(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 794 021**

(51) Int. CI.:

C12Q 1/6806 (2008.01)
C12Q 1/6869 (2008.01)
C12N 15/10 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **17.03.2014 PCT/US2014/030649**

(87) Fecha y número de publicación internacional: **18.09.2014 WO14145820**

(96) Fecha de presentación y número de la solicitud europea: **17.03.2014 E 14764477 (7)**

(97) Fecha y número de publicación de la concesión europea: **22.04.2020 EP 2969847**

(54) Título: **Etiquetado múltiple de fragmentos largos de ADN**

(30) Prioridad:

15.03.2013 US 201361801052 P
11.03.2014 US 201414205145

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
17.11.2020

(73) Titular/es:

COMPLETE GENOMICS, INC. (100.0%)
2904 Orchard Parkway
San Jose, CA 95134, US

(72) Inventor/es:

DRMANAC, RADOJE;
PETERS, BROCK A. y
ALEXEEV, ANDREI

(74) Agente/Representante:

SÁEZ MAESO, Ana

Observaciones:

Véase nota informativa (Remarks, Remarques o Bemerkungen) en el folleto original publicado por la Oficina Europea de Patentes

DESCRIPCIÓN

Etiquetado múltiple de fragmentos largos de ADN

Campo técnico

Esta descripción se refiere al análisis de ácidos nucleicos, tales como ADN genómico, que incluye la secuenciación y la determinación de haplotipos. La invención es como se define en las reivindicaciones.

Antecedentes de la invención

Existe la necesidad de métodos mejorados para determinar la contribución de los padres a los genomas de organismos superiores, es decir, separar por fases el haplotipo de los genomas. Los métodos para separar por fases el haplotipo, incluidos los métodos computacionales y separación experimental por fase, se revisan en Browning y Browning, Nature Reviews Genetics 12: 703-7014, 2011.

La mayoría de los mamíferos, incluidos los humanos, son diploides, con la mitad de los cromosomas homólogos derivados de cada progenitor. Muchas plantas tienen genomas que son poliploides. Por ejemplo, el trigo (*Triticum* spp.) tiene una ploidía que varía desde diploide (trigo Einkorn) hasta cuadriploide (trigo emmer y trigo duro) hasta hexaploide (trigo de espelta y trigo común [*T. aestivum*]).

El contexto en el que se producen variaciones en cada cromosoma individual puede tener profundos efectos sobre la expresión y regulación de genes y otras regiones transcritas del genoma. Además, determinar si ocurren dos mutaciones potencialmente perjudiciales dentro de uno o ambos alelos de un gen es de suma importancia clínica. Para las especies de plantas, el conocimiento de la contribución genética de los padres es útil para la reproducción de la progenie con rasgos deseables.

Algunos de los métodos actuales para la secuenciación del genoma completo carecen de la capacidad de ensamblar por separado los cromosomas parentales de una manera rentable y describir el contexto (haplotipos) en el que coexisten variaciones. Los experimentos de simulación muestran que la haplotipificación a nivel de cromosomas requiere información de enlace de alelos a través de un intervalo de al menos 70-100 kb.

La secuenciación de moléculas individuales de fragmentos de ADN de más de 100 kb sería útil para haplotipificación si el procesamiento de moléculas tan largas fuera factible, si la precisión de la secuenciación de moléculas individuales fuera alta y los costos de detección/instrumentos fueran bajos. Esto es muy difícil de lograr en moléculas cortas con alto rendimiento, y mucho menos en fragmentos de 100 kb.

La secuenciación más reciente del genoma humano se ha llevado a cabo en sistemas de longitud de lectura corta (<200 pb), altamente paralelizados, comenzando con cientos de nanogramos de ADN. Estas tecnologías son excelentes para generar grandes volúmenes de datos de forma rápida y económica. Desafortunadamente, las lecturas cortas, a menudo combinadas con tamaños de acoplamiento de huecos pequeños (500 bp-10 kb), eliminan la mayoría de la información de la fase SNP más allá de unas pocas kilobases (McKernan et al., Genome Res. 19: 1527, 2009). Además, es muy difícil mantener largos fragmentos de ADN en múltiples etapas de procesamiento sin fragmentación como resultado del corte.

Hasta hace poco, solo se habían secuenciado y ensamblado como diploides aproximadamente tres genomas personales: aquellos de J. Craig Venter (Levy et al., PLoS Biol. 5: e254, 2007), un indio de Gujarati (muestra HapMap NA20847; Kitzman et al., Nat. Biotechnol. 29:59, 2011) y dos europeos (Max Planck One [MP1]; Suk et al., Genome Res., 2011; y la muestra HapMap NA 12878; Duitama et al., Nucl. Acids Res.40: 2041-2053, 2012). Todos han implicado la clonación de fragmentos largos de ADN en constructos en un proceso similar a la secuenciación del cromosoma artificial bacteriano (BAC) utilizada durante la construcción del genoma de referencia humano (Venter et al., Science 291: 1304, 2001; Lander et al., Nature 409 : 860, 2001). Si bien estos procesos generan cóntigos en fases largas (N50 de 350 kb [Levy et al., PLoS Biol. 5: e254, 2007], 386 kb [Kitzman et al., Nat. Biotechnol. 29: 59-63, 2011] y 1Mb [Suk et al., Genome Res. 21: 1672-1685, 2011]) requieren una gran cantidad de ADN inicial, un extenso procesamiento de la biblioteca y son demasiado costosos para usar en un entorno clínico de rutina.

Además, se ha demostrado la haplotipificación de cromosomas completos a través del aislamiento directo de los cromosomas de la metafase (Zhang et al., Nat. Genet. 38: 382-387, 2006; Ma et al., Nat. Methods 7: 299-301, 2010; Fan et al., Nat. Biotechnol. 29: 51-57, 2011; Yang et al., Proc. Natl. Acad. Sci. USA 108: 12-17, 2011). Estos métodos son útiles para la haplotipificación de largo alcance, pero aún no se han utilizado para la secuenciación del genoma completo; requieren la preparación y el aislamiento de cromosomas de la metafase completa, lo que puede ser un desafío para algunas muestras clínicas. El documento WO 2012/106546 A2 describe métodos de captura masivamente paralela de información de contigüidad a diferentes escalas para lograr el ensamblaje de *novo* de genomas de mamíferos y la resecuenciación de genomas humanos resuelta por haplotipo. El documento WO 2014/108810 A2 (que se publicó después de la fecha de presentación de esta solicitud de patente) describe métodos para usar transposasa inmovilizada y un extremo de transposón para generar una biblioteca inmovilizada de ADN objetivo de cadena doble etiquetado en 5' en una superficie. Los métodos son útiles para generar fragmentos de ADN etiquetados en 5' y 3' para su uso en secuenciación de ADN masivamente paralela.

También existe la necesidad de métodos mejorados para obtener información de secuencia a partir de mezclas de organismos tales como en metagenómica (por ejemplo, bacterias intestinales u otros microbiomas). También existe la necesidad de métodos mejorados para la secuenciación y el ensamblaje del genoma, incluido el ensamblaje *de novo* sin uso o con un uso mínimo de una secuencia de referencia), o el ensamblaje de genomas que incluye varios tipos de secuencias repetidas, incluyendo la resolución de pseudogenes, variaciones en el número de copias y variaciones estructurales, especialmente en genomas de cáncer.

Se han descrito métodos de lectura de fragmentos largos (LFR) que proporcionan un ensamblaje preciso de secuencias separadas de cromosomas parentales (es decir, haplotipificación completo) en genomas diploides a costos experimentales y computacionales significativamente reducidos y sin clonación en vectores y replicación basada en células. La LFR se basa en la separación física de fragmentos largos de ADN genómico (u otros ácidos nucleicos) a través de muchas alícuotas diferentes, de modo que existe una baja probabilidad de que cualquier región determinada del genoma del componente materno y paterno esté representada en la misma alícuota. Al colocar un identificador único en cada alícuota y analizar muchas alícuotas en el agregado, los datos de la secuencia de ADN se pueden ensamblar en un genoma diploide, por ejemplo, se puede determinar la secuencia de cada cromosoma parental. La LFR no requiere la clonación de fragmentos de un ácido nucleico complejo en un vector, como en los enfoques de haplotipificación que usan bibliotecas de fragmentos grandes (por ejemplo, BAC). La LFR tampoco requiere el aislamiento directo de los cromosomas individuales de un organismo. Además, la LFR se puede realizar en un organismo individual y no requiere una población del organismo para lograr la separación por fases del haplotipo.

Los métodos de LFR se han descrito en las solicitudes de patente de los Estados Unidos Nos. 12/329.365 y 13/447.087, las publicaciones de patente de los Estados Unidos 2011-0033854 y 2009-0176234, y las patentes delos Estados Unidos Nos. 7.901.890, 7.897.344, 7.906.285, 7.901.891 y 7.709.197.

Sumario de la invención

La invención proporciona métodos para el etiquetado múltiple de fragmentos largos de ADN individuales (a los que se hace referencia en el presente documento mediante la abreviatura marcación múltiple, o MT). MT es útil para el análisis de ácidos nucleicos, como el ADN genómico, incluida la secuenciación y para analizar la información de secuencia resultante para reducir errores, realizar separación por fases de haplotipos, entre otras cosas, y realizar llamadas de variantes precisas, especialmente para heterocigotos. La invención es como se define en las reivindicaciones. La presente invención proporciona un método para preparar ADN genómico para análisis de secuencia, por reacción homogénea sin el uso de compartimentación física tal como nanogotas, comprendiendo el método:

(a) combinar una pluralidad de fragmentos largos que comprenden secuencias de ADN genómico en una sola mezcla con una población de perlas, en la que (i) los fragmentos largos son de 5 kilobases a 750 kilobases de longitud, (ii) cada perla comprende al menos 1000 copias del mismo oligonucleótido inmovilizado sobre ella, dicho oligonucleótido que comprende una secuencia que contiene una etiqueta, (iii) cada secuencia que contiene una etiqueta comprende una secuencia de etiqueta, y (iv) la población de perlas comprende, en conjunto, al menos 10.000 secuencias de etiquetas diferentes;

(b) producir fragmentos largos etiquetados incorporando en cada uno de una pluralidad de fragmentos largos individuales múltiples copias de una secuencia de etiqueta, en la que en una pluralidad de dichos fragmentos largos individuales dichas copias múltiples son de una sola perla;

(c) producir subfragmentos de los fragmentos largos etiquetados, en la que una pluralidad de subfragmentos del mismo fragmento largo etiquetado comprende secuencias de etiquetas de la misma perla;

en la que la etapa (b) se lleva a cabo bajo condiciones que promueven la interacción de solo una secuencia de etiqueta por fragmento largo

Se proporcionan métodos en esta descripción para secuenciar un ácido nucleico objetivo mediante: (a) combinación en un único recipiente de reacción (i) una pluralidad de fragmentos largos del ácido nucleico objetivo, y (ii) una población de polinucleótidos, en la que cada polinucleótido comprende una etiqueta y la mayoría de los polinucleótidos comprende una etiqueta diferente; (b) introducir en la mayoría de los fragmentos largos secuencias que contienen una etiqueta de dicha población de polinucleótidos para producir fragmentos largos etiquetados, en los que cada uno de los fragmentos largos etiquetados comprende una pluralidad de secuencias que contienen una etiqueta en un espaciado promedio seleccionado, y cada secuencia que contiene una etiqueta comprende una etiqueta; y (c) producir una pluralidad de subfragmentos a partir de cada fragmento largo etiquetado, en el que cada subfragmento comprende una o más etiquetas. Dichos métodos son adecuados para preparar un ácido nucleico objetivo para la secuenciación de ácido nucleico, y pueden comprender la secuenciación de los subfragmentos para producir una pluralidad de lecturas de secuencia; asignar la mayoría de las lecturas de secuencia a los fragmentos largos correspondientes; y el ensamblaje de la secuencia se lee para producir una secuencia ensamblada del ácido nucleico objetivo.

La producción de los subfragmentos etiquetados por tales métodos puede comprender realizar una reacción de amplificación para producir una pluralidad de amplicones a partir de cada fragmento largo. Cada amplicón puede comprender una etiqueta de cada una de las secuencias introducidas adyacentes y una región del fragmento largo entre las secuencias introducidas adyacentes. Tales métodos pueden comprender combinar los fragmentos largos

con un exceso de la población de secuencias que contienen una etiqueta; y/o la combinación de los fragmentos largos con la solución que contiene la etiqueta en condiciones adecuadas para la introducción de una secuencia que contiene una sola etiqueta en la mayoría de los fragmentos largos.

Tales métodos pueden comprender combinar los fragmentos largos con la solución que contiene la etiqueta en condiciones que sean adecuadas para la introducción de diferentes secuencias que contienen una etiqueta en la mayoría de los fragmentos largos. La población de secuencias que contienen una etiqueta puede comprender una población de perlas, en la que cada perla comprende múltiples copias de una secuencia que contiene una sola etiqueta. En tales métodos, las secuencias que contienen una etiqueta comprenden típicamente extremos de transposones, el método típicamente comprende combinar los fragmentos largos y las secuencias que contienen una etiqueta bajo condiciones que son adecuadas para la transposición de las secuencias que contienen una etiqueta en cada uno de los fragmentos largos. Alternativamente, las secuencias que contienen una etiqueta pueden ser una secuencia de horquilla. El ácido nucleico objetivo puede ser un ácido nucleico complejo, tal como el genoma de un organismo. Tales métodos pueden hacerse para determinar un haplotipo del genoma, o para cualquier otro propósito que valga la pena.

A menos que se indique o requiera lo contrario, cualquier método para analizar o secuenciar de acuerdo con esta invención puede comprender amplificar porciones del ácido nucleico objetivo para formar los fragmentos iniciales. Esto se puede hacer, por ejemplo, insertando transposones en el ácido nucleico objetivo; y replicar el ácido nucleico objetivo usando cebadores que se unen dentro de los transposones, formando así los fragmentos iniciales. Por lo tanto, la amplificación puede comprender las etapas de ligar oligonucleótidos adaptadores en una pluralidad de muescas o huecos; y replicar el ácido nucleico objetivo usando cebadores que se unen dentro de los oligonucleótidos adaptadores, formando así los fragmentos iniciales. La amplificación puede realizarse con transposones, muescas o huecos introducidos en el polinucleótido objetivo a una frecuencia de uno en aproximadamente cada 3 a 20 kb, o como se ejemplifica en otra parte de esta descripción.

El ácido nucleico objetivo puede ser un ácido nucleico complejo, como el genoma de un organismo. El análisis puede incluir determinar el haplotipo de un genoma, determinar los patrones de metilación de un genoma; y/o determinar la variación del número de copias en una muestra celular presente, por ejemplo en la muestra de biopsia tomada de un paciente con cáncer. Los métodos de esta invención pueden usarse para diagnosticar o evaluar el cáncer en un paciente, o para el diagnóstico genético previo a la implantación.

Los productos para llevar a cabo un método de esta invención incluyen cualquier nuevo constructo o complejo de ácido nucleico descrito a continuación o mostrado en las figuras, opcionalmente en combinación con otros componentes útiles para la secuenciación o análisis de ADN complejo. Tales componentes pueden incluir un sustrato o reactivo de partida, un producto intermedio o final de un método de la invención como se describe a continuación. Por ejemplo, un sistema para secuenciar o analizar un ácido nucleico objetivo incluye (a) fragmentos del ácido nucleico objetivo que tienen un tamaño específico (por ejemplo, aproximadamente 2 a 5 o aproximadamente 5 a 750 pares de bases de longitud), una pluralidad de cada uno de los cuales contiene o está hibridado con múltiples copias de una secuencia de inserción que comprende una etiqueta particular, en la que diferentes fragmentos contienen secuencias de inserción con una etiqueta diferente y una secuencia de cebador común; y (b) un conjunto de cebadores que comprende una secuencia que se hibrida específicamente con la secuencia del cebador común.

Se describe un método para el análisis de secuencia de un ácido nucleico objetivo que comprende: (a) combinar una pluralidad de fragmentos largos de ADN del ácido nucleico objetivo con una población de secuencias que contienen una etiqueta, en la que la población comprende al menos 1000 secuencias de etiquetas diferentes; (b) producir fragmentos largos etiquetados, en los que cada fragmento largo etiquetado comprende una secuencia de ácido nucleico objetivo y múltiples secuencias de etiquetas intercaladas, en las que las múltiples secuencias de etiquetas intercaladas en un fragmento largo etiquetado individual pueden ser iguales o diferentes; (c) producir a partir de cada fragmento largo etiquetado una pluralidad de subfragmentos etiquetados, en los que los subfragmentos etiquetados comprenden cada uno una o más secuencias de etiquetas; (d) obtener la secuencia de subfragmentos etiquetados individuales, en los que la secuencia obtenida incluye la secuencia de ácido nucleico objetivo y al menos una secuencia de etiqueta; (e) combinar secuencias obtenidas en (d) para producir secuencia o secuencias ensambladas del ácido nucleico objetivo, en el que la combinación comprende (i) determinar que las secuencias obtenidas en (d) se originaron a partir del mismo fragmento largo de ADN si dichas secuencias comprenden la misma secuencia de etiqueta y/o (ii) identificar pares de secuencias como secuencias adyacentes en el ácido nucleico objetivo si el par comprende la misma secuencia de etiqueta. En un aspecto, las etapas (a) - (c) se llevan a cabo en un solo recipiente o mezcla. En un aspecto, las etapas de la pluralidad de fragmentos largos de ADN son secuencias de ADN genómico. En un aspecto, las etapas de la pluralidad de fragmentos largos de ADN tienen al menos 50 kb, opcionalmente al menos 100 kb, de longitud, o están en el intervalo de 50 kb a 200 kb. En algunas realizaciones, los fragmentos largos etiquetados comprenden una pluralidad de secuencias que contienen una etiqueta a un espaciado promedio seleccionado. En algunas realizaciones, la separación promedio está en el intervalo de 100 a 5000 bases. En algunas realizaciones, la separación promedio está en el intervalo de 200 y 1500 bases. En algunas realizaciones, la separación promedio está en el intervalo de 250 y 1000 bases.

En un aspecto, las etapas (a) - (c) se llevan a cabo en un único recipiente o mezcla y el recipiente único o la mezcla comprende más de una cantidad de haploides (N) de ADN genómico. En algunas realizaciones, el ADN genómico es

de un solo organismo. En algunas realizaciones, el ADN genómico comprende ADN fetal y ADN materno. En algunas realizaciones, el ADN genómico es ADN de 1-100 células eucariotas. En algunas realizaciones, el ADN genómico es ADN de 2 a 10 células eucariotas. En algunas realizaciones, el ADN genómico es ADN de más de 50 células eucariotas. En algunas realizaciones, el ADN genómico se obtiene de una mezcla que comprende más de un tipo celular. En algunas realizaciones, el ADN se obtiene de una mezcla que comprende más de un tipo celular de la misma especie. En algunas realizaciones, las células son (i) células fetales y células maternas o (ii) células tumorales y células normales.

En algunas realizaciones, los fragmentos largos de ADN son fragmentos de ADN cromosómico. En algunas realizaciones, los fragmentos largos de ADN son amplicones de ADN celular. En algunas realizaciones, los fragmentos largos de ADN son productos de amplificación del genoma completo. En algunas realizaciones, comprende porciones de amplificación del ácido nucleico objetivo para formar los fragmentos largos de ADN usados en la Etapa (a).

En un aspecto, las secuencias que contienen una etiqueta son etiquetas clonales y la población de secuencias que contienen una etiqueta es una población de fuentes de etiquetas clonales. Las fuentes de las etiquetas clonales son perlas, en las que cada perla tiene múltiples copias de una secuencia de una sola etiqueta inmovilizada luego. Las fuentes de etiquetas clonales comprenden cada una al menos 1000 copias de una única secuencia de etiqueta. En algunas realizaciones, las secuencias que contienen una etiqueta comprenden extremos de transposones. En algunas realizaciones, las secuencias que contienen una etiqueta comprenden extremos de transposones. En algunas realizaciones, las secuencias que contienen una etiqueta son oligonucleótidos que adoptan una conformación en horquilla. En algunas realizaciones, cada oligonucleótido comprende dos secuencias de etiqueta. En algunas realizaciones, las dos secuencias de etiquetas son iguales. En algunas realizaciones, las secuencias de la población que contienen una etiqueta comprenden secuencias de unión a cebador. En algunas realizaciones, cada una de las secuencias de la población que contienen una etiqueta comprende las mismas secuencias de unión a cebador o combinación de secuencias de unión a cebador.

En un aspecto, la Etapa (a) comprende combinar los fragmentos largos de ADN y las secuencias que contienen una etiqueta en condiciones que sean adecuadas para la transposición de las secuencias de etiquetadas en los fragmentos largos de ADN. En un aspecto relacionado, la etapa (a) comprende combinar los fragmentos largos de ADN y una población de fuentes de etiquetas clonales. En algunas realizaciones, el método comprende combinar los fragmentos largos de ADN con un exceso de secuencias que contienen una etiqueta o fuentes de secuencias que contienen una etiqueta.

La invención comprende combinar los fragmentos largos de ADN con las secuencias que contienen una etiqueta en condiciones adecuadas para la introducción de copias múltiples de una secuencia de una sola etiqueta en los fragmentos largos de ADN, en los que al menos el 20% de los fragmentos largos de ADN comprenden solo una secuencia de etiqueta. En algunas realizaciones, las condiciones son tales que la mayoría de los fragmentos largos de ADN en los que se introduce una secuencia de etiqueta comprenden una secuencia de etiqueta introducida única.

En un aspecto, en promedio, cada fragmento largo etiquetado que comprende secuencias de etiquetas intercaladas comprende al menos 10 secuencias de etiqueta. En algunas realizaciones, las secuencias de etiquetas intercaladas múltiples en un fragmento largo etiquetado individualmente son las mismas. En algunas realizaciones, más del 10% de la longitud de ADN de un fragmento largo de ADN está representado por subfragmentos etiquetados.

En algunas realizaciones, las etapas (b) y (c) comprenden: (i) crear muescas o huecos en los fragmentos largos de ADN que producen los extremos 3’ libres, (ii) ligar una secuencia adaptadora común 3' a los terminales 3’ libres, (iii) hibridar oligonucleótidos con la secuencia adaptadora común 3’, en la que los oligonucleótidos comprenden cada uno una secuencia de etiqueta; y luego (iv) extender el primer oligonucleótido para formar subfragmentos etiquetados. En algunas realizaciones en la etapa (b), los transposones, muescas o huecos se introducen en el fragmento largo de ADN con una frecuencia de uno en aproximadamente cada 300 a 1000 bases.

En un aspecto, producir los subfragmentos comprende realizar una reacción de amplificación para producir amplicones a partir de los fragmentos largos etiquetados. En algunas realizaciones, la reacción de amplificación es PCR. En algunas realizaciones, cada amplicón comprende una etiqueta de cada una de las secuencias introducidas adyacentes y una región del fragmento largo etiquetado entre las secuencias introducidas adyacentes.

En algunas realizaciones, la Etapa (c) comprende formar múltiples subfragmentos etiquetados que contienen cada uno una porción de un transposón etiquetado y una porción del fragmento largo de ADN. En algunas realizaciones, los transposones etiquetados tienen una secuencia de etiqueta en o cerca de un extremo que es lo mismo que una etiqueta o secuencia en o cerca del otro extremo.

En un aspecto, los subfragmentos etiquetados se forman por amplificación, usando un cebador o cebadores que se hibridan con una secuencia o secuencias dentro de una secuencia o secuencias que contienen una etiqueta.

En un aspecto, el método comprende: (i) proporcionar cebadores que comprenden cada uno una secuencia de etiqueta y una secuencia de sonda aleatoria; (ii) hibridar los cebadores mediante sus respectivas secuencias de sonda con los fragmentos largos de ADN; y (iii) extender los cebadores para formar múltiples subfragmentos etiquetados. En algunas realizaciones, la etapa (i) comprende: hibridar copias de un oligonucleótido adaptador común con una secuencia de

etiqueta en cada una de una pluralidad de perlas que es diferente de las secuencias de etiquetas en otras perlas; e hibridar las copias con una pluralidad de diferentes secuencias de sonda aleatorias; y extender las copias para formar dichos cebadores. En algunas realizaciones, la formación de la muesca o hueco y la liberación de las secuencias de etiquetas de las perlas se realiza en la misma mezcla de reacción.

En un aspecto, los subfragmentos etiquetados se forman escindiendo entre dos códigos de barras presentes en la misma secuencia que contiene una etiqueta.

En algunos aspectos de la invención, la secuencia de subfragmentos etiquetados individuales se obtiene mediante secuenciación por hibridación, secuenciación por ligadura, secuenciación por síntesis, secuenciación de una sola molécula, detección óptica de la secuencia, detección electromagnética de la secuencia o detección de la secuencia por cambio de voltaje

En algunas realizaciones, la combinación en la Etapa (e) comprende determinar que las secuencias obtenidas en (d) se originaron a partir del mismo fragmento largo de ADN si dichas secuencias comprenden la misma secuencia de etiqueta.

En aspectos de la invención, el método comprende determinar un haplotipo del genoma. En aspectos de la invención, el método comprende el análisis de metilación de un genoma. En aspectos de la invención, el método comprende determinar la variación del número de copias en las células cancerosas. En aspectos de la invención, el método comprende el diagnóstico genético previo a la implantación.

En la invención, el ácido nucleico objetivo es el ADN genómico de un organismo. En algunas realizaciones, el ADN genómico es de una planta o animal. En algunas realizaciones, el animal es un mamífero. En algunas realizaciones, el animal es un humano.

En un aspecto, la divulgación proporciona un método para el análisis de secuencia de una o más moléculas de ácido nucleico objetivo que comprende: (a) producir una población de subfragmentos de un solo fragmento largo etiquetado del ácido nucleico objetivo, en el que el fragmento largo etiquetado comprende la secuencia de ácido nucleico objetivo y secuencias de etiquetas intercaladas múltiples, en las que la mayoría de los subfragmentos comprenden secuencia de ácido nucleico objetivo y al menos una secuencia de etiqueta; (b) obtener la secuencia de subfragmentos etiquetados individuales, en los que la secuencia obtenida incluye la secuencia de ácido nucleico objetivo y al menos una secuencia de etiqueta; (c) combinar secuencias obtenidas en (b) para producir una secuencia o secuencias ensambladas del ácido nucleico objetivo, en las que la combinación comprende determinar que las secuencias obtenidas en (b) se originaron a partir del mismo fragmento largo de ADN si dichas secuencias comprenden la misma secuencia de etiqueta.

En un aspecto, la divulgación proporciona un método para el análisis de secuencia de una o más moléculas de ácido nucleico objetivo que comprende: (a) obtener una población de subfragmentos de un fragmento largo etiquetado del ácido nucleico objetivo, en el que el fragmento largo etiquetado comprende la secuencia de ácido nucleico objetivo y múltiples secuencias de etiquetas intercaladas, en las que la mayoría de los subfragmentos comprenden secuencia de ácido nucleico objetivo y al menos una secuencia de etiqueta; (b) obtener la secuencia de subfragmentos etiquetados individuales, en los que la secuencia obtenida incluye la secuencia de ácido nucleico objetivo y al menos una secuencia de etiqueta; (c) combinar secuencias obtenidas en (b) para producir una secuencia o secuencias ensambladas del ácido nucleico objetivo, en el que la combinación comprende determinar que las secuencias obtenidas en (b) se originaron a partir del mismo fragmento largo de ADN si dichas secuencias comprenden la misma secuencia de etiqueta.

En un aspecto, la divulgación proporciona un método para el análisis de secuencia de una o más moléculas de ácido nucleico objetivo que comprende: (a) obtener lecturas de secuencia de subfragmentos de un fragmento largo etiquetado del ácido nucleico objetivo, en el que los subfragmentos de los que se obtienen las lecturas de secuencias comprenden una secuencia de ácido nucleico objetivo y una secuencia de etiqueta; (b) ensamblar las lecturas de secuencia para producir una secuencia o secuencias ensambladas del ácido o ácidos nucleicos objetivo, en los que el ensamblaje comprende ensamblar secuencias objetivo adyacentes basadas en la presencia de secuencias de etiquetas comunes en pares de lecturas correspondientes a pares de secuencias objetivo adyacentes.

En un aspecto, la divulgación proporciona un método de secuenciación de un ácido nucleico objetivo que comprende: combinar en un único recipiente de reacción (i) una pluralidad de fragmentos largos del ácido nucleico objetivo, y (ii) una población de polinucleótidos, en los que cada polinucleótido comprende una etiqueta y la mayoría de los polinucleótidos comprenden una etiqueta diferente; introducir en la mayoría de los fragmentos largos secuencias que contienen una etiqueta de dicha población de polinucleótidos para producir fragmentos largos etiquetados, en los que cada uno de los fragmentos largos etiquetados comprende una pluralidad de secuencias que contienen una etiqueta en un espaciado promedio seleccionado, y cada secuencia que contiene una etiqueta comprende una etiqueta. En un aspecto, la invención proporciona un método para el análisis de secuencia de un ácido nucleico objetivo que comprende: (a) combinar una pluralidad de fragmentos largos de ADN del ácido nucleico objetivo con una población de secuencias que contienen una etiqueta; (b) producir fragmentos largos etiquetados, en los que cada fragmento largo etiquetado comprende una secuencia de ácido nucleico objetivo y múltiples secuencias de etiquetas intercaladas,

en las que las múltiples secuencias de etiquetas intercaladas en un fragmento largo etiquetado individual pueden ser iguales o diferentes. En algunas realizaciones, las etapas (a) y (b) se llevan a cabo en un solo tubo o mezcla. En un aspecto, el método incluye (c) producir a partir de cada fragmento largo etiquetado una pluralidad de subfragmentos etiquetados, en los que los subfragmentos etiquetados comprenden cada uno una o más secuencias de etiquetas.

En un aspecto, la divulgación proporciona un método de secuenciación de un ácido nucleico objetivo que comprende: combinar en un único recipiente de reacción (i) una pluralidad de fragmentos largos del ácido nucleico objetivo, y (ii) una población de polinucleótidos, en la que cada polinucleótido comprende una etiqueta y la mayoría de los polinucleótidos comprenden una etiqueta diferente; introducir en la mayoría de los fragmentos largos secuencias que contienen una etiqueta de dicha población de polinucleótidos para producir fragmentos largos etiquetados, en los que cada uno de los fragmentos largos etiquetados comprende una pluralidad de secuencias que contienen una etiqueta en un espaciado promedio seleccionado, y cada etiqueta que contiene la secuencia comprende una etiqueta; producir una pluralidad de subfragmentos de cada fragmento largo etiquetado, en el que cada subfragmento comprende una o más etiquetas; secuenciar los subfragmentos para producir una pluralidad de lecturas de secuencia; asignar la mayoría de las lecturas de secuencia a fragmentos largos correspondientes; y ensamblar las lecturas de secuencia para producir una secuencia ensamblada del ácido nucleico objetivo. En algunas realizaciones, la producción de los subfragmentos comprende realizar una reacción de amplificación para producir una pluralidad de amplicones a partir de cada fragmento largo. En algunas realizaciones, cada amplicón comprende una etiqueta de cada una de las secuencias introducidas adyacentes y una región del fragmento largo entre las secuencias introducidas adyacentes. En algunas realizaciones, el método comprende combinar los fragmentos largos con un exceso de la población de secuencias que contienen una etiqueta. En algunas realizaciones, el método comprende combinar los fragmentos largos con la solución que contiene la etiqueta en condiciones que son adecuadas para la introducción de una secuencia única que contiene la etiqueta en la mayoría de los fragmentos largos. En algunas realizaciones, el método comprende combinar los fragmentos largos con la solución que contiene la etiqueta en condiciones que son adecuadas para la introducción de diferentes secuencias que contienen una etiqueta en la mayoría de los fragmentos largos. La población de secuencias que contienen una etiqueta comprende una población de perlas, en la que cada perla comprende múltiples copias de una sola secuencia que contiene una etiqueta. En algunas realizaciones, las secuencias que contienen una etiqueta comprenden extremos de transposones, comprendiendo el método combinar los fragmentos largos y las secuencias que contienen una etiqueta en condiciones que son adecuadas para la transposición de las secuencias que contienen una etiqueta en cada uno de los fragmentos largos. En algunas realizaciones, las secuencias que contienen una etiqueta comprenden una secuencia de horquilla. En algunas realizaciones, el ácido nucleico objetivo es un ácido nucleico complejo. En algunas realizaciones, el ácido nucleico objetivo es un genoma de un organismo. En algunas realizaciones, el método comprende determinar un haplotipo del genoma. En algunas realizaciones, la población de secuencias que contienen una etiqueta comprende al menos 10.000 secuencias de etiquetas diferentes. En algunas realizaciones, la población de secuencias que contienen una etiqueta comprende al menos 100.000 secuencias de etiquetas diferentes.

Una composición puede comprender al menos 10^3 elementos de ácido nucleico que contienen una etiqueta diferente y al menos una de (i) ADN genómico y (ii) cebadores que se unen a los elementos de ácido nucleico que contienen una etiqueta. En algunos casos, la composición comprende al menos 5 equivalentes de genoma de ADN genómico. En algunos casos, la composición comprende tanto ADN genómico como cebadores. En algunos casos, la composición que comprende fragmentos largos etiquetados comprende una secuencia de ácido nucleico genómico y múltiples secuencias de etiquetas intercaladas.

Un kit comprende una biblioteca que comprende 10^3 o más códigos de barras distintos o fuentes de códigos de barras clonales: i) una biblioteca de códigos de barras asociada con extremos de transposones, y opcionalmente secuencias adaptadoras; ii) una biblioteca de códigos de barras clonales, opcionalmente con secuencias adaptadoras, que comprende una pluralidad de 10^4 o más fuentes distintas de códigos de barras clonales; iii) una biblioteca de concatámeros que comprende monómeros, en la que los monómeros comprenden códigos de barras; iv) una biblioteca de plantillas adecuadas para la amplificación de círculo rodante, en la que las plantillas comprenden un monómero como se describe en (iii); y/o v) una biblioteca de oligonucleótidos de horquilla, comprendiendo cada oligonucleótido dos copias de una secuencia de código de barras, en la que la biblioteca comprende una pluralidad de al menos aproximadamente 10^4 códigos de barras. En algunos casos, el kit comprende una enzima seleccionada de una transposasa, una polimerasa, una ligasa, una endonucleasa y una exonucleasa. En algunos casos, el kit comprende al menos aproximadamente 10^4, al menos aproximadamente 10^5, al menos aproximadamente 10^6 o al menos aproximadamente 10^7 códigos de barras diferentes. En algunos casos, el kit comprende al menos aproximadamente 10^4, al menos aproximadamente 10^5, al menos aproximadamente 10^6, o al menos aproximadamente 10^7 códigos de barras diferentes o fuentes de códigos de barras clonales. En algunos casos, los miembros de la biblioteca comprenden una o dos secuencias comunes para la unión del cebador. En algunos casos, el kit comprende un cebador o cebadores que se hibridan con una secuencia o secuencias dentro de la secuencia que contiene una etiqueta.

Otros aspectos de la invención serán evidentes a partir de la descripción que sigue.

Breve descripción de los dibujos

Las Figuras 1A y 1B muestran un método para el etiquetado y la fragmentación de fragmentos largos de un ácido nucleico objetivo con códigos de barras mediados por transposones.

Las Figuras 2A y 2B muestran un método para el etiquetado y la fragmentación de fragmentos largos de un ácido nucleico objetivo con códigos de barras mediados por horquillas.

La Figura 3 muestra un método para el etiquetado y fragmentación mediado por transposones de fragmentos largos de un ácido nucleico objetivo.

La Figura 4A muestra un método para etiquetar fragmentos largos de un ácido nucleico objetivo usando un adaptador etiquetado. La Nickasa y la exonucleasa Klenow 3'-5', sin dNTPS, se utilizan para crear sitios aleatorios a lo largo del ADNbc para la ligadura del adaptador común 3'. Se pueden obtener resultados similares usando cualquier otra enzima de corte y/o actividad de exonucleasa. Las perlas con muchas copias de la etiqueta y complementariedad con el adaptador común 3' se agregan y se fragmentan con endonucleasa de restricción. DNB o perlas está cerca del ADNbc, por lo que la mayoría de las copias de etiquetas no se difunden, sino que se hibridan con un adaptador 3'. El ADNbc largo y DNB o perla se pueden etiquetar en un extremo para forzar la interacción si es necesario. La extensión del cebador crea un fragmento de ADN genómico etiquetado. La extensión adicional del cebador crea ADNbc que puede ligarse y amplificarse y secuenciarse mediante PCR. El ensamblaje *in silico* en fragmentos largos de ADN es similar a la Figura 3.

La Figura 4B muestra un método alternativo para etiquetar fragmentos largos de un ácido nucleico objetivo usando un adaptador etiquetado.

Las Figuras 4C y 4D muestran un segundo método alternativo para etiquetar fragmentos largos de un ácido nucleico objetivo usando un adaptador etiquetado.

Las Figuras 4E y 4F muestran métodos para crear una serie de subfragmentos etiquetados con regiones cada vez más cortas de los fragmentos largos de ADN.

La Figura 4G muestra un método para etiquetar fragmentos largos de un ácido nucleico objetivo usando traducción controlada del corte.

Las Figuras 5A y 5B muestran ejemplos de sistemas de secuenciación.

La Figura 6 muestra un ejemplo de un dispositivo informático que se puede utilizar en, o junto con una máquina de secuenciación y/o un sistema informático.

La Figura 7 muestra la arquitectura general del algoritmo MT.

La Figura 8 muestra el análisis por parejas de SNP heterocigotos cercanos.

La Figura 9 muestra un ejemplo de la selección de una hipótesis y la asignación de un puntaje a la hipótesis.

La Figura 10 muestra la construcción del gráfico.

La Figura 11 muestra la optimización del gráfico.

La Figura 12 muestra la alineación del cóntigo.

La Figura 13 muestra la separación universal por fases asistida por los padres.

La Figura 14 muestra separaciones naturales de cóntigos.

La Figura 15 muestra la separación universal por fases.

La Figura 16 muestra la detección de errores usando MT.

La Figura 17 muestra un ejemplo de un método para disminuir el número de falsos negativos en el que se podría realizar una llamada segura de SNP heterocigóticos a pesar del pequeño número de lecturas.

Descripción detallada

Como se usa en este documento y en las reivindicaciones adjuntas, las formas singulares "un", "uno, una" y "el, la" incluyen referentes plurales a menos que el contexto indique claramente lo contrario. Así, por ejemplo, la referencia a "una polimerasa" se refiere a un agente o mezclas de tales agentes, y la referencia al "método" incluye referencia a etapas equivalentes y/o métodos conocidos por los expertos en la técnica, y así sucesivamente.

A menos que se defina lo contrario, todos los términos técnicos y científicos utilizados en este documento tienen el mismo significado que el entendido comúnmente por un experto en la materia a la que pertenece esta invención. Todas las publicaciones mencionadas en el presente documento tienen el propósito de describir y divulgar dispositivos, composiciones, formulaciones y metodologías que se describen en la publicación y que podrían usarse en relación con la invención actualmente descrita.

Cuando se proporciona un intervalo de valores, se entiende que cada valor interviniente, hasta la décima parte de la unidad del límite inferior, a menos que el contexto indique claramente lo contrario, entre el límite superior e inferior de ese intervalo y cualquier otro establecido o el valor interviniente en ese intervalo establecido está incluido dentro de la invención. Los límites superior e inferior de estos intervalos más pequeños pueden incluirse independientemente en los intervalos más pequeños también pueden estar abarcados dentro de la invención, sujeto a cualquier límite específicamente excluido en el intervalo establecido. Cuando el intervalo establecido incluye uno o ambos límites, los intervalos que excluyen ambos límites incluidos también se incluyen en la invención.

En la siguiente descripción, se exponen numerosos detalles específicos para proporcionar una comprensión más completa de la presente invención. Sin embargo, será evidente para un experto en la materia que la presente invención se puede practicar sin uno o más de estos detalles específicos. En otros casos, no se han descrito características y procedimientos bien conocidos por los expertos en la materia para evitar enredar la invención.

Aunque la presente invención se describe principalmente con referencia a realizaciones específicas, también se prevé que otras realizaciones serán evidentes para los expertos en la materia al leer la presente divulgación, y se pretende que tales realizaciones estén contenidas dentro de los métodos de la presente invención.

La práctica de la presente invención puede emplear, a menos que se indique lo contrario, técnicas convencionales y descripciones de química orgánica, tecnología de polímeros, biología molecular (incluyendo técnicas recombinantes), biología celular, bioquímica e inmunología, que están dentro de los conocimientos del estado de la técnica. Dichas técnicas convencionales incluyen síntesis de matriz polimérica, hibridación, ligadura y detección de hibridación usando una etiqueta. Pueden obtenerse ilustraciones específicas de técnicas adecuadas haciendo referencia al siguiente ejemplo a continuación. Sin embargo, también se pueden utilizar otros procedimientos convencionales equivalentes. Dichas técnicas y descripciones convencionales se pueden encontrar en los manuales estándar de laboratorio tales como Genome Analysis: A Laboratory Manual Series (volúmenes I-IV), Using Antibodies: A Laboratory Manual, Cells: A Laboratory Manual, PCR Primer: A Laboratory Manual, and Molecular Cloning: A Laboratory Manual (todos de Cold Spring Harbor Laboratory Press), Stryer, L. (1995) Biochemistry (4ª edición.) Freeman, New York, Gait, "Oligonucleotide Synthesis: A Practical Approach" 1984, IRL Press, Londres, Nelson y Cox (2000), Lehninger, Principles of Biochemistry 3z Ed., W. H. Freeman Pub., New York, N.Y. y Berg et al., (2002) Biochemistry, 5a Ed., W. H. Freeman Pub., New York, N.Y.

Visión general

De acuerdo con un aspecto de la invención, se proporcionan métodos para el etiquetado múltiple de fragmentos largos individuales de ácidos nucleicos objetivo, o polinucleótidos, que incluyen, sin limitación, ácidos nucleicos complejos. Los fragmentos largos de un ácido nucleico o polinucleótido objetivo se etiquetan mediante un método que introduce una etiqueta o código de barras en múltiples sitios en cada fragmento largo. En principio, cada fragmento puede haber introducido múltiples copias de una etiqueta única, una etiqueta específica de fragmento, o un patrón único de inserción de etiquetas múltiples, un patrón de etiqueta específico de fragmento. Sin embargo, esto no es requerido. Como se discute a continuación, en algunas realizaciones, algunos fragmentos largos pueden no tener etiqueta insertada. Además, en algunas realizaciones, un fragmento largo puede tener insertado en él más de una etiqueta distinta, y dos o más fragmentos pueden tener insertado en ellos la misma etiqueta.

Los "fragmentos largos" son polinucleótidos de más de 10 kb de longitud, más a menudo más de 20 kb de longitud, incluso más a menudo más de 50 kb de longitud y muy a menudo 100 kb o más. Para la haplotipificación, los fragmentos largos de 100 kb o más son particularmente útiles.

Después del etiquetado, se producen subfragmentos de los fragmentos largos. En principio, cada subfragmento puede incluir al menos una etiqueta. De nuevo, esto no es obligatorio. Como se discute a continuación, en algunas realizaciones, algunos subfragmentos pueden no tener etiqueta insertada.

Comúnmente, los subfragmentos que contienen una etiqueta se amplifican (por ejemplo, por PCR). Los subfragmentos, que incluyen las etiquetas que forman parte de cada subfragmento, se secuencian luego. La secuencia de una etiqueta permite que los datos de secuencia obtenidos de cada subfragmento se asignen al fragmento largo del que se deriva el subfragmento. Esto facilita el mapeo y ensamblaje de secuencias y la ordenación de alelos (o hets) en un haplotipo de los ácidos nucleicos objetivo.

La unión o inserción de códigos de barras en fragmentos largos de ADN se puede realizar en una única mezcla o recipiente (por ejemplo, un solo tubo o un solo pozo en una placa de múltiples pozos) y el proceso puede automatizarse. Usando MT, la mezcla única en la que se produce el etiquetado contiene más de un genoma equivalente. En diversas realizaciones, la mezcla puede comprender al menos 5 equivalentes de genoma, al menos 10 equivalentes de genoma, al menos 25 equivalentes de genoma, al menos 50 equivalentes de genoma, al menos 100 equivalentes de genoma, al menos 500 equivalentes de genoma o al menos 1000 equivalentes de genoma, tal como de 5-20 equivalentes de genoma, tal como de 5-100 equivalentes de genoma, tal como de 50-1000 equivalentes de genoma.

En algunas aplicaciones, una sola célula puede analizarse en una única mezcla de MT, proporcionando solo dos cadenas complementarias (es decir, dos equivalentes del genoma) para discriminar la variación natural de los errores

introducidos por el procesamiento del ADN, por ejemplo, amplificación de sub-fragmentos.

De acuerdo con una realización, la mayoría de los subfragmentos, o 60%, 70%, 80%, 90% o más, o sustancialmente todos los subfragmentos, incluyen una secuencia de etiqueta. En un aspecto, esta invención proporciona un método para etiquetar fragmentos largos de ADN usando códigos de barras clonales. Como se detalla a continuación, "códigos de barras clonales" se refiere a una pluralidad de códigos de barras o etiquetas que tienen una secuencia común y que están físicamente asociados entre sí (en lugar de estar físicamente separados y, por ejemplo, libres de difundirse en solución). En este enfoque, una fuente de etiquetas clonales puede asociarse con un solo fragmento largo de ADN. El resultado es que una pluralidad de etiquetas clonales identificables o códigos de barras pueden estar asociados con un fragmento de ADN y no con otros. Las etiquetas clonales o códigos de barras pueden mantenerse juntas en un portador tal como un polímero o perlas de tamaño micrométrico. El uso de códigos de barras clonales permite la preparación de millones de códigos de barras distintos a un costo relativamente modesto para su uso en MT de "de un solo tubo".

En un aspecto, MT implica (a) proporcionar (i) una biblioteca de códigos de barras clonales y (ii) fragmentos largos de ADN; (b) preparar (mediante elaboración de huecos por corte, extensión de cebador aleatorio o inserción de transposón) los fragmentos de ADN para unir códigos de barras (por ejemplo, a una distancia promedio predefinida en los fragmentos largos de ADN); (c) unir múltiples copias de códigos de barras por molécula de ADN larga (por ejemplo, a la distancia promedio predeterminada); (d) preparar (por extensión de cebador o PCR o fragmentación de ADN) múltiples fragmentos de ADN cortos a partir de un fragmento largo etiquetado con copias del mismo código de barras. Antes de la etapa (c) se producen copias de códigos de barras individuales de, por ejemplo, liberadas de un soporte (es decir, una perla).

En otro aspecto, MT implica (a) proporcionar (i) una biblioteca de códigos de barras y (ii) fragmentos largos de ADN; (b) incorporar las secuencias de código de barras en las secuencias largas de ADN (por ejemplo, a una distancia promedio predefinida en los fragmentos largos de ADN); (c) preparar múltiples subfragmentos en los que las secuencias que son del mismo fragmento largo, tales como las secuencias adyacentes entre sí en la secuencia del fragmento largo, se etiquetan con copias del mismo código de barras. En un enfoque, se usa una biblioteca de transposones de copiar y pegar que contienen códigos de barras para obtener múltiples copias de códigos de barras por un fragmento largo de ADN. Un transposón de copiar y pegar con un código de barras unido al final de un fragmento largo de ADN puede insertar una copia del código de barras y las secuencias asociadas en múltiples lugares en el fragmento largo.

Para una secuenciación clínica precisa y haplotipificación de genomas humanos individuales de un pequeño número de células, son preferibles los fragmentos genómicos largos (-100 kb o más), aunque pueden usarse fragmentos más cortos. Suponiendo fragmentos de 100 kb, un genoma humano tendría aproximadamente $6x10^4$ fragmentos por célula, y ~18 generarían 1 las células un millón de fragmentos. Las etiquetas de ADN que tienen 12 bases de largo (12 mers) o más tienen suficiente diversidad de secuencia (de 16 millones a más de mil millones) para etiquetar cada fragmento con una etiqueta única.

Se proporcionan varios métodos ilustrativos para asociar copias de la misma etiqueta larga a cientos de subregiones de ~1 kb de fragmentos genómicos de ~100 kb en una reacción homogénea sin compartimentación física (por ejemplo, gotitas en una emulsión). Se reconocerá que el MT no se limita a estos métodos particulares.

En algunas realizaciones de la invención, tales métodos conducen a una mayoría (por ejemplo, 50% o 60%, 70%, 80%, 90% o más de los fragmentos largos de un ácido nucleico objetivo que se etiqueta con múltiples secuencias que contienen una etiqueta que incluyen la misma secuencia de etiqueta. Dichos métodos minimizan el etiquetado con diferentes secuencias de etiquetas, por ejemplo: seleccionar la proporción adecuada de secuencias que contienen una etiqueta con respecto a fragmentos largos; seleccionar la dilución o concentración de ADN adecuada; minimizar el movimiento de la molécula después del inicio del proceso de etiquetado, por ejemplo, mezclando fragmentos de ADN, secuencias que contienen una etiqueta y enzimas y tampones a baja temperatura, esperar que se detengan los movimientos del líquido y luego aumentar la temperatura de la mezcla para activar procesos enzimáticos); atar una secuencia que contiene una sola etiqueta a un solo fragmento largo de ADN mediante unión covalente o no covalente; y otras técnicas. Hay varias maneras de unir o atar una sola perla o nanobola con múltiples copias de una secuencia que contiene una etiqueta particular a un solo fragmento largo de un ácido nucleico objetivo. Por ejemplo, se puede agregar una secuencia de homopolímero (por ejemplo, y cola A) al fragmento largo usando una transferasa terminal o se puede ligar un adaptador con una secuencia seleccionada a un extremo o extremos del fragmento largo. Se puede agregar una secuencia complementaria al final o incluirse dentro de la secuencia que contiene una etiqueta o nanobola de tal manera que, en condiciones apropiadas seleccionadas, la secuencia que contiene una etiqueta o el ensamblaje de la etiqueta se hibrida con la secuencia complementaria correspondiente en el fragmento largo. Preferiblemente, un fragmento largo puede hibridar con solo una secuencia que contiene una etiqueta o un conjunto de etiquetas.

El MT evita la subclonación de fragmentos de un ácido nucleico complejo en un vector y la replicación posterior en una célula huésped, o la necesidad de aislar cromosomas individuales (por ejemplo, cromosomas de metafase). Tampoco requiere fragmentos de alícuotas de un ácido nucleico objetivo. El MT puede ser totalmente automatizado, lo que lo hace adecuado para aplicaciones rentables y de alto rendimiento. El etiquetado de subregiones de ~1 kb de

fragmentos genómicos largos (~100 kb o más) con la misma etiqueta única tiene muchas aplicaciones, que incluyen haplotipificación de genomas diploides o poliploides, ensamblaje eficiente de la secuencia del genoma *de novo*, resolución de repeticiones genómicas, llamadas de variantes precisas y corrección de error.

Las ventajas del MT incluyen:

- Un número prácticamente ilimitado de fragmentos de ADN individuales se puede etiquetar de forma única, proporcionando información máxima para el ensamblaje *de novo*, por ejemplo.

- El MT puede realizarse en un único recipiente de reacción (por ejemplo, tubo, pozo en una placa de múltiples pozos, etc.) en un pequeño número de etapas y es fácil de escalar y automatizar; no hay necesidad de una gran cantidad de alícuotas o nanogotas.

- Un método de MT, que emplea corte y un proceso de ligadura de cebadores, usa ambas cadenas del ADNbc, duplicando así la cobertura de secuencia por fragmento (pares de empalme más largos para la misma longitud de lectura).

- El MT reduce las demandas computacionales y los costos asociados de mapeo y ensamblaje de secuencias.

- Reducción sustancial de errores o llamadas de base cuestionables que pueden resultar de las tecnologías de secuenciación actuales, incluidos, por ejemplo, errores sistemáticos que son característicos de una plataforma de secuenciación dada o mutaciones introducidas por la amplificación de ADN. El MT proporciona una secuencia altamente precisa de un genoma humano u otro ácido nucleico complejo, minimizando la necesidad de confirmación de seguimiento de las variantes detectadas y facilita la adopción de la secuenciación del genoma humano para aplicaciones de diagnóstico.

El MT puede usarse como un método de procesamiento previo con cualquier tecnología de secuenciación conocida, incluidos los métodos de lectura corta y de lectura más larga. Por ejemplo, los subfragmentos etiquetados de 1-10 kb se pueden secuenciar con métodos de molécula única, sin la necesidad de hacer pares de empalme, y se pueden usar en el ensamblaje genómico preciso o la detección de variantes genéticas a pesar de tener una alta tasa de error en las lecturas sin procesar. El MT también se puede usar junto con varios tipos de análisis, incluido, por ejemplo, el análisis del transcriptoma, el metiloma, etc. Debido a que requiere muy poca entrada de ADN, el MT se puede usar para secuenciar y haplotipificar una o una pequeña cantidad de células, que puede ser particularmente útil para el cáncer, el diagnóstico prenatal y la medicina personalizada. Esto puede facilitar la identificación de enfermedades genéticas familiares, etc. Al hacer posible distinguir las llamadas de los dos conjuntos de cromosomas en una muestra diploide, el MT también permite llamadas de mayor confianza de posiciones variantes y no variantes con baja cobertura. Las aplicaciones adicionales de MT incluyen la resolución de reordenamientos extensos en genomas de cáncer y la secuenciación completa de transcripciones empalmadas alternativamente.

El MT puede usarse para procesar y analizar ácidos nucleicos complejos, que incluyen pero no se limitan a ADN genómico, que está purificado o no purificado, que incluye células y tejidos que se rompen suavemente para liberar tales ácidos nucleicos complejos sin cizallamiento y fragmentación excesiva de dichos ácidos nucleicos complejos.

En un aspecto, el MT produce longitudes de lectura virtual de aproximadamente 100-1000 kb o más de longitud, por ejemplo.

Además de ser aplicable a todas las plataformas de secuenciación, la secuenciación basada en MT es adecuada para una amplia variedad de aplicaciones, que incluyen, entre otras, el estudio de reordenamientos estructurales en genomas de cáncer, análisis completo del metiloma que incluyen los haplotipos de sitios metilados y aplicaciones de ensamblaje *de novo* para genomas humanos individuales, metagenómica o secuenciación genómica novedosa, incluso de genomas poliploides complejos como los que se encuentran en las plantas.

El MT proporciona la capacidad de obtener secuencias reales de cromosomas individuales en lugar de solo las secuencias de consenso de los cromosomas parentales o relacionados (a pesar de sus altas similitudes y presencia de repeticiones largas y duplicaciones segmentarias). Para generar este tipo de datos, la continuidad de la secuencia se establece en general en largos intervalos de ADN.

El software y los algoritmos se pueden usar para utilizar eficientemente los datos de MT para el mapeo de haplotipos cromosómicos completos y de variación estructural y la corrección de errores falsos positivos/negativos.

La extensión controlada del cebador y la traducción controlada del corte se pueden usar para aleatorizar los extremos de los fragmentos clonales generados por la amplificación inicial del ADN etiquetado.

Una o más de las siguientes características pueden ser parte del protocolo de MT:

1) Minimizar los casos en los que se inserta más de un código de barras diferente por fragmento largo de ADN, seleccionando las concentraciones apropiadas (es decir, la dilución apropiada) de códigos de barras clonales y fragmentos largos de ADN de modo que menos del 0,1%, menos del 1% o menos del 10% de los fragmentos largos

de ADN están etiquetados con múltiples secuencias de códigos de barra. Las diluciones óptimas dependerán de factores como la cantidad de ADN objetivo disponible. Por ejemplo, se puede usar un exceso de fragmentos largos de ADN sobre códigos de barras cuando el ADN no es limitante (por ejemplo, usando una muestra de sangre o saliva), mientras que se puede usar un exceso de códigos de barras sobre el ADN cuando el material de partida se limita a unas pocas células y es deseable etiquetar cada fragmento. Cuando el ADN no es limitante (por ejemplo, más de 20, más de 50, más de 100 o más de 500 equivalentes de genoma en la mezcla de reacción) no es necesario etiquetar de manera óptima cada fragmento individual. En cambio, puede ser ventajoso sacrificar algo de rendimiento para minimizar el etiquetado de un fragmento de ADN con diferentes códigos de barras. En un enfoque, los fragmentos de ADN están presentes en exceso en relación con los códigos de barras clonales. El uso de un exceso de fragmentos de ADN sobre el número de códigos de barras clonales diferentes aumenta la probabilidad de que solo uno o unos pocos fragmentos de ADN estén cerca de cualquier código de barras para permitir el etiquetado. También permite tener más espacio entre los códigos de barras clonales para minimizar tener dos códigos de barras diferentes por fragmento de ADN. Se puede usar un exceso de aproximadamente 3, 10, 30, 100 o incluso 300 veces en diferentes configuraciones de reacción. Se pueden usar más de 10.000, más de 100.000 o más de 1.000.000 de códigos de barras diferentes.

2) Se puede usar un medio similar a un gel (por ejemplo, bloques de gel de agarosa de bajo punto de fusión u otros polímeros como PEG) para minimizar el movimiento del líquido, limitando la mezcla e interacciones de diferentes códigos de barras clonales y diferentes moléculas largas de ADN.

3) Usar ADN previamente separado, en el que, antes de mezclar ADN con códigos de barras clonales, se introducen espacios en el ADN, por ejemplo, por transposón y listo para la ligadura del código de barras. En este enfoque, se prepara el ADN con huecos (por ejemplo, cortando y separando), seguido de la adición de códigos de barras y una etapa de etiquetado (ligadura) realizada en el ADN previamente separado. Este enfoque reduce la complejidad enzimática de la mezcla de reacción.

4) Ligadura en el espacio del código de barras del adaptador al extremo 3’ del ADN sin complementariedad con el ADN objetivo o al extremo 5' usando 2-8 bases degeneradas; y/o

5) Separar ADN largo, liberando las copias individuales del código de barras y la unión de los códigos de barras al ADN se realiza como una reacción (es decir, todos necesitan enzimas, códigos de barras clonales y ADN largo presente en la mezcla antes de comenzar la incubación). Alternativamente, solo la liberación y unión del código de barras se realiza como una reacción; la separación del ADN se realiza como una etapa anterior.

6) Cuando se usan transposones de copiar y pegar, la concentración de ADN debe ser apropiadamente baja para minimizar los transposones de "salto" entre fragmentos largos de ADN. Una molécula de transposón de copiar y pegar con una secuencia de código de barras proporciona códigos de barras clonales dentro de un fragmento largo de ADN y no los otros fragmentos de ADN si dichos fragmentos de ADN están lo suficientemente separados como para evitar que el transposón salte de un ADN a otro.

En un enfoque, se etiqueta una pequeña cantidad de ADN (por ejemplo, de 10 células) en un único recipiente. Las etiquetas clonales interactúan con el ADN en un volumen pequeño para dar una alta probabilidad de que casi todos los fragmentos largos de ADN encuentren un código de barras clonal (una perla con un oligonucleótido de código de barras adaptador amplificado clonalmente). Los códigos de barras clonales y los portadores asociados proporcionan un exceso de capacidad de unión al ADN. La capacidad de unión al ADN se define por el número de perlas u otro portador y el número de fragmentos de ADN que pueden unirse por perla u otro portador. Para ilustración, los genomas de 10 células humanas son equivalentes a 1 millón de fragmentos de 60 kb. Obsérvese que incluso en el caso de tener un exceso de ADN y la necesidad de etiquetar 10 millones de fragmentos, si el portador puede unir múltiples fragmentos de ADN, entonces $\sim 10^6$ etiquetas clonales pueden ser suficientes.

Casi todos los fragmentos largos de ADN de una cantidad de ADN limitada (por ejemplo, 3-30 células humanas) pueden etiquetarse con múltiples copias del mismo código de barras por fragmento largo de ADN. Esto se puede hacer: (a) proporcionando códigos de barras clonales > 10 K o 100 K o 1 M y una pequeña cantidad de fragmentos largos de ADN, en los que la capacidad total de unión al ADN de las "entidades de código de barras clonales" excede la cantidad proporcionada de ADN; (b) asociar casi todo el fragmento o fragmentos largos de ADN en alta concentración al portador de los códigos de barras clonales o entidades de códigos de barras clonales modificados (para ciertas aplicaciones, cada entidad de código de barras clonal tiene una capacidad de unión al ADN limitada de < 100 kb, < 300 kb, menos de 1 Mb); (c) diluir o espaciar entidades de códigos de barras clonales antes de cortar/liberar códigos de barras y etiquetar el ADN (para minimizar tener más de un código de barras distinto por ADN); (d) etiquetar fragmentos largos de ADN (a una distancia promedio predefinida) con copias de códigos de barras de la entidad de código de barras clonal asociada. En algunas realizaciones, se prefieren, lecturas de secuencia más largas, tales como 2 x 100-300 bases o subfragmentos enteros de 1-3 kb, de modo que se leen más bases por cada fragmento largo.

El uso de exceso de partículas de código de barras clonal o el exceso de capacidad de unión total puede ayudar a garantizar que se usen casi todos los fragmentos de ADN y que (i) es raro que más de un fragmento esté unido al mismo portador de código de barras clonal o (ii) un número promedio predefinido de fragmentos está unido a la misma entidad de código de barras clonal. Los códigos de barras clonales o sintéticos se pueden unir en una superficie de

manera que no haya dilución después de la unión del ADN a los puntos del código de barras. Los códigos de barras clonales espaciados unidos a la superficie se pueden organizar de la siguiente manera: un chip de 1 cm^2 con 10^6 etiquetas distintas unidas a puntos que tienen un tamaño de ~ 0,5-2 µm y espaciados a 10 µm, con una capacidad de unión total > 66 pg de ADN, preferiblemente > 100 pg, más de 300 pg, o > 1 ng. El ADN se carga con 10 células en un volumen de 5-10 µL. Con un tiempo de incubación adecuado y una mezcla opcional, la mayoría de los fragmentos de ADN pueden unirse a puntos de código de barras clonal. Los códigos de barras en dicho chip se pueden preparar por síntesis de oligo en lugar de por proceso de clonación. La superficie está configurada de modo que el ADN largo no se adhiera a la superficie entre puntos.

Los productos adecuados para el código de barras de ADN avanzado incluyen una biblioteca de ADN con código de barras (opcionalmente preparada en una reacción única) que comprende códigos de barras > 10 k,> 100 k,> 1 M o > 10 M y en promedio > 15%, > 20%, > 25%, > 30%, 40% o > 50% de la secuencia de un fragmento largo de ADN se representan en fragmentos de código de barras (opcionalmente, los fragmentos largos de ADN no se amplifican). El fragmento largo se representa en fragmentos cortos de ADN etiquetados con las copias del mismo código de barras. Los fragmentos de ADN etiquetados pueden amplificarse y, opcionalmente, uno o ambos extremos del fragmento son aleatorios. Se utilizan tan solo 100, 50, 30, 20 o 10 células o menos para hacer la biblioteca.

Preparación de fragmentos largos de ácido nucleico

Los ácidos nucleicos objetivo, que incluyen pero no se limitan a ácidos nucleicos complejos, pueden aislarse usando técnicas convencionales, por ejemplo, como se describe en Sambrook y Russell, Molecular Cloning: A Laboratory Manual, citado anteriormente. En algunos casos, particularmente si se emplean pequeñas cantidades de los ácidos nucleicos en una etapa particular, es ventajoso proporcionar ADN portador, por ejemplo, ADN de cadena doble sintético circular no relacionado, para mezclar y usar con los ácidos nucleicos de la muestra siempre que solo pequeñas cantidades de ácidos nucleicos de muestra se encuentren disponibles y exista el peligro de pérdidas a través de la unión no específica, por ejemplo, a las paredes del recipiente y similares.

De acuerdo con algunas realizaciones de la invención, el ADN genómico u otros ácidos nucleicos complejos se obtienen de una célula individual o un pequeño número de células con o sin purificación, mediante cualquier método conocido.

Los fragmentos largos son deseables para los métodos de la presente invención. Se pueden aislar fragmentos largos de ADN genómico de una célula mediante cualquier método conocido. Un protocolo para el aislamiento de fragmentos largos de ADN genómico de células humanas se describe, por ejemplo, en Peters et al., Nature 487: 190-195 (2012). En una realización, las células se lisan y los núcleos intactos se sedimentan con una etapa de centrifugación suave. El ADN genómico se libera a través de la digestión de proteinasa K y RNasa durante varias horas. El material puede tratarse para reducir la concentración de residuos celulares restantes, por ejemplo, por diálisis durante un período de tiempo (es decir, de 2 a 16 horas) y/o dilución. Dado que tales métodos no necesitan emplear muchos procesos disruptivos (como la precipitación con etanol, la centrifugación y la agitación vorticial), el ácido nucleico genómico permanece en gran parte intacto, produciendo una mayoría de fragmentos que tienen longitudes superiores a 150 kilobases. Los fragmentos son de aproximadamente 5 a aproximadamente 750 kilobases de longitud. En algunas realizaciones, los fragmentos tienen una longitud de aproximadamente 150 a aproximadamente 600, aproximadamente 200 a aproximadamente 500, aproximadamente 250 a aproximadamente 400 y aproximadamente 300 a aproximadamente 350 kilobases. El fragmento más pequeño que puede usarse para haplotipificación es uno que contiene al menos dos hets (aproximadamente 2-5 kb); no existe un tamaño teórico máximo, aunque la longitud del fragmento puede estar limitada por el cizallamiento resultante de la manipulación de la preparación de ácido nucleico inicial.

En otras realizaciones, se aíslan y manipulan fragmentos largos de ADN de una manera que minimiza el cizallamiento o la absorción del ADN en un recipiente, que incluye, por ejemplo, aislar células en agarosa en tapones de gel de agarosa, o aceite, o usando tubos y placas especialmente recubiertos.

El uso controlado de una exonucleasa 5' (ya sea antes o durante la amplificación) puede promover múltiples replicaciones del ADN original de una sola célula y minimizar así la propagación de errores tempranos mediante el copiado de copias.

El ADN fragmentado de una sola célula puede duplicarse ligando un adaptador con un saliente de cebado de cadena sencilla y usando un cebador específico de adaptador y polimerasa phi29 para hacer dos copias de cada fragmento largo. Esto puede generar cuatro células de ADN a partir de una sola célula.

De acuerdo con una realización de la invención, se comienza con más fragmentos largos de los necesarios para la secuenciación para lograr una cobertura de secuencia adecuada y etiquetar solo una parte de los fragmentos largos con un número limitado de secuencias que contienen una etiqueta, o conjuntos de etiquetas que incluyen muchas, quizás cientos, de copias de una secuencia de etiqueta, para aumentar la probabilidad de etiquetado único de los fragmentos largos. Subfragmentos no etiquetados que carecen de secuencias introducidas que proporcionan unión de cebador o unión de oligo de captura y pueden eliminarse en el procesamiento posterior. Dichos conjuntos de etiquetas incluyen perlas a las que se unen muchas copias de las secuencias que contienen una etiqueta.

De acuerdo con otra realización, para obtener una cobertura uniforme del genoma en el caso de muestras con un pequeño número de células (por ejemplo, 1, 2, 3, 4, 5, 10, 10, 15, 20, 30, 40 , 50 o 100 células de una micro biopsia o tumor circulante o células fetales, por ejemplo), todos los fragmentos largos obtenidos de las células están etiquetados.

Preservación de los extremos del fragmento

Una vez que se aísla el ADN, es ventajoso evitar la pérdida de secuencias de los extremos de cada fragmento, ya que la pérdida de dicho material puede dar como resultado huecos en el ensamblaje final del genoma. En una realización, la pérdida de secuencia se evita mediante el uso de una enzima de corte poco frecuente, que crea sitios de partida para una polimerasa, como la polimerasa phi29, a distancias de aproximadamente 100 kb entre sí. A medida que la polimerasa crea una nueva cadena de ADN, desplaza la cadena anterior, creando secuencias superpuestas cerca de los sitios de inicio de la polimerasa. Como resultado, hay muy pocas eliminaciones de secuencia.

El MT puede llevarse a cabo utilizando códigos de barras clonales, incluidos códigos de barras clonales elaborados sintéticamente

Los términos "código de barras", "etiqueta", "secuencia de código de barras", "secuencia de etiqueta" y variaciones obvias de estos se usan indistintamente, tienen el significado normal en la técnica y se refieren generalmente a un nombre identificable (generalmente único) o secuencia de nucleótidos, heteróloga a la secuencia objetivo. En una población o biblioteca de etiquetas, los códigos de barras únicos a veces se asocian con secuencias de adaptador comunes, en uno o ambos lados del código de barras, que pueden ser compartidos por muchos o todos los miembros de la población o biblioteca.

"Etiquetado" se refiere a asociar (por ejemplo, insertar) una secuencia de etiqueta con un polinucleótido. Etiquetar fragmentos largos implica introducir en fragmentos largos múltiples copias de secuencias (adaptadores, transposones, etc.) que incluyen etiquetas. Tales "secuencias introducidas" están separadas en el fragmento. Típicamente, el espaciado promedio entre secuencias introducidas adyacentes se selecciona para permitir la creación de subfragmentos que contienen una etiqueta de los fragmentos largos. Los subfragmentos se pueden elaborar por cualquier método adecuado, por ejemplo, por amplificación por PCR usando cebadores que tienen sitios de unión de cebadores en secuencias introducidas adyacentes; por digestión de restricción; o por otros métodos conocidos en la técnica. Posteriormente, las lecturas de secuencia se generan secuenciando subfragmentos de los fragmentos largos etiquetados. Dichas lecturas de secuencia pueden asignarse al fragmento largo individual del que derivan en última instancia.

En algunas realizaciones de MT, se usa una fuente de etiquetas clonales o códigos de barras. Por "clonal" se entiende etiquetas o códigos de barras que contienen (es decir, comprenden) la misma secuencia y se asocian físicamente entre sí (en lugar de separarse y pueden difundirse libremente en la solución) de modo que una fuente de etiquetas clonales se pueda asociar con un solo fragmento de ADN largo. El resultado es que una pluralidad de etiquetas clonales identificables o códigos de barras pueden estar asociados con un fragmento de ADN y no con otros. Las etiquetas clonales o códigos de barras pueden mantenerse juntas en un portador tal como microperlas. Los términos "partícula" y "fuente de códigos de barras clonales" también se usan en esta divulgación para referirse a un sistema de suministro para copias múltiples, de una secuencia de etiqueta.

Otro ejemplo de una fuente de códigos de barras clonales es una partícula (por ejemplo, una perla u otra estructura de soporte) con una pluralidad de oligonucleótidos inmovilizados sobre ella. En un enfoque, los oligonucleótidos se unen covalentemente al soporte, por ejemplo, mediante un enlazador escindible. En otro enfoque, los oligonucleótidos están unidos no covalentemente al soporte. Los oligonucleótidos pueden liberarse del soporte usando cualquier método adecuado, tal como tratamiento con una enzima de restricción que liberó un fragmento del oligonucleótido unido. Alternativamente, se puede escindir un enlazador. En un enfoque, un enlazador puede ser ácido nucleico con bases modificadas, como el uracilo, que se puede escindir enzimática o químicamente. Se puede usar cualquiera de varios métodos de disociación de oligonucleótidos.

También se contempla el uso de fuentes con un pequeño número (por ejemplo, 2 o 3) de diferentes secuencias de etiquetas, tales como una perla asociada con una "secuencia a" y una "secuencia b". En este caso, se reconocería que la "secuencia a" y la "secuencia b" se insertarán en el mismo fragmento largo, y la "secuencia a" y la "secuencia b" se tratarán como equivalentes en el proceso de ensamblaje de secuencia.

Elaboración de fuentes de etiquetas clonales o códigos de barras

Secuencias de etiquetas/transposones asociadas con perlas u otros soportes

Una fuente de códigos de barras clonales tales como una perla u otro soporte asociado con múltiples copias de etiquetas puede prepararse mediante PCR en emulsión o CPG (vidrio de poro controlado) o síntesis química de otras partículas con copias de un código de barras adaptado preparado para ese fin. Una población de secuencias de ADN que contienen una etiqueta puede amplificarse por PCR en perlas en una emulsión de agua en aceite (w/o) mediante métodos conocidos. Véase, por ejemplo, Tawfik y Griffiths Nature Biotechnology 16: 652-656 (1998); Dressman et al., Proc. Natl. Acad. Sci. USA 100: 8817-8820, 2003; y Shendure et al., Science 309: 1728-1732 (2005). Esto da como resultado muchas copias de cada secuencia que contiene una sola etiqueta en cada perla.

Otro método para elaborar una fuente de códigos de barras clonales es mediante síntesis de oligonucleótidos en microperlas o CPG en un proceso combinatorio de "mezclar y dividir". Usando este proceso, se puede crear un conjunto de perlas cada una con una población de copias de un código de barras. Por ejemplo, para elaborar todo el $B_{20}Ni_5B_{20}$ en el que cada uno de aproximadamente 1.000 millones está representado en -1000+ copias en cada una de 100 perlas, en promedio, se puede comenzar con ~100.000 millones de perlas, sintetizar la secuencia común B_{20} (adaptador) en todas ellas y luego dividirlas en 1024 columnas de síntesis para elaborar un 5 mer diferente en cada uno, luego mézclelos y luego divídalos nuevamente en 1024 columnas y haga 5-mer adicionales, y luego se repite eso una vez más para completar N15, y luego se mezclan y en una gran columna, se sintetiza el último B_{20} como un segundo adaptador. Por lo tanto, en las síntesis de 3050 se pueden elaborar los mismos conjuntos de códigos de barras "de tipo clonal" que en una gran reacción PCR de emulación con -1.000 millones de perlas (10^{12} perlas) porque solo 1 de cada 10 perlas tendrá una plantilla inicial (las otras 9 no tendrían ninguna) para evitar tener dos plantillas con diferentes códigos de barras por perla.

Características de las etiquetas

De acuerdo con una realización, se usa una secuencia que contiene código de barras o etiqueta que tiene dos, tres o más segmentos de los cuales, uno, por ejemplo, es la secuencia de código de barras. Por ejemplo, una secuencia introducida puede incluir una o más regiones de secuencia conocida y una o más regiones de secuencia degenerada que sirven como código o códigos de barras o etiquetas. La secuencia conocida (B) puede incluir, por ejemplo, sitios de unión del cebador de PCR, extremos del transposón, secuencias de reconocimiento de endonucleasas de restricción (por ejemplo, sitios para cortadores raros, por ejemplo, Not I, Sac II, Mlu I, BssH II, etc.), u otras secuencias. La secuencia degenerada (N) que sirve como etiqueta es lo suficientemente larga como para proporcionar una población de etiquetas de secuencia diferente que es igual o, preferiblemente, mayor que el número de fragmentos de un ácido nucleico objetivo a analizar.

De acuerdo con una realización, la secuencia que contiene una etiqueta comprende una región de secuencia conocida de cualquier longitud seleccionada. De acuerdo con otra realización, la secuencia que contiene una etiqueta comprende dos regiones de secuencia conocida de una longitud seleccionada que flanquean una región de secuencia degenerada de una longitud seleccionada, es decir, $B_nN_nB_n$, en la que N puede tener cualquier longitud suficiente para etiquetar fragmentos largos de un nucleico objetivo ácido, que incluye, sin limitación, N = 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 o 20, y B puede tener cualquier longitud que acomode secuencias deseadas tales como extremos de transposones, sitios de unión de cebadores, etc. Por ejemplo, tal realización puede ser $B_{20}N_{15}B_{20}$.

En una realización, se utiliza un diseño de dos o tres segmentos para los códigos de barras utilizados para etiquetar fragmentos largos. Este diseño permite una gama más amplia de posibles códigos de barras al permitir que se generen segmentos combinatorios de códigos de barras al ligar diferentes segmentos de códigos de barras para formar el segmento completo de código de barras o al usar un segmento como reactivo en la síntesis de oligonucleótidos. Este diseño combinatorio proporciona un repertorio más grande de posibles códigos de barras al tiempo que reduce la cantidad de códigos de barras de tamaño completo que deben generarse. En realizaciones adicionales, la identificación única de cada fragmento largo se logra con códigos de barras de 8-12 pares de bases (o más largos).

En una realización, se usan dos segmentos de código de barras diferentes. Los segmentos A y B se pueden modificar fácilmente para que cada uno contenga la mitad de una secuencia de código de barras diferente para producir miles de combinaciones. En una realización adicional, las secuencias de código de barras se incorporan en el mismo adaptador. Esto se puede lograr dividiendo el adaptador B en dos partes, cada una con la mitad de la secuencia de código de barras separada por una secuencia de superposición común utilizada para la ligadura. Los dos componentes de la etiqueta tienen 4-6 bases cada uno. Un conjunto de etiquetas de 8 bases (2 x 4 bases) es capaz de etiquetar únicamente 65.000 secuencias. Tanto las etiquetas de 2 x 5 bases como las de 2 x 6 bases pueden incluir el uso de bases degeneradas (es decir, "comodines") para lograr una eficiencia de decodificación óptima.

En realizaciones adicionales, la identificación única de cada secuencia se logra con códigos de barras de corrección de errores de 8-12 pares de bases. Los códigos de barras pueden tener una longitud, para ilustración y no limitativa, de 5 a 20 bases informativas, generalmente de 8 a 16 bases informativas.

Etiquetar fragmentos largos individuales

Los métodos de la presente invención emplean diversos enfoques para introducir múltiples copias de una etiqueta en múltiples sitios separados a lo largo de un fragmento largo (por ejemplo, 100 kb o más) del ácido nucleico objetivo sin la necesidad de dividir los fragmentos largos en alícuotas (como en la tecnología de lectura de fragmentos largos): todo el proceso puede realizarse en un solo tubo o en una placa de microtitulación.

De acuerdo con una realización de la invención, las etiquetas se introducen a intervalos de entre aproximadamente 300 pb y 1000 pb a lo largo del fragmento. Este espacio puede ser más corto o más largo, dependiendo del tamaño de fragmento deseado para el procesamiento posterior, por ejemplo, la construcción y secuenciación de la biblioteca. Después del etiquetado, cada subfragmento del fragmento largo y cualquier información de secuencia derivada de él se puede asignar a un solo fragmento largo.

Fragmentos largos que contienen la misma etiqueta o código de barras

En algunas realizaciones de la invención, tales métodos dan como resultado que la mayoría (por ejemplo, 50%, 60%, 70%, 80%, 90% o más) de los fragmentos largos de un ácido nucleico objetivo se etiqueten con múltiples secuencias que contienen etiquetas que incluyen la misma secuencia de etiqueta. Se pueden tomar medidas para minimizar el etiquetado con diferentes secuencias de etiqueta, por ejemplo: seleccionar la relación adecuada de secuencias que contienen una etiqueta con respecto a los fragmentos largos; seleccionar la dilución o concentración de ADN adecuada; minimizar el movimiento de la molécula después del inicio del proceso de etiquetado, por ejemplo, mezclando fragmentos de ADN, secuencias que contienen una etiqueta y enzimas y tampones a baja temperatura, esperar que se detengan los movimientos del líquido y luego aumentar la temperatura de la mezcla para activar los procesos enzimáticos); atar una secuencia que contiene una sola etiqueta a un solo fragmento largo de ADN mediante unión covalente o no covalente; y otras técnicas

Fragmentos largos que contienen una huella digital única

En otras realizaciones de la invención, en lugar de maximizar el número de fragmentos largos con una única secuencia de etiqueta insertada en múltiples ubicaciones a lo largo del fragmento largo, el MT implica proporcionar condiciones bajo las cuales se insertan múltiples etiquetas con diferentes secuencias en múltiples ubicaciones, creando un patrón único o "huella digital" para cada fragmento largo que es proporcionado por un patrón único de inserción de las etiquetas de secuencia diferente.

A continuación se describen ejemplos de métodos para etiquetar fragmentos individuales.

(1) Etiquetado con transposones

Varios enfoques para MT hacen uso de secuencias de transposones y/o transposasas. Se puede usar cualquier sistema adecuado de transposón/transposasa o de transposón/integrasa para introducir transposones etiquetados. Los ejemplos incluyen *in vitro* la transposición de Mu (Haapa et al., Nucl. Acids Res., 27: 2777-2784, 1999; Savilahti et al., EMBO J. 14: 4893-4903, 1995); Tyl (Devine y Boeke, Nucl. Acids Res., 22: 3765-3772, 1994; solicitud internacional de patente WO 95/23875); Tn7 (Craig, Curr. Topics Microbiol. Immunol. 204: 27-48, 1996); Tn 10 e IS 10 (Kleckner et al., Curr. Top. Microbiol. Immunol. 204: 49-82, 1996); Mariner (Lampe et al., EMBO J. 15: 5470-5479, 1996); Tcl (Vos et al., Genes Dev., 10: 755-761, 1996); Tn5 (Park et al., Taehan Misaengmul Hakhoechi 27: 381-389, 1992); Elemento P (Kaufman y Rio, Cell 69: 27-39, 1992); Tn3 (Ichikawa y Ohtsubo, J. Biol. Chem. 265: 18829-18832, 1990); secuencias de inserción bacteriana (Ohtsubo y Sekine, Curr. Top. Microbiol. Immunol., 204: 1-26, 1996); retrovirus (Varmus y Brown, "Retroviruses", en Mobile DNA, Berg y Howe, eds., American Society for Microbiology, Washington, DC, páginas. 53-108, 1989); y retrotransposones de levadura (Boeke, "Transposable elements in Saccharomyces cerevisiae", en Mobile DNA, Berg y Howe, eds., American Society for Microbiology, Washington, DC, páginas 53-108, 1989). Otros transposones conocidos incluyen, sin limitación, AC7, Tn5SEQ1, Tn916, Tn951, Tn1721, Tn 2410, Tn1681, Tn1, Tn2, Tn4, Tn6, Tn9, Tn30, Tn101, Tn903, Tn501, Tn1000 (γ6), Tn1681, Tn2901, transposones AC, transposones Mp, transposones Spm, transposones En, transposones Punteados, transposones Ds, transposones dSpm y transposones I. Se pueden usar formas modificadas de los extremos del transposón y/o transposasas, por ejemplo, una transposasa Tn5 modificada como en la tecnología Nextera^{MR} (Epicentre Biotechnologies, Madison, WI).

Muchas transposasas reconocen diferentes secuencias de inserción, y por lo tanto debe entenderse que un vector basado en transposasa contendrá secuencias de inserción reconocidas por la transposasa particular también encontrada en el vector basado en transposasa. Las transposasas y las secuencias de inserción de vectores basados en transposones eucariotas pueden incluso modificarse y usarse. Sin embargo, los elementos basados en transposones no eucariotas reducen la probabilidad de que una transposasa eucariota en el organismo receptor (por ejemplo, sujeto humano) reconozca secuencias de inserción procariotas que agrupan el transgén.

Un primer enfoque implica la transposición *in vitro* (véanse las Figuras 1A y 1B). Se utiliza una población de transposones etiquetados **21a, 21b, 21c, 21d**. Los transposones etiquetados son constructos de ADN que incluyen los extremos del transposón **24**, y cerca de cada uno de los extremos, pares de secuencias únicas de etiqueta (código de barras) **22a**, **22b**, **22c**, **22d** (la misma secuencia de etiqueta cerca de ambos extremos) y un sitio **23** de unión de cebador de PCR común. La población de transposones se combina con fragmentos largos **1** de un ácido nucleico objetivo. La adición de transposasa provoca la transposición *in vitro* de varios de los transposones etiquetados en los fragmentos largos **2**. Cada fragmento largo tiene un patrón único de inserción de transposón, y cada transposón insertado tiene una secuencia de una sola etiqueta (código de barras). Además, el acto de transposición replica 9 pb de secuencia en cada extremo del transposón que distingue aún más cada evento de inserción de transposón (y puede considerarse otra forma de "etiquetado").

La PCR se realiza usando cebadores que se unen a los sitios **23** de unión del cebador de PCR de cada transposón insertado. Los amplicones **3** de PCR resultantes incluyen una porción del fragmento largo **31a, 31b, 31c, 31d** que se encuentra entre las porciones de transposón adyacente **32a**, **32b**, **32c**, **32d**. En cada extremo de los amplicones (es decir, que flanquean la secuencia objetivo o de fragmento largo) hay secuencias desde el extremo de un transposón adyacente, incluida la secuencia de una sola etiqueta (código de barras) para ese transposón **22a, 22b, 22c, 22d**.

Después de secuenciar los amplicones de PCR, es posible no solo mapear las lecturas de secuencia a un genoma de

referencia, suponiendo que esté disponible, sino usar las etiquetas para construir cóntigos para guiar el ensamblaje *de novo*. Cada lectura de secuencia **42a** a **42h** está asociada con una secuencia de etiqueta **22a**, **22b**, **22c**, **22d**. Una secuencia de etiqueta particular (o patrón de etiquetas, por ejemplo, un par de etiquetas o una cadena de etiquetas) corresponde a un solo fragmento. Por lo tanto, las lecturas de secuencia del mismo fragmento deben mapearse dentro de la misma región del ácido nucleico objetivo. En general, dos amplicones diferentes (como **31a** y **31b**) tienen la misma etiqueta única **22b** de un transposón en sus extremos y, por lo tanto, son adyacentes entre sí en el fragmento largo del que derivan.

Las lecturas de secuencia se ensamblan usando códigos de barras adyacentes coincidentes para construir lecturas largas, cada una de las cuales comprende una secuencia de etiqueta **21a** a **21d** junto con una secuencia de la parte **42a**, **42b**, **42c**, **42d**, **42e** del fragmento inicial. Las lecturas de secuencia son continuas o discontinuas dependiendo de la longitud de lectura de secuencia. Si se analiza más de un equivalente genómico de fragmentos largos (por ejemplo, 2, 3, 4, 5, 10 o 20 o más equivalentes genómicos), la construcción de cóntigos sin lecturas de secuencia derivadas de fragmentos largos superpuestos es sencilla.

(2) Etiquetado con horquillas

Este enfoque comienza con fragmentos largos de un ácido nucleico objetivo 1 que se desnaturalizan para formar dos cadenas simples complementarias de cada fragmento **11**, **12**. Véanse las Figuras 2A y 2B. También utiliza una población de oligonucleótidos (**25a**, **25b**, **25c**, **25d**) que forman horquillas, cada una de las cuales incluye secuencias de etiquetas **22a**, **22b**, **22c**, **22d** en el bucle que flanquea los sitios de unión del cebador de PCR y tienen un corto tramo de bases aleatorias (por ejemplo, 3-5 bases) **26a** a **26h** en cada extremo. Los oligos de horquilla se hibridan **2a** con la forma de cadena sencilla **11** de los fragmentos largos iniciales separados, por ejemplo, en aproximadamente 300 a 1000 pb. Cada fragmento largo tiene un patrón único de horquillas hibridadas. Después de la hibridación, la región de cadena sencilla entre las horquillas adyacentes se rellena **2b** con una polimerasa 5'-3' que carece de desplazamiento de cadena, seguido de un tratamiento con ligasa para sellar el corte **2c** restante.

La amplificación **3** por PCR usando cebadores que se unen a los sitios **23** de unión de PCR entre las secuencias de código de barras de cada horquilla crea amplicones que tienen una porción del fragmento largo **31a**, **31b**, **31c**, **31d** que se encuentra entre los sitios de unión de los oligonucleótidos de la horquilla adyacente. En cada extremo, tales amplicones incluyen secuencias del bucle de un oligonucleótido de horquilla adyacente, que incluye la secuencia de una sola etiqueta para ese oligonucleótido **22a**, **22b**, **22c**, **22d**. De la misma manera que el método (1) anterior, las secuencias de código de barras en los extremos de los amplicones de PCR se pueden usar para construir cóntigos **4** para guiar el mapeo y ensamblaje *de novo*.

(3) Etiquetado con transposones en una nanobola o perla de ADN

Este enfoque, y varios otros discutidos en el presente documento, usan una partícula (tal como una perla) que contiene muchas copias de la misma secuencia de etiqueta. En algunos casos, las etiquetas incluyen secuencias de transposones. La asociación de polinucleótidos con perlas es bien conocida en la técnica, y se describió brevemente anteriormente.

El enfoque ilustrado en La Figura 3 emplea una partícula **15** de la que pueden liberarse secuencias de transposón: por ejemplo, perlas cubiertas con secuencias de transposón. Como en el método (1) anterior, las "secuencias de transposón" son constructos de ADN que incluyen (i) extremos de transposones **24** y, (ii) en una ubicación seleccionada en las secuencias de transposón **31a** a **31e** entre los extremos del transposón (cerca de cada uno de los extremos del transposón), secuencias de etiquetas **22** (opcionalmente, la misma secuencia de etiqueta puede estar cerca de ambos extremos), y (iii) un sitio **23** de unión al cebador de PCR común.

La perla **15** que contiene transposón se combina con los fragmentos **1** largos de un ácido nucleico objetivo de cadena doble. Las condiciones se seleccionan para promover la interacción de un solo conjunto de etiqueta, es decir, una perla que lleva una secuencia de transposón único, con cada fragmento largo. Por ejemplo, en la dilución correcta, solo una perla **15** interactúa con cada fragmento largo en la mayoría de los casos, ya que la difusión es lenta y la mayoría de los transposones no viajan lejos de un fragmento largo. Alternativamente, la secuencia de transposón u otra secuencia en el ensamblaje de transposón (por ejemplo, un adaptador ligado a un extremo de la secuencia de transposón o concatámero; una secuencia de homopolímero añadida por una transferasa terminal) puede usarse para unir mediante hibridación un oligonucleótido que contiene una secuencia de código de barras que representa una molécula de transposón. Tras la adición de transposasa, se produce la transposición (no mostrada). En la mayoría de los casos, cada fragmento recibió múltiples copias de la misma secuencia de transposón. Una minoría de los fragmentos largos puede recibir copias de más de un transposón. Además, en una minoría de casos, un transposón con una etiqueta particular puede transponerse en más de un fragmento largo.

Como en el método (1), la amplificación por PCR se realiza usando cebadores que se unen a los sitios **23** de unión del cebador de PCR de cada transposón insertado. Los amplicones **3** de PCR resultantes (entre aproximadamente 300 pb y 1000 pb de longitud) incluyen una porción del fragmento largo **31a**, **31b**, **31c**, **31d**, **31e** que se encuentra entre transposones adyacentes; en cada extremo, tales amplicones incluyen secuencias del extremo de un transposón adyacente **32a**, **32b**, que incluye la secuencia de una sola etiqueta (código de barras) para ese transposón **22** en una

o en ambas porciones de transposasa **32a** y **32b**. Los constructos son amplificados y secuenciados. Después de la secuenciación, las lecturas de secuencia **42a**, **42b**, **42c**, **42d**, **42e** se mapean y se ensamblan. El código de barras **22** es una etiqueta para el fragmento **1** largo particular.

En este método, debido a que la mayoría de los fragmentos largos están etiquetados con múltiples copias de un solo transposón, los amplicones resultantes tienen la misma etiqueta en cada extremo. Las etiquetas permiten que cada lectura de secuencia se asocie con el mismo fragmento largo, aunque no es posible construir cóntigos basados en el orden de las secuencias de etiquetas solo como en los métodos (1) y (2). Si más de un transposón se inserta en un solo fragmento largo, lo más probable es que todos los transposones que se insertan en un fragmento largo se inserten solo en ese fragmento largo y no en otros fragmentos. Como resultado, las lecturas de secuencia asociadas con cada una de las etiquetas insertadas se mapean muy juntas en el genoma (u otro ácido nucleico objetivo). Incluso si este no es el caso, y el mismo transposón salta en más de un fragmento, existe una alta probabilidad de que los fragmentos en los que se inserta dicho transposón no se superpongan, en cuyo caso las lecturas de secuencia resultantes se asignan a regiones ampliamente separadas del genoma. El software de mapeo y ensamblaje puede dar cuenta de estos eventos y mapear y ensamblar correctamente las lecturas de secuencia en una secuencia del genoma y ordenar los polimorfismos de secuencia (hets) en un haplotipo.

(4) Etiquetado con adaptadores etiquetados

En este método, los fragmentos 1 largos de cadena doble de un genoma (u otro ácido nucleico objetivo) se cortan en ubicaciones aleatorias en ambas cadenas usando un agente tal como DNasa I que corta las cadenas dobles de ADN (es decir, una "nickasa") y el fragmento grande de ADN polimerasa I (Klenow), que retiene la polimerización y la actividad exonucleasa 3'→ 5', pero ha perdido la actividad exonucleasa 5'→ 3'. Véase la Figura 4A. No se incluyen dNTP en la reacción. Un adaptador **27** común 3' está ligado al extremo 3' de cada cadena en un corte. Se agrega una partícula **15** (tal como una perla o ADN) con muchas copias de una secuencia (por ejemplo, oligonucleótido) que incluye (i) la secuencia **22** de etiqueta y (ii) se agrega una secuencia **28** que es complementaria del adaptador **27** común 3' bajo condiciones que permiten que el adaptador común 3' hibride con la secuencia complementaria. Por ejemplo, el oligonucleótido puede liberarse de la perla.

Como también se describe en otra parte del presente documento, a la relación adecuada de fragmentos largos con respecto a perlas y a la dilución adecuada, la mayoría de los fragmentos están espacialmente asociados con una (o con menos frecuencia 2 o más) perlas, y copias del adaptador común 3' hibridan con la secuencia complementaria en una sola perla, ya que un solo evento de hibridación conduce a una interacción física entre el fragmento largo y la perla, lo que lleva a otras secuencias complementarias a una proximidad cercana. En otras palabras, una partícula está cerca al fragmento de ADNbc, por lo que la mayoría de las copias de etiquetas no se difunden, sino que se hibridan con el adaptador 3'.

Alternativamente, y como también se describe a continuación, los fragmentos de ADNbc largos y las perlas se pueden etiquetar en un extremo para forzar la interacción si es necesario. Por ejemplo, se pueden usar secuencias de ADN complementarias como una cola A en el fragmento largo y una región de cola T o poli-T en las secuencias que contienen una etiqueta, u otras fracciones interactivas, para forzar la interacción del fragmento largo y secuencias que contienen una etiqueta para aumentar la probabilidad de que cada fragmento largo haya introducido múltiples copias de una sola secuencia que contiene etiquetas. A continuación, los ácidos nucleicos que contienen la etiqueta en la perla se fragmentan, por ejemplo, con una endonucleasa de restricción, lo que da como resultado adaptadores comunes ligados al fragmento largo que hibrida con secuencias complementarias que se incluyen en los ácidos nucleicos liberados de la perla. La extensión del cebador utilizando el fragmento grande de ADN polimerasa I (Klenow) o una ADN polimerasa similar da como resultado la creación de una molécula etiquetada en 3' espaciada en el fragmento largo cada 300 - 1000 pb.

La molécula larga de ADN puede luego desnaturalizarse y un oligonucleótido puede hibridarse con el adaptador común 3' ; la extensión con el fragmento Klenow o una polimerasa similar da como resultado una molécula de ADN de cadena doble de extremo romo que puede ligarse a un adaptador común 5' y amplificarse por PCR. Los amplicones de PCR resultantes (subfragmentos efectivamente etiquetados de los fragmentos largos de ADN) se secuencian, mapean y ensamblan de una manera similar a la descrita en el método (3).

Por lo tanto, de acuerdo con este método de la invención, el proceso de MT puede comprender:

I) Copia "clonal" de plantillas de código de barras y adaptadores necesarios, por ejemplo, mediante PCR en emulsión en perlas para crear miles de copias. Opcionalmente, la unidad copiada puede representar un transposón.

II) Mezclar fragmentos genómicos largos y perlas de adaptador de etiqueta en la relación adecuada y en concentraciones apropiadas para tener principalmente, la mayoría o casi todos los fragmentos genómicos asociados espacialmente con un concatámero y con poca frecuencia con dos o más.

III) Agregar un cebador universal al ADN genómico mediante: (a) corte de ADN genómico a una frecuencia predefinida (por ejemplo, 1 kb) usando corte parcial con un cortador frecuente u otros métodos; puede usarse una traducción de corte controlada para aleatorizar aún más los sitios de inicio de fragmentos; opcionalmente, se puede crear un pequeño espacio en el sitio de corte, por ejemplo, por actividad exo de Pol I o Klenow sin dNTP; (b) ligar un cebador por el

extremo 5’ al extremo 3' del ADN cortado proporcionando el cebador hibridado con un didesoxi oligo corto complementario en el extremo 5’; este cebador es complementario a un adaptador al lado del código de barras. Opcionalmente, esta etapa se puede realizar antes de la etapa dos o mezclar ADN genómico con etiquetas clonales;

IV) Copiar la etiqueta del donante de etiqueta (perla) y otro adaptador por extensión de cebador utilizando plantillas de etiqueta. Después de desnaturalizar el ADN, esto da como resultado fragmentos de ADNmc de aproximadamente 1 kb con una extensión de adaptador - código de barras -adaptador en el extremo 3’. Estos fragmentos pueden usarse como plantillas de secuenciación mediante un cebador complementario al adaptador del extremo 3’ o convertirse en ADNbc por el mismo cebador y un proceso adicional (por ejemplo, ligar un adaptador en el otro extremo, amplificar, circularizar) antes de la secuenciación.

Opcionalmente, las etapas 3 y 4 pueden reemplazarse por inserción de transposón y fragmentación o amplificación si las perlas representan clones de transposones etiquetados.

(5) Inserción sin corte - Método 1

Un enfoque alternativo para insertar secuencias que contienen una etiqueta no se basa en cortes. Véase Figura 4B. Los fragmentos **1** largos 1 se desnaturalizan (por ejemplo, por calentamiento) para producir cadenas sencillas complementarias. Los cebadores aleatorios (N mers) **29a**, **29b**, **29c** se hibridan a las cadenas individuales y se extienden con polimerasa. Se agrega una fosfatasa alcalina (por ejemplo, fosfatasa alcalina de camarón, SAP), y se usa polimerasa que tiene una función exonucleasa 3’ → 5' (por ejemplo, Klenow) para crear huecos. El producto resultante parcialmente de doble cadena que comprende N mers aleatorios **29a**, **29b**, **29c** entre las porciones del fragmento largo a secuenciar **31a**, **31b**, **31c** se maneja como se describió anteriormente y en la Figura 4A, comenzando con la ligadura 3’ de un adaptador común.

(6) Inserción sin cortes - Método 2

Un segundo enfoque alternativo inserta secuencias que contienen una etiqueta sin mellar, usando una partícula. Véanse las Figuras 4C y 4D. En este enfoque, se hibridan dos oligonucleótidos con una secuencia que contiene una etiqueta transportada en una partícula **15**, tal como una perla: (i) un cebador **23** común, que se hibrida secuencia arriba de la secuencia **22** de etiqueta o código de barras, y (ii) un adaptador **27** común que se hibrida secuencia abajo de la etiqueta. El cebador se extiende y se agrega ligasa para ligar el producto de extensión del cebador al adaptador **27** común. Este producto de ligadura, por lo tanto, incluye la secuencia **22** de etiqueta y, en su extremo 3’, el adaptador **27** común.

Una población de oligonucleótidos que incluye (i) una secuencia **29a** degenerada (N mer aleatorio) en su extremo 5’, (ii) una secuencia complementaria al adaptador **28** común y (iii) una secuencia no complementaria (no mostrada en la Figura 4C) se hibrida con el producto de ligadura de la etapa anterior y se realiza una extensión del cebador, añadiendo al extremo 3’ del producto de ligadura una secuencia **29a** degenerada complementaria a la de cada oligonucleótido (que posteriormente se elimina, por ejemplo, mediante digestión). El producto resultante (una población de "adaptadores etiquetados", cada uno con una secuencia **29a** degenerada en sus extremos 3’) se libera luego de la perla **15**, por ejemplo, por desnaturalización por calor. Los adaptadores etiquetados se hibridan con una cadena sencilla del fragmento **1** largo (producido desnaturalizando el fragmento largo de doble cadena); como se muestra en la Figura 4D, las diferentes secuencias degeneradas en los extremos de varios adaptadores **29a**, **29b** etiquetados se hibridan con las secuencias complementarias espaciadas a lo largo del fragmento **1** largo. Como se describió anteriormente, se agrega una polimerasa para extender el adaptador etiquetado y el producto de extensión incluye una secuencia complementaria a una región del fragmento **31a**, **31b** largo. Las moléculas resultantes, que incluyen un adaptador etiquetado unido a una secuencia del fragmento largo, se pueden usar para crear subfragmentos etiquetados del fragmento largo como se describió anteriormente (Figura 4A).

(7) Inserción utilizando traducción de corte controlado

La Figura 4G ilustra un enfoque similar a (4) anterior. Se corta el ADN largo de doble cadena y luego se abren los cortes en huecos cortos para facilitar la ligadura posterior del adaptador al extremo 3’ del hueco. El corte puede lograrse mediante digestión parcial con cualquier endonucleasa de corte (nickasa). Una nickasa adecuada es *Nt.CviPII*. El sitio de reconocimiento para *Nt.CviPII* es la secuencia corta CCD, en la que D = A, G o T. El hueco se puede abrir utilizando la actividad 3'-exo de la polimerasa Klenow de corrección de pruebas, que se unirá a los cortes y degradará la cadena cortada en una dirección de 3’ a 5', dejando un pequeño espacio en ausencia de nucleótidos, o mediante traducción controlada del corte (CNT), que utiliza la polimerasa Pol I que traduce el corte y una cantidad limitada de nucleótidos para traducir el corte a corta distancia. Esta reacción deja un pequeño espacio (1-3 bases) en lugar de un corte.

Este enfoque de corte proporciona una buena cobertura de lectura por fragmento largo de ADN porque usa ambas cadenas de ADN. En algunas realizaciones, los procesos pueden llevarse a cabo en un bloque de gel (u otro bloque de polímero o relleno en el fondo de un tubo o un pozo de placa de microtitulación), opcionalmente llevando a cabo las etapas enzimáticas en serie. Por ejemplo, si los fragmentos largos de ADN quedan atrapados en los tapones de gel y luego se lleva a cabo un corte controlado, por ejemplo por Vvn, seguido por la realización de lavado de la nickasa, entonces se puede hacer una reacción CNT para 20-100 bases (por ejemplo, aproximadamente 20 bases) para crear

un pequeño hueco (y aleatorizar aún más la posición de los sitios de corte) seguido por el lavado de la polimerasa. Otras formas de crear un hueco incluyen el uso de Klenow o exonucleasas. El producto resultante en esta etapa estaría listo para el ADN para la ligadura del adaptador "en el hueco" y los DNB en su mayoría intactos. El ADN se fragmentaría en segmentos de ~ 10 kb pero no podrían moverse significativamente en los tapones de gel. La etapa final de este tubo sencillo del método en gel LFR es a) fragmentar los DNB, crear un adaptador por hibridación/ligadura de componentes añadidos en la solución y b) ligar dichos adaptadores con código de barras en los huecos preparados en el ADN genómico. Esta etapa requiere mezclar nickasa específica para fragmentar DNB y ligasa. La reacción se puede detener lavando las enzimas y/o por destrucción de un solo golpe que también liberaría el ADN de los tapones de gel. Las perlas con código de barras clonal se utilizan en este método. Se pueden usar perlas a una distancia de 2 a 40 micrómetros, preferiblemente de 5 a 20 µm, en promedio. Las moléculas de ADN largas, que generalmente ocupan un espacio tridimensional que tiene aproximadamente 0,5-2 μm^3 o más en volumen, pueden usarse en una concentración creando una distancia promedio de 1 a 3 µm o incluso 4-10 µm, en promedio.

Serie de subfragmentos

Los métodos de (4), (5) y (6) (mostrados en las Figuras 4A a 4D) dan como resultado amplicones de PCR que son, efectivamente, subfragmentos etiquetados de los fragmentos largos de ADN. Esto es ventajoso si se utilizan métodos de secuenciación de lectura corta. Hay una variedad de formas de crear una serie de fragmentos de este tipo.

Por ejemplo, es posible crear una serie de dichos subfragmentos con regiones cada vez más cortas de los fragmentos largos de ADN como se muestra en la Figura 4E. Esto comienza con el subfragmento etiquetado con cebador extendido de extremos romos que resulta de la amplificación por PCR, que comprende una porción **31a** de un fragmento largo unido a una secuencia **22** de etiqueta, Un adaptador 3' **27** está ligado a los subfragmentos etiquetados. Un extremo del adaptador incluye una saliente; el otro extremo es un extremo romo que incluye un nucleótido bloqueado (por ejemplo, un ddNTP). Después de la ligadura del adaptador 3', el subfragmento se desnaturaliza y se realiza otra ronda de extensión del cebador utilizando traducción de corte controlado. La extensión del cebador se detiene antes de la finalización de modo que el cebador no se extienda hasta el final de la cadena complementaria. Un adaptador 3' **27** está ligado al extremo de la cadena extendida. Este proceso puede repetirse tantas veces como se desee con la extensión de la extensión del cebador variada para crear una serie de fragmentos **33** que tienen un extremo 5' común que se acorta en sus extremos 3'. Se proporcionan detalles de la estrategia de adaptador bloqueado y de la traducción de corte controlado, por ejemplo, en las solicitudes de patente de los Estados Unidos 12/329.365 (publicada como US 2012-0100534 A1) y 12/573.697 (publicada como US-2010-0105052-A1).

Otro enfoque para crear una serie de tales subfragmentos con regiones cada vez más cortas de los fragmentos largos de ADN como se muestra en La Figura 4F. Este enfoque también utiliza traducción de corte controlado. Los subfragmentos se ciclan y luego se dividen en dos o más pozos separados. La traducción de corte controlado se realiza de forma diferente en los distintos pozos para crear subfragmentos con un extremo común 5' que se acortan en sus extremos 3' en varios grados. Los subfragmentos se pueden agrupar y continuar el proceso. Otro enfoque utiliza la exonucleasa III u otras exonucleasas.

Estrategias para obtener una alta proporción de fragmentos largos etiquetados con exactamente una secuencia de etiqueta

El uso óptimo de los fragmentos largos se produce cuando la mayoría de ellos han sido etiquetados. El ensamblaje de lecturas en una secuencia de longitud completa o el análisis de regiones adyacentes del genoma suele ser más fácil si, en la mayoría o en todos los fragmentos largos, hay múltiples copias de una secuencia de una sola etiqueta que difieren de la secuencia de etiqueta en otros fragmentos largos. Por lo tanto, después de la amplificación de cada uno de los subfragmentos etiquetados, dos lecturas que tengan la misma secuencia de etiqueta habrían provenido del mismo fragmento largo. Las secciones que describen el uso del exceso de fragmentos de ADN o el exceso de capacidad de unión del ADN en los portadores de códigos de barras clonales, anteriormente, describieron cómo ajustar las diluciones y las proporciones de la reacción de fragmentos largos de ADN e introdujeron secuencias de etiquetas clonales para optimizar el etiquetado.

Una estrategia para obtener una alta proporción de fragmentos largos etiquetados con exactamente una secuencia de etiqueta implica el anclaje y otras formas de asociar partículas con fragmentos largos individuales. Hay varias formas de unir o atar una sola perla con múltiples copias de una secuencia que contiene una etiqueta particular a un solo fragmento largo de un ácido nucleico objetivo. Por ejemplo, se puede agregar una secuencia de homopolímero (por ejemplo, una cola A) al fragmento largo usando la transferasa terminal o se puede ligar un adaptador con una secuencia seleccionada a un extremo o extremos del fragmento largo. Se puede agregar una secuencia complementaria al final o incluirse dentro de la secuencia que contiene una etiqueta o nanobola de tal manera que, en condiciones apropiadas seleccionadas, la secuencia que contiene una etiqueta o el ensamblaje de la etiqueta se hibrida con la secuencia complementaria correspondiente en el fragmento largo. Preferiblemente, un fragmento largo puede hibridarse con solo una secuencia que contiene una etiqueta o un conjunto de etiquetas.

Varias perlas utilizadas para la amplificación clonal de oligonucleótidos adaptador-código de barras adaptador pueden tener una capacidad de unión a ADN temporal débil adicional: por ejemplo, una superficie cargada positivamente o una superficie que se une a bases de ADN. Se pueden usar perlas de diferentes tamaños para proporcionar suficientes

copias de códigos de barras y también suficiente superficie para unir fragmentos largos de ADN.

Características de fragmentos largos etiquetados

Se describe una molécula de ADN que comprende una secuencia genómica (G) y una pluralidad de secuencias discretas introducidas (IS), en la que dichas secuencias introducidas no son naturalmente contiguas con la secuencia de ADN genómico. La molécula de ADN es de cadena sencilla o cadena doble. El ADN tiene una longitud de 5 kb a 750 kb. En algunas realizaciones, el ADN tiene una longitud de al menos 5 kb, al menos 7,5 kb o al menos 10 kb, tal como una longitud en el intervalo de 5-20 kb, 7,5-15 kb o 10-12,5 kb. En algunas realizaciones, el ADN tiene una longitud de al menos 50 kb, al menos 75 kb o al menos 100 kb, tal como una longitud en el intervalo de 50-200 kb, 75-150 kb o 100-125 kb. Generalmente, cada fragmento comprende al menos 5, al menos 10, al menos 25, o al menos 50 secuencias introducidas en las que cada una de dichas secuencias introducidas tiene la misma secuencia o comprende una subsecuencia común.

El espaciado promedio entre las secuencias introducidas es de 100 pb, 200 pb, 300 pb, 400 pb, 500 pb, 600 pb, 700 pb, 800 pb, 900 pb, 1000 pb, 1500 pb, 2000 pb, 2500 pb, 3000 pb, 3500 pb, 4000 pb o 5000 pb. De acuerdo con otra realización, el espaciado promedio es de entre aproximadamente 100 pb y aproximadamente 5000 pb, o entre aproximadamente 200 pb y aproximadamente 4000 pb, o entre aproximadamente 300 pb y aproximadamente 3000 pb, o entre aproximadamente 300 pb y aproximadamente 2000 pb, o entre aproximadamente 300 pb y aproximadamente 1000 pb.

La longitud del ADN, N, es de 50 a 150 kb y el número de secuencias introducidas, IS, está en el intervalo [(N x 1) a (N x 4)]. Alternativamente, la longitud del ADN, N, es de 50 a 150 kb y el número de secuencias introducidas, IS, está en el intervalo de [(N x 2) - (N x 10)]. Alternativamente, la longitud del ADN, N, es de 50 a 150 kb y el número de secuencias introducidas, IS, está en el intervalo [(N x 1) a (N x 0,2)].

En algunas realizaciones, la longitud del ADN, N, es de 5 a 15 kb y el número de secuencias introducidas, IS, está en el intervalo [(N x 1) a (N x 4)]. Alternativamente, la longitud del ADN, N, es de 5 a 15 kb y el número de secuencias introducidas, IS, está en el intervalo [(N x 2) - (N x 10)]. Alternativamente, la longitud del ADN, N, es de 5 a 15 kb y el número de secuencias introducidas, IS, está en el intervalo [(N x 1) a (N x 0,2)].

La secuencia de ADN genómico puede ser, por ejemplo, de un animal, tal como un mamífero (por ejemplo, humano), una planta, un hongo o una bacteria.

Las secuencias introducidas pueden comprender secuencias de transposón y/o secuencias de unión a cebador.

Las composiciones que comprenden una población de moléculas de ADN como se describió anteriormente pueden en conjunto representar esencialmente toda la secuencia genómica de un organismo (por ejemplo, al menos 80%, al menos 90%, al menos 95% o al menos 99%). La composición puede comprender además (i) una transposasa, (ii) una ADN polimerasa y/o (iii) cebadores de amplificación que se unen a una secuencia en la secuencia introducida o el complemento de una secuencia en la secuencia introducida.

Las composiciones que comprenden una población de fragmentos largos etiquetados como se describió anteriormente pueden comprender en conjunto al menos aproximadamente 10^4, al menos aproximadamente 10^5, al menos aproximadamente 10^6 o al menos aproximadamente 10^7 diferentes códigos de barras insertados.

En un caso, la mayoría o esencialmente todas (por ejemplo, al menos 50%, al menos 80%, al menos 90%, al menos 95% o al menos 99%) de las moléculas de ADN en la población comprenden una secuencia introducida única (es decir, una secuencia introducida no compartida por otras moléculas). En algunos casos, las secuencias introducidas únicas comparten una subsecuencia común. La subsecuencia común puede ser una secuencia de unión a cebador.

En algunos casos, más del 10%, más del 20%, más del 30% o más del 50% de la longitud del ADN en un fragmento largo puede representarse en fragmentos de ADN cortos etiquetados con las copias de los mismos códigos de barras. Por ejemplo, el uso de un fragmento de 100 kb de largo para generar 100 subfragmentos con una longitud promedio de 1 kb, daría como resultado 20-50 fragmentos etiquetados útiles. Los otros fragmentos se perderían como sin etiquetar o demasiado cortos o demasiado largos.

En algunos casos, un número sustancial (más del 25%) o la mayoría (más del 50%) de fragmentos largos de ADN en una composición están etiquetados con más de una (por ejemplo, dos o tres) etiquetas diferentes.

Producción de subfragmentos de fragmentos largos etiquetados

Después del etiquetado, los fragmentos largos del ácido nucleico objetivo se subfragmentan a un tamaño deseado mediante amplificación (por ejemplo, por PCR, extensión de cebador, RCA), digestión con enzimas de restricción (por ejemplo, usando un cortador raro que tiene un sitio de reconocimiento dentro una secuencia que contiene una etiqueta introducida en fragmentos largos), o por otras técnicas convencionales, que incluyen digestión enzimática, cizallamiento, sonicación, etc.

Los tamaños de los subfragmentos pueden variar dependiendo del ácido nucleico fuente objetivo y los métodos de

construcción de la biblioteca usados, pero para la secuenciación estándar del genoma completo, dichos fragmentos típicamente varían de 50 a 2000 nucleótidos de longitud. En otras realizaciones, los fragmentos tienen de 300 a 600 nucleótidos de longitud, de 200 a 2000 nucleótidos de longitud o de 1000 a 5000 nucleótidos de longitud. En otra realización más, los fragmentos tienen de 10-100, 50-100, 50-300, 100-200, 200-300, 50-400, 100-400, 200-400, 300-400, 400-500, 400- 600, 500-600, 50-1000, 100-1000, 200-1000, 300-1000, 400-1000, 500-1000, 600-1000, 700-1000, 700-900, 700-800, 800-1000, 900-1000, 1500-2000, 1750-2000 y 50-2000 nucleótidos de longitud.

En una realización adicional, se aíslan fragmentos de un tamaño particular o en un intervalo particular de tamaños. Tales métodos son bien conocidos en la técnica. Por ejemplo, el fraccionamiento en gel se puede usar para producir una población de fragmentos de un tamaño particular dentro de un intervalo de pares de bases, por ejemplo para 500 pares de bases + 50 pares de bases.

Dependiendo de la selección de las condiciones de etiquetado y procesamiento posterior y diferentes longitudes de lectura de secuencia que comienzan con aproximadamente 5 a aproximadamente 1.000.000 equivalentes de genoma de ADN de fragmento largo aseguran que la población de fragmentos largos cubra todo el genoma. Las bibliotecas que contienen plantillas de ácido nucleico generadas a partir de una población de fragmentos superpuestos proporcionarán la mayor parte o la totalidad de la secuencia de un genoma completo.

Características de los subfragmentos

Una composición puede comprender una población de polinucleótidos, cada uno de los cuales comprende (1) la secuencia correspondiente al segmento de un ADN genómico; (2) secuencias introducidas (por ejemplo, códigos de barras clonales) en uno o ambos terminales, en los que la población comprende una pluralidad de segmentos diferentes de secuencia de ADN genómico, y las secuencias introducidas comprenden en conjunto una pluralidad de diferentes secuencias de etiquetas o códigos de barras de manera que algunos polinucleótidos que comprenden diferentes segmentos de ADN genómico comprenden al menos una secuencia de etiqueta o código de barras en común; y (3) la población de polinucleótidos comprende al menos 10^4 secuencias de códigos de barras diferentes, al menos 10^5 secuencias de códigos de barras diferentes, al menos 10^6 secuencias de códigos de barras diferentes o al menos 10^7 secuencias de códigos de barras diferentes. En algunas realizaciones, el polinucleótido tiene un tamaño promedio (en bases o pares de bases) en el intervalo de 50-5000, tal como 50-100, 100-200, 200-300, 300-500, 500-700, 700-1000, 1000-1500, 1500-2000, 2000-3000, 3000-4000 o 4000-5000. En algunos casos, al menos un par de segmentos de ADN que comprenden una etiqueta o código de barras en común son adyacentes en el genoma. En algunos casos, los polinucleótidos que comprenden al menos una secuencia de etiqueta o código de barras en común comprenden solo segmentos de ADN genómico que no se superponen, en los que "que no se superponen" significa que los segmentos no se superponen en el genoma. A veces, la composición comprende una pluralidad (por ejemplo, al menos 10, al menos 100 o al menos 500) de polinucleótidos en una o múltiples copias que comparten la misma etiqueta o código de barras y son adyacentes en el genoma. A veces, la composición comprende una pluralidad (por ejemplo, al menos 10, al menos 100 o al menos 500) de pares de polinucleótidos que comparten la misma etiqueta o código de barras y son adyacentes en la secuencia objetivo (por ejemplo, genoma). Las secuencias de etiquetas pueden comprender códigos de barras en combinación con secuencias de transposón y/o sitios de unión de cebador. La secuencia introducida no es naturalmente contigua al segmento de ADN genómico.

El ADN genómico puede ser de una planta, animal (por ejemplo, un mamífero tal como un humano), bacterias u hongos. Para las bacterias puede ser una mezcla (metagenomas, para permitir el ensamblaje de cepas y genomas de especies sin cultivar cepa por cepa) o cepas o especies aisladas. Los polinucleótidos pueden ser amplicones.

Amplificación

Antes o después de cualquier etapa descrita en este documento, se puede usar una etapa de amplificación para asegurar que haya suficiente ácido nucleico disponible para las etapas posteriores.

De acuerdo con una realización de la divulgación, se proporcionan métodos para secuenciar pequeñas cantidades de ácidos nucleicos complejos, incluidos los de organismos superiores, en los que dichos ácidos nucleicos complejos se amplifican para producir suficientes ácidos nucleicos para la secuenciación mediante los métodos descritos en este documento. Una sola célula humana incluye aproximadamente 6,6 picogramos (pg) de ADN genómico. La secuenciación de ácidos nucleicos complejos de un organismo superior se puede lograr usando 1 pg, 5 pg, 10 pg, 30 pg, 50 pg, 100 pg o 1 ng o más, de un ácido nucleico complejo como material de partida, que se amplifica mediante cualquier método de amplificación de ácido nucleico conocido en la técnica, para producir, por ejemplo, 200 ng, 400 ng, 600 ng, 800 ng, 1 µg, 2 µg, 3 µg, 4 µg, 5 µg, 10 µg o cantidades mayores del ácido nucleico complejo. También se divulgan protocolos de amplificación de ácido nucleico que minimizan el sesgo de GC. Sin embargo, la necesidad de amplificación y el sesgo de GC posterior se pueden reducir aún más simplemente aislando una célula o un pequeño número de células, cultivándolas durante un tiempo suficiente en condiciones de cultivo adecuadas conocidas en la técnica, y usando la progenie de la célula o células iniciales para secuenciación.

Tales métodos de amplificación incluyen, sin limitación: amplificación de desplazamiento múltiple (MDA), reacción en cadena de la polimerasa (PCR), reacción en cadena de ligadura (a veces denominada como amplificación por oligonucleótido ligasa OLA), tecnología de sonda de ciclación (CPT), ensayo de desplazamiento de cadena (SDA),

amplificación mediada por transcripción (TMA), amplificación basada en secuencia de ácido nucleico (NASBA), amplificación de círculo rodante (RCA) (para fragmentos circulares) y tecnología de escisión invasiva.

La amplificación se puede realizar después de fragmentar o antes o después de cualquier etapa descrita en el presente documento.

Protocolos de amplificación de genoma completo

Los protocolos de amplificación particulares que se han usado durante el desarrollo de esta invención incluyen los siguientes.

Un primer protocolo de amplificación es la inserción de adaptadores mediada por transposón para la amplificación por PCR larga. Para maximizar la cobertura del genoma mediante una secuencia de lectura corta, se amplificaron fragmentos relativamente grandes del genoma. Esto permite generar fragmentos superpuestos más cortos, que luego pueden secuenciarse. Los transposones se insertan a una frecuencia de 3-20 kb en ADN genómico largo. Luego se puede realizar una PCR de cebador único o dos cebadores para un pequeño número de ciclos para generar una amplificación de más de 10 veces. En una versión de este proceso, solo se realiza una ronda de PCR larga seguida de la fragmentación hasta la superposición de -300 pb a 1,5 kilobases mediante la incorporación de uracilo durante la amplificación (CoRE), fragmentación ultrasónica, digestión con nucleasa, fragmentación de transposón u otro método adecuado.

En otra versión de este proceso, primero se generan productos de PCR más largos de ~ 10 kb de tamaño y se realiza una amplificación de menos de 100 veces. Una segunda ronda de inserción de transposones se realiza a una frecuencia de -3 kb. Se realizan rondas adicionales de amplificación por PCR para generar una amplificación de más de 1000 veces. Ahora se puede realizar la fragmentación como se describió anteriormente.

Un segundo protocolo es insertar adaptadores en los huecos generados en los fragmentos largos. Los cortes se introducen por primera vez a una frecuencia de 3-20 kb. Los cortes se abren para huecos de más de un 1 pb usando nucleasas o polimerasas en ausencia de nucleótidos en el tampón. Luego, los adaptadores se ligan en el OH de 3' y el PO_4 de 5’ del hueco. En el lado 3’ no es necesaria la hibridación con bases en el hueco. En el lado 5’, será necesario un adaptador con N6 en el extremo 3' para hibridar primero adyacente al PO_4 de 5’ antes de la ligadura. Una vez que se completa la ligadura, se pueden realizar PCR de 1 o 2 cebador. Los cebadores largos de PCR se fragmentan nuevamente en fragmentos superpuestos más pequeños de 300-1,5 kb como se indicó anteriormente.

Protocolo de amplificación MDA con sesgo de GC reducido

Se describen métodos de amplificación de ácido nucleico en los que el ácido nucleico se amplifica fielmente, por ejemplo, aproximadamente 30.000 veces dependiendo de la cantidad de ADN inicial.

De acuerdo con una realización de los métodos de MT, MT comienza con el tratamiento de ácidos nucleicos genómicos, generalmente ADN genómico, con una exonucleasa 5’ para crear salientes de cadena sencilla 3'. Tales salientes de cadena sencilla sirven como sitios de iniciación de MDA. El uso de la exonucleasa también elimina la necesidad de una etapa de desnaturalización alcalina o térmica antes de la amplificación sin introducir sesgo en la población de fragmentos. En otra realización, la desnaturalización alcalina se combina con el tratamiento con exonucleasa 5’, lo que da como resultado una reducción en el sesgo que es mayor de lo que se observa con cualquiera de los tratamientos solos. Los fragmentos son luego amplificados.

En una realización, se usa una amplificación de desplazamiento múltiple (MDA) basada en phi29. Numerosos estudios han examinado el intervalo de sesgos de amplificación no deseados, formación de productos de fondo y artefactos quiméricos introducidos a través de la MDA basada en phi29, pero muchos de estos inconvenientes se han producido en condiciones extremas de amplificación (más de 1 millón de veces). Comúnmente, MT emplea un nivel de amplificación sustancialmente más bajo y comienza con fragmentos largos de ADN (por ejemplo, -100 kb), lo que resulta en una MDA eficiente y un nivel más aceptable de sesgos de amplificación y otros problemas relacionados con la amplificación.

Se ha desarrollado un protocolo de MDA mejorado para superar los problemas asociados con la MDA que utiliza varios aditivos (por ejemplo, enzimas modificadoras del ADN, azúcares y/o productos químicos tales como DMSO), y/o diferentes componentes de las condiciones de reacción para la MDA se reducen, aumentan o sustituyen para mejorar aún más el protocolo. Para minimizar las quimeras, también se pueden incluir reactivos para reducir la disponibilidad del ADN de una sola cadena desplazado para que actué como una plantilla incorrecta para la extensión de la cadena de ADN, que es un mecanismo común para la formación de quimeras. Una fuente importante de sesgo de cobertura introducido por MDA es causada por las diferencias en la amplificación entre regiones ricas en GC frente a regiones ricas en AT. Esto puede corregirse utilizando diferentes reactivos en la reacción de MDA y/o ajustando la concentración del cebador para crear un entorno para cebar incluso en todo el porcentaje de las regiones de GC del genoma. En algunas realizaciones, se usan hexámeros aleatorios en el cebado de MDA. En otras realizaciones, se utilizan otros diseños de cebadores para reducir el sesgo. En realizaciones adicionales, el uso de exonucleasa 5’ antes o durante la MDA puede ayudar a iniciar un cebado exitoso de bajo sesgo, particularmente con fragmentos más largos (es decir, de 200 kb a 1Mb) que son útiles para secuenciar regiones caracterizadas por una larga duplicación segmentaria (es

decir, en algunas células cancerosas) y repeticiones complejas.

En algunas realizaciones, se usan etapas de fragmentación y ligadura mejoradas y más eficientes que reducen el número de rondas de amplificación de MDA requeridas para preparar muestras hasta 10.000 veces, lo que reduce aún más el sesgo y la formación de quimeras resultantes de la MDA.

En algunas realizaciones, la reacción de MDA está diseñada para introducir uracilos en los productos de amplificación en preparación para la fragmentación de CoRE. En algunas realizaciones, se usa una reacción de MDA estándar que utiliza hexámeros aleatorios para amplificar los fragmentos en cada pozo; alternativamente, se pueden usar cebadores aleatorios de 8 mers para reducir el sesgo de amplificación (por ejemplo, el sesgo de GC) en la población de fragmentos. En realizaciones adicionales, también se pueden agregar varias enzimas diferentes a la reacción de MDA para reducir el sesgo de la amplificación. Por ejemplo, pueden usarse bajas concentraciones de exonucleasas 5' no procesadoras y/o proteínas de unión de cadena sencilla para crear sitios de unión para los 8 mers. Los agentes químicos como la betaína, el DMSO y la trehalosa también se pueden usar para reducir el sesgo.

Después de la amplificación de los ácidos nucleicos en una muestra, los productos de amplificación pueden opcionalmente fragmentarse. En algunas realizaciones, el método CoRE se usa para fragmentar más los fragmentos después de la amplificación. En tales realizaciones, la amplificación MDA de fragmentos está diseñada para incorporar uracilos en los productos de MDA. El producto de MDA se trata luego con una mezcla de uracilo ADN glicosilasa (UDG), ADN glicosilasa liasa endonucleasa VIII y polinucleótido quinasa T4 para escindir las bases de uracilo y crear huecos de una sola base con grupos funcionales 5' fosfato y 3' hidroxilo. La traducción de corte mediante el uso de una polimerasa tal como la Taq polimerasa da como resultado rompimiento de extremos romos de doble cadena, lo que da como resultado fragmentos ligables de un intervalo de tamaño que depende de la concentración de dUTP añadida en la reacción de MDA. En algunas realizaciones, el método CoRE utilizado implica la eliminación de uracilos por polimerización y el desplazamiento de la cadena por phi29. La fragmentación de los productos de MDA también se puede lograr mediante sonicación o tratamiento enzimático. El tratamiento enzimático que podría usarse en esta realización incluye, sin limitación, DNasa I, endonucleasa I T7, nucleasa micrococcal y similares.

Después de la fragmentación de los productos de MDA, los extremos de los fragmentos resultantes pueden repararse. Muchas técnicas de fragmentación pueden dar como resultado terminales con extremos sobresalientes y terminales con grupos funcionales que no son útiles en reacciones de ligadura posteriores, tales como grupos hidroxilo 3' y 5' y/o grupos fosfato 3' y 5'. Puede ser útil tener fragmentos reparados para tener extremos romos. También puede ser deseable modificar los terminales para agregar o eliminar grupos fosfato e hidroxilo para evitar la "polimerización" de las secuencias objetivo. Por ejemplo, una fosfatasa puede usarse para eliminar grupos fosfato, de modo que todos los extremos contengan grupos hidroxilo. Cada extremo puede ser alterado selectivamente para permitir la ligadura entre los componentes deseados. Un extremo de los fragmentos se puede "activar" por tratamiento con fosfatasa alcalina.

Secuenciación de ácido nucleico

Los métodos de MT descritos en este documento pueden usarse como una etapa de procesamiento previo para secuenciar genomas diploides usando cualquier método de secuenciación conocido en la técnica, que incluye por ejemplo, sin limitación, secuenciación por síntesis basada en polimerasa (por ejemplo, sistema HiSeq 2500, Illumina, San Diego, CA), secuenciación basada en ligadura (por ejemplo, SOLiD 5500, Life Technologies Corporation, Carlsbad, CA), secuenciación de semiconductores iónicos (por ejemplo, secuenciadores Ion PGM o Ion Proton, Life Technologies Corporation, Carlsbad, CA), Guías de onda de modo cero (por ejemplo, secuenciador PacBio RS, Pacific Biosciences, Menlo Park, CA), secuenciación de nanoporos (por ejemplo, Oxford Nanopore Technologies Ltd., Oxford, Reino Unido), pirosecuenciación (por ejemplo, 454 Life Sciences, Branford, CT), u otras tecnologías de secuenciación. Algunas de estas tecnologías de secuenciación son tecnologías de lectura corta, pero otras producen lecturas más largas, por ejemplo, GS FLX+ (454 Life Sciences; hasta 1000 pb), PacBio RS (Pacific Biosciences; aproximadamente 1000 pb) y secuenciación de nanoporos (Oxford Nanopore Technologies Ltd.; 100 kb). Para el ajuste de fase del haplotipo, las lecturas más largas son ventajosas, ya que requieren mucho menos cómputo, aunque tienden a tener una tasa de error más alta y los errores en lecturas tan largas pueden necesitar ser identificadas y corregidas de acuerdo con los métodos establecidos en el presente documento antes del ajuste de fase del haplotipo.

[0168] De acuerdo con una realización, la secuenciación se realiza usando la ligadura de anclaje de sonda combinatoria (cPAL) como se describe, por ejemplo, en las publicaciones de solicitud de patente de los Estados Unidos Nos. 2010/0105052; US 2007099208; US 2009/0264299; US 2009/0155781; US 2009/0005252; US 2009/0011943; US 2009-0118488; US 2007/0099208; US 2008/0234136; US 2009/0137404; US 2009/0137414; US 2007/0072208; US 2010/0081128; US 2008/0318796; US 2009/0143235; US 2008/0213771; US 2008/0171331; US 2007/0037152; US 2009/0005259; US 2009/0036316; US 2009/0011416; US 2009/0075343; US 2009/0111705; US 2009/0111706; US 2009/0203551; US 2009/0105961; US 2008/0221832; US 2009/0318304; US 2009/0111115; US 2009/0176652; US 2009/0311691; US 2009/0176234; US 2009/0263802; US 2011/0004413; y las solicitudes internacionales de patente publicadas Nos. WO2007120208, WO2006073504 y WO2007133831, y las solicitudes de patente de los Estados Unidos Nos. 13/448.279 (publicada como US 20140051588), 13/447.087 (publicada como 20130124100).

Ejemplos de métodos para llamar variaciones en una secuencia de polinucleótidos comparada con una secuencia de polinucleótidos de referencia y para el ensamblaje (o reensamblaje) de secuencias de polinucleótidos, por ejemplo, se

proporcionan en la publicación de patente de los Estados Unidos No. 2011-0004413, (solicitud No. 12/770.089). Véase también Drmanac et al., Science 327,78-81, 2010. También se hace referencia a la solicitud No. 61/623,876, titulada "Identification Of DNA Fragments And Structural Variations"; la solicitud No. 13/649.966, publicado como la publicación de patente de los Estados Unidos No. 2013-0096841; y la solicitud No.13/447,087, titulada "Processing and Analysis of Complex Nucleic Acid Sequence Data" publicada como la publicación de patente de los Estados Unidos No. 2013/0124100.

Al secuenciar el 50% de cada fragmento de aproximadamente 1 kb, se generaría una cobertura de secuencia de aproximadamente 1x para cada fragmento genómico porque los fragmentos etiquetados se generan a partir de ambas cadenas de ADNbc. Si se secuencia un 25% (1/2 de cobertura de lectura por fragmento), se observaría el enlace de dos regiones en el 25% de los fragmentos. Para el mismo presupuesto de lectura, se puede aumentar el número de fragmentos dos veces y tener solamente una reducción de dos veces en los enlaces observados. Para un 25% de lectura (125 bases forman cada extremo del fragmento de 1 kb) y 36 células iniciales, se observarán nueve enlaces en lugar de aproximadamente 18 enlaces para 18 células si se lee el 50% del ADN (250 bases de cada extremo y un fragmento de aproximadamente 1 kb). Si solo se pueden leer -60 bases de cada fragmento, es mejor usar fragmentos de 300-500 pb que todavía sean pares de empalme muy útiles.

Si se secuencia una fracción de ADN de cada fragmento de aproximadamente 1 kb, se necesitan más fragmentos iniciales. Por ejemplo, si se secuencia la mitad, se requieren 4 veces más fragmentos.

Reducción de complejidad

En un aspecto adicional, las técnicas de MT de la invención reducen la complejidad del ADN a secuenciar para centrarse en la secuencia de interés tal como un panel de genes dirigido para diferentes enfermedades, exomas o cepas bacterianas raras. La reducción de la complejidad y la separación del haplotipo en ADN de más de 100 kb de largo pueden ser útiles para un ensamblaje de secuencias más eficiente y rentable y para la detección de variaciones de secuencia en genomas humanos y otros diploides y poliploides o mezclas de genomas bacterianos y otros (metagenomas). Una forma de reducir la complejidad de los fragmentos de ADN etiquetados es usar oligonucleótidos de captura para secuencias genómicas de interés, por ejemplo, secuencias de codificación (por ejemplo, selección de exoma para obtener variantes de exoma "en fase").

Los fragmentos de ADN de interés pueden enriquecerse mediante: (a) proporcionar una mezcla de fragmentos de ADN etiquetados por una pluralidad de códigos de barras, (b) capturar fragmentos de ADN utilizando oligonucleótidos complementarios al menos en parte a las secuencias de códigos de barras de interés, (c) descartar ADN no capturado preferiblemente que no tiene oligonucleótidos de captura que coincidan con el código de barras, enriqueciendo así fragmentos de ADN de interés. La mezcla puede comprender fragmentos de ADN de más de 30, 100 o 1000 cepas bacterianas o especies en representación variable. Se puede usar un código de barras para etiquetar el ADN de una o unas pocas células bacterianas. Se pueden usar más de 10, 30, 50, 100 o 300 oligonucleótidos de captura diferentes, cada uno específico para un código de barras. En ambos enfoques, en lugar de la selección positiva, se puede usar una selección negativa para eliminar secuencias no deseadas, generalmente frecuentes, tales como las repeticiones de Alu en el genoma humano o las bacterias frecuentes en el microbioma.

Definiciones

Como se usa en el presente documento, un "fragmento" o subfragmento de un ácido nucleico objetivo, tal como un fragmento de ADN genómico, un fragmento de ADN cromosómico, un fragmento largo o un subfragmento (de un fragmento largo, de un ADN objetivo, etc.) se refiere a la relación de las secuencias, en vez de necesariamente a una derivación física directa. Un "fragmento" de una secuencia más larga (por ejemplo, un "fragmento de una molécula de ácido nucleico objetivo" o "un subfragmento de un fragmento largo etiquetado") comprende la secuencia de una porción de la secuencia más larga, o alternativamente, comprende el complemento exacto de una secuencia de una parte de la secuencia más larga, independientemente de cómo se produce el fragmento. Por ejemplo, se puede producir un "subfragmento" de un ácido nucleico objetivo o de un fragmento largo por amplificación o replicación de una porción del ácido nucleico objetivo, lo que da como resultado una nueva molécula que comprende una secuencia que es igual o exactamente complementaria a, la secuencia de ácido nucleico objetivo. En otro ejemplo, se puede producir un "subfragmento" de un ácido nucleico objetivo o de un fragmento largo por fragmentación física o enzimática del ácido nucleico objetivo o fragmento largo.

El término "fragmento largo" se usa en esta descripción para referirse a un polímero de ácido nucleico de partida que se usa en un protocolo de esta invención para etiquetar, secuenciar, analizar o procesar adicionalmente. El término no requiere que el ácido nucleico se obtenga de ninguna fuente en particular o por ningún proceso en particular. El ácido nucleico puede tener cualquier longitud y tener características que sean compatibles con el protocolo al que se hace referencia. Puede haber sido previamente amplificado, fragmentado, recombinado o procesado antes de la etapa inicial del protocolo al que se hace referencia. En algunas de las ilustraciones proporcionadas en esta descripción, un "fragmento largo" inicial está entre aproximadamente 10 a 100 kb o 30-300 kb o la mayor parte o la mayoría son más largos que 10 o más largos que 20 kb o más largos que 30 kb o 50 kb.

Como se usa en este documento, la "amplificación" de ácido nucleico se refiere a métodos en los que se producen

copias de polinucleótidos a través de ciclos de polimerización o ligadura, a una velocidad geométrica o conveniente, a diferencia de la replicación de plantilla, en la que una copia única de una plantilla se realiza o en la que se obtienen una o más copias de la plantilla por RCR.

Un fragmento de un ácido nucleico objetivo a secuenciar y analizar a veces se denomina "fragmento inicial". El término se usa solo para indicar que un fragmento que tiene las características especificadas es un producto usado tempranamente en un protocolo particular.

Como se usa en el presente documento, "insertar", "introducir" e "incorporar" no se limitan a incorporar físicamente un polinucleótido (por ejemplo, un oligonucleótido que contiene un código de barras) en otro polinucleótido (por ejemplo, un ADN largo). Por ejemplo, como se describió anteriormente, se puede introducir una secuencia de código de barras en un fragmento largo de ADN replicando toda o parte de la secuencia de fragmento largo de ADN junto con secuencias de etiquetas intercaladas (véase, por ejemplo, la Fig. 2A). Se puede introducir una secuencia de código de barras en un fragmento largo de ADN mediante la transposición física de una secuencia de código de barras en la molécula de ADN larga (véase, por ejemplo, la Figura 1A).

Como se usa en el presente documento, "intercalado" tiene su significado normal en la técnica. Por ejemplo, un fragmento largo etiquetado que contiene "secuencia de ácido nucleico objetivo y múltiples secuencias de etiquetas intercaladas", es un polinucleótido que comprende una secuencia de ácido nucleico objetivo, tal como una secuencia genómica, interrumpida por múltiples secuencias de etiquetas (por ejemplo, copias múltiples de una secuencia de etiqueta) de modo que las secuencias de etiquetas se encuentran entre secuencias objetivo que son contiguas en el ácido nucleico objetivo. Como se discutió anteriormente, el espacio promedio entre las secuencias de etiquetas introducidas adyacentes puede ser, por ejemplo, 100 pb, 200 pb, 300 pb, 400 pb, 500 pb, 600 pb, 700 pb, 800 pb, 900 pb, 1000 pb, 1500 pb , 2000 pb, 2500 pb, 3000 pb, 3500 pb, 4000 pb o 5000 pb.

El término "transposón", como se usa en el presente documento, se refiere a un segmento de ácido nucleico que es reconocido por una transposasa o una enzima integrasa y es capaz de transposición.

El término "transposasa", como se usa en el presente documento, se refiere a una enzima que es un componente de un complejo funcional de ácido nucleico-proteína capaz de transposición y que media en la transposición. El término "transposasa" también se refiere a integrasas de retrotransposones o de origen retroviral.

La expresión "reacción de transposición" usada en el presente documento se refiere a una reacción en la que un transposón se inserta en un ácido nucleico objetivo. Los componentes primarios en una reacción de transposición son un transposón y una transposasa o una enzima integrasa.

El término "secuencia del extremo del transposón" o "extremos del transposón", como se usa en el presente documento, se refiere a las secuencias de nucleótidos en los extremos distales de un transposón. Las secuencias finales del transposón son responsables de identificar el transposón para la transposición; son las secuencias de ADN que requiere la enzima de transposición para formar un complejo de transpososomas y realizar una reacción de transposición. Un ADN transponible puede comprender solo una secuencia del extremo del transposón o más de una secuencia del extremo del transposón. La secuencia del extremo del transposón en la secuencia de ADN transponible, por lo tanto, no está unida a otra secuencia del extremo del transposón por la secuencia de nucleótidos, es decir, el ADN transponible contiene solo una secuencia de unión a la transposasa. Por lo tanto, el ADN transponible comprende un "extremo de transposón" (véase, por ejemplo, Savilahti et al., EMBO J. 14: 4893-4903, 1995).

El término "secuencia de unión a la transposasa" o "sitio de unión a la transposasa" como se usa en el presente documento se refiere a las secuencias de nucleótidos que siempre están dentro de la secuencia del extremo del transposón a la que se une específicamente una transposasa cuando media la transposición. La secuencia de unión a la transposasa puede comprender más de un sitio para la unión de subunidades de transposasa.

El término "cadena de unión al transposón" o "extremo de unión", como se usa en el presente documento, significa el extremo de esa cadena del ADN del transposón de cadena doble, que está unido por la transposasa al ADN objetivo en el sitio de inserción.

Los complejos de transposones se forman entre una enzima transposasa y un fragmento de ADN de cadena doble que contiene una secuencia de unión específica para la enzima, denominada "extremo del transposón". La secuencia del sitio de unión al transposón puede modificarse con otras bases, en ciertas posiciones, sin afectar la capacidad del complejo de transposón para formar una estructura estable que pueda transponerse eficientemente en el ADN objetivo. Al manipular la secuencia del extremo del transposón, el método proporcionó propiedades al ADN objetivo fragmentado que se puede utilizar en aplicaciones posteriores, particularmente cuando se usa el método para la preparación de la biblioteca antes de la secuenciación.

El término "adaptador" o "cola de adaptador", como se usa en el presente documento, se refiere a un componente de ácido nucleico no objetivo, generalmente ADN, que proporciona un medio para abordar un fragmento de ácido nucleico al que está unido. Por ejemplo, en las realizaciones, un adaptador comprende una secuencia de nucleótidos que permite la identificación, el reconocimiento y/o la manipulación molecular o bioquímica del ADN al que está unido el adaptador (por ejemplo, proporcionando un sitio para la hibridación de un oligonucleótido, tal como un cebador para

la extensión por una ADN polimerasa, o un oligonucleótido para captura o para una reacción de ligadura).

El término "partícula" como se usa en esta descripción se refiere a un sistema de suministro para múltiples copias de un oligonucleótido pequeño, tal como un transposón o cebador. El oligonucleótido se une a la partícula o se incorpora a ella de una manera que lo hace liberable con el fin de participar en una reacción o recombinación, por ejemplo, usando una nucleasa de restricción. Los ejemplos no limitantes incluyen nanoperlas a las que se unen múltiples copias de un oligonucleótido. Las copias de oligonucleótidos en la partícula típicamente comprenden una secuencia de etiqueta que difiere de las secuencias de etiquetas en otras partículas. Cuando esta divulgación se refiere a un concatámero o perla que participa en una reacción, a menos que se indique o se requiera lo contrario, la descripción debe considerarse que se refiere ampliamente a partículas de cualquier naturaleza que tienen oligonucleótidos liberables y son compatibles con los protocolos descritos, ejemplificados pero no limitados al tipo de partícula utilizada con fines ilustrativos.

Como se usa en el presente documento, el término "ácido nucleico complejo" se refiere a grandes poblaciones de ácidos nucleicos o polinucleótidos no idénticos. En ciertas realizaciones, el ácido nucleico objetivo es ADN genómico; ADN del exoma (un subconjunto de ADN genómico completo enriquecido para secuencias transcritas que contiene el conjunto de exones en un genoma); un exoma (es decir, regiones codificantes de proteínas de un genoma seleccionado por un método de captura o enriquecimiento de exón); un microbioma; una mezcla de genomas de diferentes organismos; una mezcla de genomas de diferentes tipos de células de un organismo; y otras mezclas complejas de ácido nucleico que comprende un gran número de moléculas de ácido nucleico diferentes (los ejemplos incluyen, sin limitación, un microbioma, un xenoinjerto, una biopsia de tumor sólido que comprende células normales y tumorales, etc.), incluidos subconjuntos de los tipos de ácidos nucleicos complejos mencionados anteriormente. En una realización dicho ácido nucleico complejo tiene una secuencia completa que comprende al menos una gigabase (Gb) (un genoma humano diploide comprende aproximadamente 6 Gb de secuencia).

Los ejemplos no limitantes de ácidos nucleicos complejos incluyen "ácidos nucleicos circulantes" (CNA), que son ácidos nucleicos que circulan en la sangre humana u otros fluidos corporales, que incluyen pero no se limitan a líquido linfático, licor, ascitis, leche, orina, heces y lavado bronquial, por ejemplo, y se pueden distinguir como ácidos nucleicos libres de células (CF) o asociados a células (revisado en Pinzani et al., Methods 50: 302-307, 2010), por ejemplo, células fetales en circulación en el torrente sanguíneo de una madre embarazada (véase, por ejemplo, Kavanagh et al., J. Chromatol. B 878: 1905-1911, 2010) o las células tumorales en circulación (CTC) del torrente sanguíneo de un paciente con cáncer (véase, por ejemplo, Allard et al., Clin Cancer Res. 10: 6897-6904, 2004). Otro ejemplo es el ADN genómico de una sola célula o un pequeño número de células, tales como, por ejemplo, de biopsias (por ejemplo, células fetales de biopsias del trofectodermo de un blastocisto; células cancerosas por aspiración con aguja de un tumor sólido; etc.). Otro ejemplo son los patógenos, por ejemplo, células bacterianas, virus u otros patógenos, en un tejido, en la sangre u otros fluidos corporales, etc.

Como se usa en el presente documento, el término "ácido nucleico objetivo" (o polinucleótido) o "ácido nucleico de interés" se refiere a cualquier ácido nucleico (o polinucleótido) adecuado para el procesamiento y secuenciación por los métodos descritos en el presente documento. El ácido nucleico puede ser de cadena sencilla o de cadena doble y puede incluir ADN, ARN u otros ácidos nucleicos conocidos. Los ácidos nucleicos objetivo pueden ser los de cualquier organismo, incluidos, entre otros, virus, bacterias, levaduras, plantas, peces, reptiles, anfibios, aves y mamíferos (incluidos, entre otros, ratones, ratas, perros, gatos, cabras, ovejas, vacas, caballos, cerdos, conejos, monos y otros primates no humanos y humanos). Se puede obtener un ácido nucleico objetivo de un individuo o de múltiples individuos (es decir, una población). Una muestra de la que se obtiene el ácido nucleico puede contener ácidos nucleicos de una mezcla de células o incluso organismos, tales como: una muestra de saliva humana que incluye células humanas y células bacterianas; un xenoinjerto de ratón que incluye células de ratón y células de un tumor humano trasplantado; etc.

Los ácidos nucleicos objetivo pueden no amplificarse o pueden amplificarse mediante cualquier método de amplificación de ácido nucleico adecuado conocido en la técnica. Los ácidos nucleicos objetivo pueden purificarse de acuerdo con métodos conocidos en la técnica para eliminar contaminantes celulares y subcelulares (lípidos, proteínas, carbohidratos, ácidos nucleicos distintos de los que se van a secuenciar, etc.), o pueden no estar purificados, es decir, incluir al menos algunos contaminantes celulares y subcelulares, que incluyen, entre otros, células intactas que se rompen para liberar sus ácidos nucleicos para su procesamiento y secuenciación. Los ácidos nucleicos objetivo se pueden obtener de cualquier muestra adecuada utilizando métodos conocidos en la técnica. Dichas muestras incluyen, pero no se limitan a: tejidos, células aisladas o cultivos celulares, fluidos corporales (incluidos, entre otros, sangre, orina, suero, linfa, saliva, secreciones anales y vaginales, transpiración y semen); muestras de aire, agrícolas, agua y suelo, etc.

Se desea una alta cobertura en la secuenciación de escopeta porque puede superar los errores en la base de llamadas y ensamblaje. Como se usa en el presente documento, para cualquier posición dada en una secuencia ensamblada, el término "redundancia de cobertura de secuencia", "cobertura de secuencia" o simplemente "cobertura" significa el número de lecturas que representa esa posición. Se puede calcular a partir de la longitud del genoma original (G), el número de lecturas (N) y la longitud de lectura promedio (L) como N x L/G. La cobertura también se puede calcular directamente haciendo un recuento de las bases para cada posición de referencia. Para una secuencia de genoma completo, la cobertura se expresa como un promedio para todas las bases en la secuencia ensamblada. La cobertura

de secuencia es el número promedio de veces que se lee una base (como se describió anteriormente). A menudo se expresa como "cobertura de plegado", por ejemplo, como en "cobertura de 40 veces (o 40x)", lo que significa que cada base en la secuencia ensamblada final se representa en un promedio de 40 lecturas.

Como se usa en este documento, el término "tasa de llamada" significa una comparación del porcentaje de bases del ácido nucleico complejo que se llaman completamente, comúnmente con referencia a una secuencia de referencia adecuada tal como, por ejemplo, un genoma de referencia. Por lo tanto, para un genoma humano completo, la "tasa de llamada del genoma" (o simplemente "tasa de llamada") es el porcentaje de las bases del genoma humano que se llaman completamente con referencia a una referencia de genoma humano completo. Una "tasa de llamadas de exoma" es el porcentaje de las bases del exoma que se llaman completamente con referencia a una referencia de exoma. Se puede obtener una secuencia de exoma secuenciando porciones de un genoma que se ha enriquecido mediante varios métodos conocidos que capturan selectivamente regiones genómicas de interés de una muestra de ADN antes de la secuenciación. Alternativamente, se puede obtener una secuencia de exoma secuenciando un genoma humano completo, que incluye secuencias de exoma. Por lo tanto, una secuencia completa del genoma humano puede tener tanto una "tasa de llamada de genoma" como una "tasa de llamada de exoma". También hay una "tasa de llamadas de lectura sin procesar" que refleja el número de bases que obtienen una designación de A/C/G/T en comparación con el número total de bases intentadas (ocasionalmente, el término "cobertura" se usa en lugar de "tasa de llamadas", pero el significado será evidente por el contexto).

Como se usa en este documento, el término "haplotipo" significa una combinación de alelos en ubicaciones adyacentes (loci) en el cromosoma que se transmite junto o, alternativamente, un conjunto de variantes de secuencia en un solo cromosoma de un par de cromosomas que están estadísticamente asociados. Cada individuo humano tiene dos conjuntos de cromosomas, uno paterno y otro materno. Por lo general, la secuenciación del ADN solo da como resultado información genotípica, la secuencia de alelos no ordenados a lo largo de un segmento de ADN. Inferir los haplotipos para un genotipo separa los alelos en cada par desordenado en dos secuencias separadas, cada una llamada haplotipo. La información del haplotipo es necesaria para muchos tipos diferentes de análisis genéticos, incluidos los estudios de asociación de enfermedades y la inferencia sobre los ancestros de la población.

Como se usa en el presente documento, el término "separación por fases" (o resolución) significa clasificar los datos de secuencia en los dos conjuntos de cromosomas o haplotipos parentales. La separación por fases del haplotipo se refiere al problema de recibir como entrada un conjunto de genotipos para un individuo o una población, es decir, más de un individuo, y generar un par de haplotipos para cada individuo, uno paterno y el otro materno. La separación por fases puede implicar la resolución de datos de secuencia sobre una región de un genoma, o tan solo dos variantes de secuencia en una lectura o cóntigo, lo que puede denominarse separación por fases local o microseparación por fases. También puede implicar la separación por fases de cóntigos más largos, que generalmente incluyen más de aproximadamente diez variantes de secuencia, o incluso una secuencia completa del genoma, que puede denominarse "separación por fases universal". Opcionalmente, la separación por fases de variantes de secuencia tienen lugar durante el ensamblaje del genoma.

Como se usa en este documento, el término "transposón" o "elemento transponible" significa una secuencia de ADN que puede cambiar su posición dentro del genoma. En una reacción de transposición clásica, una transposasa cataliza la inserción aleatoria de transposones extirpados en objetivos de ADN. Durante la transposición de cortar y pegar, una transposasa realiza rompimientos aleatorias y escalonadas de doble cadena en el ADN objetivo y une covalentemente el extremo 3’ de la cadena del transposón transferido al extremo 5' del ADN objetivo. El complejo transposasa/transposón inserta una secuencia de ADN arbitraria en el punto de inserción del transposón en el ácido nucleico objetivo. Se prefieren los transposones que se insertan aleatoriamente en la secuencia de ácido nucleico objetivo. Se han descrito y utilizado varios transposones en sistemas de transposición *in vitro*. Por ejemplo, en la tecnología Nextera^{MR} (Nature Methods 6, noviembre de 2009; Epicenter Biotechnologies, Madison, WI), todo el complejo no es necesario para la inserción; los extremos libres del transposón son suficientes para la integración. Cuando se usan los extremos libre del transposón, el ADN objetivo se fragmenta y la cadena transferida del oligonucleótido del extremo del transposón se une covalentemente al extremo 5’ del fragmento objetivo. Los extremos del transposón se pueden modificar mediante la adición de secuencias deseadas, tales como sitios de unión de cebadores de PCR, códigos de barras/etiquetas, etc. La distribución del tamaño de los fragmentos se puede controlar cambiando las cantidades de extremos de transposasa y transposón. La explotación de los extremos del transposón con secuencias adjuntas que dan como resultado bibliotecas de ADN que se pueden usar en la secuenciación de alto rendimiento. Los extremos del transposón pueden variar en longitud, pero generalmente tienen de 9 a 40 bases de largo. Los pares de extremos del transposón pueden ser complementos invertidos entre sí (es decir, los extremos del transposón pueden ser repeticiones terminales invertidas).

Como se usa en el presente documento, el término "horquilla" (también conocido como bucle de tallo) tiene su significado normal en la técnica y se refiere a una conformación de ácido nucleico en la que dos regiones de la misma cadena, generalmente complementarias en la secuencia de nucleótidos cuando se lee en direcciones opuestas, se aparean las bases para formar una doble hélice que termina en un bucle no emparejado.

En algunas realizaciones, los métodos de la presente invención se realizan en dispositivos de microfluidos.

Una reducción de los volúmenes hasta los niveles de pico litros puede lograr una reducción aún mayor en los reactivos

y los costos computacionales. En algunas realizaciones, este nivel de reducción de costos se logra mediante la combinación del proceso de MT con dispositivos de tipo de microfluidos. La capacidad de realizar todos las etapas enzimáticas en la misma reacción sin purificación de ADN facilita la capacidad de miniaturizar y automatizar este proceso y da como resultado la adaptabilidad a una amplia variedad de plataformas y métodos de preparación de muestras.

Estudios recientes también han sugerido una mejora en el sesgo de GC después de la amplificación y una reducción en la amplificación de fondo al disminuir los volúmenes de reacción hasta el tamaño de nano litros.

Actualmente hay varios tipos de dispositivos de microfluidos (por ejemplo, dispositivos vendidos por Advanced Liquid Logic, Morrisville, NC) o pico/nano-gotas (por ejemplo, RainDance Technologies, Lexington, MA) que tienen pico/nano-gotas, que realizan fusión (3000/segundo) y funciones de recolección y podrían usarse en tales realizaciones de MT.

Amplificación

De acuerdo con una realización, el proceso de MT comienza con un tratamiento corto de ADN genómico con una exonucleasa 5’ para crear salientes de cadena sencilla 3' que sirven como sitios de inicio de MDA. El uso de la exonucleasa elimina la necesidad de una etapa de desnaturalización alcalina o térmica antes de la amplificación sin introducir sesgo en la población de fragmentos. La desnaturalización alcalina se puede combinar con el tratamiento con exonucleasa 5’, lo que da como resultado una reducción adicional del sesgo. Los fragmentos se amplifican, por ejemplo, usando un método de MDA. En ciertas realizaciones, la reacción de MDA es una reacción de amplificación basada en la polimerasa phi29 modificada, aunque se puede usar otro método de amplificación conocido.

En algunas realizaciones, la reacción de MDA está diseñada para introducir uracilos en los productos de amplificación. En algunas realizaciones, se usa una reacción de MDA estándar que utiliza hexámeros aleatorios para amplificar los fragmentos en cada pozo. En muchas realizaciones, en lugar de los hexámeros aleatorios, se usan cebadores aleatorios de 8 mer para reducir el sesgo de amplificación en la población de fragmentos. En realizaciones adicionales, también se pueden agregar varias enzimas diferentes a la reacción de MDA para reducir el sesgo de la amplificación. Por ejemplo, pueden usarse bajas concentraciones de exonucleasas 5’ no procesadoras y/o proteínas de unión de cadena sencilla para crear sitios de unión para los 8 mer. Los agentes químicos como la betaína, DMSO y trehalosa también se pueden usar para reducir el sesgo a través de mecanismos similares.

Fragmentación

De acuerdo con una realización, después de la amplificación del ADN de ADN, el producto de amplificación, o amplicones, se somete a una ronda de fragmentación. En algunas realizaciones, el método CoRE se usa para fragmentar más los fragmentos en cada pozo después de la amplificación. Para usar el método CoRE, la reacción de MDA utilizada para amplificar los fragmentos en cada pozo está diseñada para incorporar uracilos en los productos de MDA. La fragmentación de los productos de MDA también se puede lograr mediante sonicación o tratamiento enzimático.

Si se usa un método CoRE para fragmentar los productos de MDA, el ADN amplificado se trata con una mezcla de uracil ADN glicosilasa (UDG), ADN glicosilasa-liasa endonucleasa VIII y polinucleótido quinasa T4 para escindir las bases de uracilo y crear huecos de una sola base con grupos funcionales fosfato 5’ e hidroxilo 3'. La traducción de corte mediante el uso de una polimerasa como la polimerasa Taq da como resultado rompimientos de extremos romos de cadena doble, lo que da como resultado fragmentos ligables de un intervalo de tamaño que depende de la concentración de dUTP añadida en la reacción de MDA. En algunas realizaciones, el método CoRE utilizado implica la eliminación de uracilos por polimerización y el desplazamiento de la cadena por phi29.

Después de la fragmentación de los productos de MDA, los extremos de los fragmentos resultantes pueden repararse. Dichas reparaciones pueden ser necesarias, porque muchas técnicas de fragmentación pueden dar como resultado terminales con extremos sobresalientes y terminales con grupos funcionales que no son útiles en reacciones de ligadura posteriores, tales como grupos hidroxilo 3’ y 5' y/o grupos fosfato 3’ y 5'. En muchos aspectos de la presente invención, es útil tener fragmentos reparados para que tengan extremos romos, y en algunos casos, puede ser deseable alterar la química de los terminales de manera que la orientación correcta de los grupos fosfato e hidroxilo no este presente, evitando así la "polimerización" de las secuencias objetivo. El control sobre la química de los terminales puede proporcionarse usando métodos conocidos en la técnica. Por ejemplo, en algunas circunstancias, el uso de fosfatasa elimina todos los grupos fosfato, de modo que todos los extremos contienen grupos hidroxilo. Cada extremo puede ser alterado selectivamente para permitir la ligadura entre los componentes deseados. Un extremo de los fragmentos puede entonces "activarse", en algunas realizaciones por tratamiento con fosfatasa alcalina.

MT utilizando un pequeño número de células como la fuente de ácidos nucleicos complejos

De acuerdo con una realización, se usa un método de MT para analizar el genoma de una célula individual o un pequeño número de células (o un número similar de núcleos aislados de las células). El proceso para aislar el ADN en este caso es similar a los métodos descritos anteriormente, pero puede ocurrir en un volumen más pequeño.

Como se discutió anteriormente, el aislamiento de fragmentos largos de ácido nucleico genómico de una célula se

puede lograr mediante varios métodos diferentes. En una realización, las células se lisan y el núcleo intacto se sedimenta con una etapa de centrifugación suave. El ADN genómico se libera a través de la digestión con proteinasa K y RNasa durante varias horas. El material puede entonces tratarse en algunas realizaciones para reducir la concentración de residuos celulares restantes; tales tratamientos son bien conocidos en la técnica y pueden incluir, sin limitación, diálisis durante un período de tiempo (por ejemplo, de 2 a 16 horas) y/o dilución. Dado que dichos métodos para aislar el ácido nucleico no implican muchos procesos disruptivos (tales como la precipitación con etanol, la centrifugación y la agitación vorticial), el ácido nucleico genómico permanece en gran parte intacto, produciendo una mayoría de fragmentos que tienen longitudes superiores a 150 kilobases. En algunas realizaciones, los fragmentos son de aproximadamente 100 a aproximadamente 750 kilobases de longitud. En realizaciones adicionales, los fragmentos tienen una longitud de aproximadamente 150 a aproximadamente 600, aproximadamente 200 a aproximadamente 500, aproximadamente 250 a aproximadamente 400 y aproximadamente 300 a aproximadamente 350 kilobases.

Una vez aislado, el ADN genómico puede fragmentarse cuidadosamente para evitar la pérdida de material, particularmente para evitar la pérdida de secuencia de los extremos de cada fragmento, ya que la pérdida de dicho material dará como resultado huecos en el ensamblaje final del genoma. En algunos casos, la pérdida de secuencia se evita mediante el uso de una enzima de corte poco frecuente, que crea sitios de partida para una polimerasa, tal como la polimerasa phi29, a distancias de aproximadamente 100 kb entre sí. A medida que la polimerasa crea la nueva cadena de ADN, desplaza la cadena anterior, con el resultado final de que hay secuencias superpuestas cerca de los sitios de inicio de la polimerasa, lo que resulta en muy pocas eliminaciones de secuencias.

En algunas realizaciones, un uso controlado de una exonucleasa 5' (ya sea antes o durante la reacción de MDA) puede promover múltiples replicaciones del ADN original de la célula única y minimizar así la propagación de errores tempranos mediante la copia de copias.

En un aspecto, los métodos de la presente invención producen datos genómicos de calidad a partir de células individuales. Suponiendo que no hay pérdida de ADN, hay un beneficio al comenzar con un número bajo de células (10 o menos) en lugar de usar una cantidad equivalente de ADN de una preparación grande. Comenzar con menos de 10 células asegura una cobertura uniforme en fragmentos largos de cualquier región determinada del genoma. Comenzar con cinco o menos células permite una cobertura cuatro veces o mayor por cada fragmento de ADN de 100 kb sin aumentar el número total de lecturas por encima de 120 Gb (20 veces la cobertura de un genoma diploide de 6 Gb). Sin embargo, una gran cantidad de fragmentos de ADN más largos (100 kb o más) son aún más beneficiosos para la secuenciación de unas pocas células, porque para cualquier secuencia dada existen solo tantos fragmentos de superposición como el número de células iniciales y la aparición de fragmentos de superposición de ambos cromosomas parentales puede ser una pérdida sustancial de información.

La primera etapa en MT es generalmente una amplificación de genoma completo de bajo sesgo, que puede ser de uso particular en el análisis genómico de células individuales. Debido a los rompimientos de la cadena de ADN y las pérdidas de ADN en el manejo, incluso los métodos de secuenciación de una sola molécula probablemente requerirían cierto nivel de amplificación de ADN de la célula única. La dificultad para secuenciar células individuales proviene de intentar amplificar todo el genoma. Los estudios realizados en bacterias que usan MDA han sufrido la pérdida de aproximadamente la mitad del genoma en la secuencia final ensamblada con una cantidad bastante alta de variación en la cobertura a través de esas regiones secuenciadas. Esto puede explicarse parcialmente como resultado de que el ADN genómico inicial tiene cortes y rompimientos de cadena que no pueden replicarse en los extremos y, por lo tanto, se pierden durante el proceso de MDA. MT proporciona una solución a este problema mediante la creación de fragmentos largos superpuestos del genoma antes de la MDA. Para lograr esto, se utiliza un proceso suave para aislar el ADN genómico de la célula. El ADN genómico en gran parte intacto puede tratarse ligeramente con una nickasa frecuente, lo que da como resultado un genoma cortado en forma semialeatoria. La capacidad de phi29 de desplazamiento de la cadena se usa luego para polimerizar a partir de los cortes creando fragmentos de superposición muy largos (> 200 kb). Estos fragmentos se utilizan luego como plantilla de inicio para MT.

Análisis de metilación utilizando MT

En otro aspecto, los métodos de la presente invención se usan para el análisis de metilación genómica. Actualmente hay varios métodos disponibles para el análisis de metilación genómica global. Un método implica el tratamiento con bisulfato de ADN genómico y la secuenciación de elementos repetitivos o una fracción del genoma obtenida por fragmentación de la enzima de restricción específica de metilación. Esta técnica proporciona información sobre la metilación total, pero no proporciona datos específicos del locus. El siguiente nivel más alto de resolución utiliza matrices de ADN y está limitado por la cantidad de características en el chip. Finalmente, la resolución más alta y el enfoque más costoso requieren tratamiento con bisulfato seguido de secuenciación de todo el genoma. Usando MT es posible secuenciar todas las bases del genoma y ensamblar un genoma diploide completo con información digital sobre los niveles de metilación para cada posición de citosina en el genoma humano (es decir, secuenciación de 5 bases). Además, MT permite que bloques de secuencia metilada de 100 kb o más se unan a haplotipos de secuencia, proporcionando haplotipificación de metilación, información que es imposible de lograr con cualquier método disponible actualmente.

En un ejemplo de realización no limitante, el estado de metilación se obtiene en un método en el que el ADN genómico

se desnaturaliza primero para MDA. A continuación, el ADN se trata con bisulfito (una etapa que requiere ADN desnaturalizado). La preparación restante sigue los métodos descritos, por ejemplo, en las solicitudes de los Estados Unidos Nos. 11/451.692, presentada el 13/06/2006 (publicada como US 2007/0072208) y 12/335.168, presentada el 15/12/2008 (publicada como US 2009/0311691), enseñanzas relacionadas con el análisis de ácido nucleico de mezclas de fragmentos de acuerdo con técnicas de lectura de fragmentos largos son particularmente relevantes.

En un aspecto, la MDA amplificará cada cadena de un fragmento específico, produciendo independientemente para cualquier posición de citosina dada, el 50% de las lecturas como no afectadas por el bisulfito (es decir, la base opuesta a la citosina, una guanina no se ve afectada por el bisulfato) y 50% que proporciona el estado de metilación. La complejidad reducida del ADN ayuda con el mapeo preciso y el ensamblaje de las lecturas menos informativas, en su mayoría de 3 bases (A, T, G).

Se ha informado que el tratamiento con bisulfito fragmenta el ADN. Sin embargo, la titulación cuidadosa de la desnaturalización y los tampones de bisulfato pueden evitar la fragmentación excesiva del ADN genómico. Se puede tolerar una conversión del 50% de la citosina en uracilo en MT, lo que permite una reducción de la exposición del ADN al bisulfito para minimizar la fragmentación. En algunas realizaciones, es aceptable cierto grado de fragmentación ya que no afectaría la haplotipificación.

Uso de MT para el análisis de genomas de cáncer

Se ha sugerido que más del 90% de los cánceres albergan pérdidas o ganancias significativas en regiones del genoma humano, denominadas aneuploidía, y se ha observado que algunos cánceres individuales contienen más de cuatro copias de algunos cromosomas. Esta mayor complejidad en el número de copias de cromosomas y regiones dentro de los cromosomas hace que la secuenciación de los genomas del cáncer sea sustancialmente más difícil. La capacidad de las técnicas de MT para secuenciar y ensamblar fragmentos muy largos (> 100 kb) del genoma lo hace muy adecuado para la secuenciación de genomas completos de cáncer.

Reducción de errores mediante secuenciación de un ácido nucleico objetivo

De acuerdo con una realización, incluso si la separación por fases basada en MT no se realiza y se usa un enfoque de secuenciación estándar, un ácido nucleico objetivo se fragmenta (si es necesario), y los fragmentos se etiquetan antes de la amplificación. Una ventaja de MT es que los errores introducidos como resultado de la amplificación (u otras etapas) se pueden identificar y corregir comparando la secuencia obtenida de múltiples fragmentos largos superpuestos. Por ejemplo, una llamada de base (por ejemplo, identificar una base particular tal como A, C, G o T) en una posición particular (por ejemplo, con respecto a una referencia) de los datos de secuencia puede aceptarse como verdadera si la llamada de base está presente en datos de secuencia de dos o más fragmentos largos (u otro número de umbral), o en una mayoría sustancial de fragmentos largos (por ejemplo, en al menos 51, 60, 70 u 80 por ciento), en los que el denominador puede restringirse a los fragmentos que tienen una llamada de base en la posición particular. Una llamada de base puede incluir cambiar un alelo de un het o het potencial. Una llamada de base en la posición particular puede aceptarse como falsa si está presente en un solo fragmento largo (u otro número de umbral de fragmentos largos), o en una minoría sustancial de fragmentos largos (por ejemplo, menos de 10, 5 o 3 fragmentos o como medida con un número relativo, tal como 20 o 10 por ciento). Los valores de umbral pueden predeterminarse o determinarse dinámicamente en función de los datos de secuenciación. Una llamada de base en la posición particular puede convertirse/aceptarse como "no llamada" si no está presente en una minoría sustancial y en una mayoría sustancial de fragmentos esperados (por ejemplo, en 40-60 por ciento). En algunas realizaciones e implementaciones, pueden usarse diversos parámetros (por ejemplo, en distribución, probabilidad y/u otras funciones o estadísticas) para caracterizar lo que puede considerarse una minoría sustancial o una mayoría sustancial de fragmentos. Los ejemplos de tales parámetros incluyen, sin limitación, uno o más de: número de llamadas de base que identifican una base particular; cobertura o número total de bases llamadas en una posición particular; número y/o identidades de fragmentos distintos que dieron lugar a datos de secuencia que incluyen una llamada de base particular; número total de fragmentos distintos que dieron lugar a datos de secuencia que incluyen al menos una llamada de base en una posición particular; la base de referencia en la posición particular; y otros. En una realización, una combinación de los parámetros anteriores para una llamada de base particular se puede ingresar a una función para determinar un puntaje (por ejemplo, una probabilidad) para la llamada de base particular. Los puntajes se pueden comparar con uno o más valores de umbral como parte de la determinación de si se acepta una llamada de base (por ejemplo, por encima de un umbral), en error (por ejemplo, por debajo de un umbral) o una no llamada (por ejemplo, si todo los puntajes para las llamadas de base están por debajo de un umbral). La determinación de una llamada de base puede depender de los puntajes de las otras llamadas de base.

Como un ejemplo básico, si se encuentra una llamada de base de A en más del 35% (un ejemplo de un puntaje) de los fragmentos que contienen una lectura para la posición de interés y se encuentra una llamada de base de C en más del 35% de estos fragmentos y las otras llamadas de base tienen un puntaje de menos del 20%, entonces la posición puede considerarse una het compuesta de A y C, posiblemente sometida a otros criterios (por ejemplo, un número mínimo de fragmentos que contienen una lectura en la posición de interés). Por lo tanto, cada uno de los puntajes puede ingresarse en otra función (por ejemplo, heurística, que puede usar lógica comparativa o difusa) para proporcionar la determinación final de la o las llamadas de base para la posición.

Como otro ejemplo, un número específico de fragmentos que contienen una llamada de base se puede usar como umbral. Por ejemplo, al analizar una muestra de cáncer, puede haber mutaciones somáticas de baja prevalencia. En tal caso, la llamada de base puede aparecer en menos del 10% de los fragmentos que cubren la posición, pero la llamada de base aún puede considerarse correcta, posiblemente sujeta a otros criterios. Por lo tanto, varias realizaciones pueden usar números absolutos o números relativos, o ambos (por ejemplo, como entradas en lógica comparativa o difusa). Y, tales números de fragmentos se pueden ingresar en una función (como se mencionó anteriormente), así como los umbrales correspondientes a cada número, y la función puede proporcionar un puntaje, que también se puede comparar con uno o más umbrales para hacer una determinación final en cuanto a la llamada de base en la posición particular.

Otro ejemplo de una función de corrección de errores se refiere a errores de secuencia en lecturas sin procesar que conducen a una supuesta llamada variante inconsistente con otras llamadas variantes y sus haplotipos. Si se encuentran 20 lecturas de la variante A en 9 y 8 fragmentos pertenecientes a los respectivos haplotipos y 7 lecturas de la variante G se encuentran en 6 pozos (5 o 6 de los cuales se comparten con fragmentos con lecturas de A), la lógica puede rechazar la variante G como un error de secuenciación porque para el genoma diploide solo una variante puede residir en una posición en cada haplotipo. La variante A está apoyada con sustancialmente más lecturas, y las lecturas de G siguen sustancialmente fragmentos de lecturas de A, lo que indica que es más probable que se generen al leer incorrectamente G en lugar de A. Si las lecturas de G están casi exclusivamente en fragmentos separados de A, esto puede indicar que las lecturas G están mal mapeadas o provienen de un ADN contaminante.

Identificación de expansiones en regiones con repeticiones en tándem cortas

Una repetición en tándem corta (STR) en el ADN es un segmento de ADN con un patrón periódico fuerte. Las STR se producen cuando se repite un patrón de dos o más nucleótidos y las secuencias repetidas son directamente adyacentes entre sí; las repeticiones pueden ser perfectas o imperfectas, es decir, puede haber algunos pares de bases que no coinciden con el motivo periódico. El patrón generalmente varía en longitud de 2 a 5 pares de bases (pb). Las STR normalmente se encuentran en regiones no codificantes, por ejemplo, en intrones. Se produce un polimorfismo de repetición en tándem corto (STRP) cuando los loci de STR homólogos difieren en el número de repeticiones entre individuos. El análisis de STR a menudo se usa para determinar los perfiles genéticos con fines forenses. Las STR que se producen en los exones de los genes pueden representar regiones hipermutables que están relacionadas con enfermedades humanas (Madsen et al., BMC Genomics 9: 410, 2008).

En genomas humanos (y genomas de otros organismos) las STR incluyen repeticiones de trinucleótidos, por ejemplo, repeticiones de CTG o CAG. La expansión de repetición de trinucleótidos, también conocida como expansión de repetición de triplete, es causada por el deslizamiento durante la replicación del ADN y está asociada con ciertas enfermedades clasificadas como trastornos de repetición de trinucleótidos, tales como la enfermedad de Huntington. En general, cuanto mayor es la expansión, mayor es la probabilidad de causar enfermedad o aumentar la gravedad de la enfermedad. Esta propiedad da como resultado la característica de "anticipación" que se observa en los trastornos por repetición de trinucleótidos, es decir, la tendencia de la edad de aparición de la enfermedad a disminuir y la gravedad de los síntomas a aumentar a través de generaciones sucesivas de una familia afectada debido a la expansión de estas repeticiones. La identificación de expansiones en las repeticiones de trinucleótidos puede ser útil para predecir con precisión la edad de inicio y la progresión de la enfermedad para los trastornos de repetición de trinucleótidos.

La expansión de las STR como repeticiones de trinucleótidos puede ser difícil de identificar usando métodos de secuenciación de próxima generación. Tales expansiones pueden no mapearse y pueden faltar o estar subrepresentadas en bibliotecas. Usando MT, es posible ver una caída significativa en la cobertura de secuencia en una región de STR. Por ejemplo, una región con las STR característicamente tendrá un nivel de cobertura más bajo en comparación con las regiones sin tales repeticiones, y habrá una caída sustancial en la cobertura en esa región si hay una expansión de la región, observable en un gráfico de cobertura versus la posición en el genoma.

Por ejemplo, si la cobertura de secuencia es de aproximadamente 20 en promedio, la región con la región de expansión tendrá una caída significativa, por ejemplo, a 10 si el haplotipo afectado tiene cobertura cero en la región de expansión. Por lo tanto, se produciría una caída del 50%. Sin embargo, si se compara la cobertura de secuencia para los dos haplotipos, la cobertura es 10 en el haplotipo normal y 0 en el haplotipo afectado, que es una caída de 10 pero una caída porcentual general del 100%. O bien, se pueden analizar las cantidades relativas, que son 2:1 (normal frente a cobertura en la región de expansión) para la cobertura de secuencia combinada, pero es 10:0 (haplotipo 1 frente a haplotipo 2), que es infinito o cero (dependiendo sobre cómo se forma la relación) y, por lo tanto, una gran distinción.

Uso diagnóstico de datos de secuencia

Los datos de secuencia generados usando los métodos de la presente invención son útiles para una amplia variedad de propósitos. De acuerdo con una realización, los métodos de secuenciación de la presente invención se usan para identificar una variación de secuencia en una secuencia de un ácido nucleico complejo, por ejemplo, una secuencia completa del genoma, que es informativa sobre el estado característico o médico de un paciente o de un embrión o feto, tal como el sexo de un embrión o feto o la presencia o el pronóstico de una enfermedad que tiene un componente genético, que incluye, por ejemplo, fibrosis quística, anemia falciforme, síndrome de Marfan, enfermedad de

Huntington y hemocromatosis o varios tipos de cáncer, tales como como cáncer de mama, por ejemplo. De acuerdo con otra realización, los métodos de secuenciación de la presente invención se usan para proporcionar información de secuencia que comienza con entre una y 20 células de un paciente (que incluye pero no se limita a un feto o un embrión) y la evaluación de una característica del paciente con base en la secuencia.

Diagnóstico de cáncer

La secuenciación del genoma completo es una herramienta valiosa para evaluar la base genética de la enfermedad. Se conocen varias enfermedades para las cuales existe una base genética, por ejemplo, fibrosis quística.

Una aplicación de la secuenciación del genoma completo es comprender el cáncer. El impacto más significativo de la secuenciación de próxima generación en la genómica del cáncer ha sido la capacidad de volver a secuenciar, analizar y comparar el tumor compatible y los genomas normales de un solo paciente, así como muestras de múltiples pacientes de un tipo de cáncer dado. Al utilizar la secuenciación del genoma completo, se puede considerar todo el espectro de variaciones de secuencia, incluidos los loci de susceptibilidad de la línea germinal, los polimorfismos somáticos de un solo nucleótido (SNP), las mutaciones de inserción y eliminación (indel) pequeñas, las variaciones del número de copias (CNV) y las variantes estructurales (SV).

En general, el genoma del cáncer está compuesto por el ADN de la línea germinal del paciente, sobre el cual se han superpuesto las alteraciones genómicas somáticas. Las mutaciones somáticas identificadas por secuenciación pueden clasificarse como mutaciones "conductor" o "pasajero". Las llamadas mutaciones conductor son aquellas que contribuyen directamente a la progresión del tumor al conferir una ventaja de crecimiento o supervivencia a la célula. Las mutaciones pasajero abarcan mutaciones somáticas neutrales que se han adquirido durante los errores en la división celular, la replicación del ADN y la reparación; estas mutaciones pueden adquirirse mientras la célula es fenotípicamente normal, o después de la evidencia de un cambio neoplásico.

Históricamente, se han hecho intentos para dilucidar el mecanismo molecular del cáncer, y se han identificado varias mutaciones "conductor" o biomarcadores, tales como HER2/neu2. Sobre la base de dichos genes, se han desarrollado regímenes terapéuticos para atacar específicamente los tumores con alteraciones genéticas conocidas. El ejemplo mejor definido de este enfoque es el ataque de HER2/neu en células de cáncer de mama por trastuzumab (Herceptina). Sin embargo, los cánceres no son simples enfermedades monogenéticas, sino que se caracterizan por combinaciones de alteraciones genéticas que pueden diferir entre los individuos. En consecuencia, estas perturbaciones adicionales en el genoma pueden hacer que algunos regímenes farmacológicos sean ineficaces para ciertos individuos.

Las células cancerosas para la secuenciación del genoma completo se pueden obtener a partir de biopsias de tumores completos (incluyendo microbiopsias de un pequeño número de células), células cancerosas aisladas del torrente sanguíneo u otros fluidos corporales de un paciente, o cualquier otra fuente conocida en la técnica.

Diagnostico genético de implantación previa

Una aplicación de los métodos de la presente invención es para el diagnóstico genético previo a la implantación. Alrededor del 2 al 3% de los bebés nacidos tienen algún tipo de defecto de nacimiento importante. El riesgo de algunos problemas, debido a la separación anormal del material genético (cromosomas), aumenta con la edad de la madre. Alrededor del 50% del tiempo, este tipo de problemas se debe al síndrome de Down, que es una tercera copia del cromosoma 21 (trisomía 21). La otra mitad es el resultado de otros tipos de anomalías cromosómicas, incluidas las trisomías, mutaciones puntuales, variaciones estructurales, variaciones en el número de copias, etc. Muchos de estos problemas cromosómicos provocan un bebé gravemente afectado o uno que no sobrevive ni siquiera al parto.

En medicina y genética (clínica), el diagnóstico genético previo a la implantación (PGD o PIGD) (también conocido como cribado de embriones) se refiere a los procedimientos que se realizan en embriones antes de la implantación, a veces incluso en los ovocitos antes de la fertilización. PGD puede permitir a los padres evitar la interrupción selectiva del embarazo. El término cribado genético previo a la implantación (PGS) se utiliza para denotar procedimientos que no buscan una enfermedad específica pero utilizan técnicas de PGD para identificar embriones en riesgo, por ejemplo, debido a una condición genética que podría conducir a una enfermedad. En cambio, los procedimientos realizados en las células sexuales antes de la fertilización pueden denominarse métodos de selección de ovocitos o selección de esperma, aunque los métodos y objetivos se superponen en parte con PGD.

El perfil genético previo a la implantación (PGP) es un método de tecnología de reproducción asistida para realizar la selección de embriones que parecen tener las mayores posibilidades de un embarazo exitoso. Cuando se usa para mujeres de edad materna avanzada y para pacientes con fallas repetitivas de fertilización *in vitro* (IVF), la PGP se lleva a cabo principalmente como un cribado para la detección de anomalías cromosómicas, tal como aneuploidía, translocaciones recíprocas y robertsonianas, y otras anomalías, tales como las inversiones o eliminaciones cromosómicas. Además, PGP puede examinar los marcadores genéticos en busca de características, incluidos varios estados de enfermedad. El principio detrás del uso de PGP es que, dado que se sabe que las anomalías cromosómicas numéricas explican la mayoría de los casos de pérdida de embarazo y una gran proporción de los embriones humanos son aneuploides, el reemplazo selectivo de embriones euploides debería aumentar las posibilidades de un tratamiento exitoso de IVF. La secuenciación del genoma completo proporciona una alternativa a dichos métodos de métodos integrales de análisis cromosómico, tal como la hibridación genómica comparativa de matrices (aCGH), PCR

cuantitativa y microarreglos SNP. La secuenciación completa del genoma completo puede proporcionar información sobre cambios de bases individuales, inserciones, eliminaciones, variaciones estructurales y variaciones de número de copias, por ejemplo.

Como PGD se puede realizar en células de diferentes etapas de desarrollo, los procedimientos de biopsia varían en consecuencia. La biopsia se puede realizar en todas las etapas previas a la implantación, incluidos, entre otros, ovocitos no fertilizados y fertilizados (para cuerpos polares, PB), en embriones en la etapa de escisión del día tres (para blastómeros) y en blastocistos (para células de trofectodermo).

Sistemas de secuenciación y análisis de datos

En algunas realizaciones, la secuenciación de muestras de ADN (por ejemplo, tales como muestras que representan genomas humanos completos) puede realizarse mediante un sistema de secuenciación. En la Figura 5 se ilustran dos ejemplos de sistemas de secuenciación.

Las Figuras 5A y 5B son diagramas de bloque de ejemplos de sistemas 190 de secuenciación que están configurados para realizar las técnicas y/o métodos para el análisis de secuencias de ácido nucleico de acuerdo con las realizaciones descritas en el presente documento. Un sistema 190 de secuenciación puede incluir o estar asociado con múltiples subsistemas tales como, por ejemplo, una o más máquinas de secuenciación como la máquina 191 de secuenciación, uno o más sistemas informáticos tales como el sistema 197 informático, y uno o más depósitos de datos tales como el depósito 195 de datos. En el caso ilustrado en la Figura 5A, los diversos subsistemas del sistema 190 pueden estar conectados comunicativamente a través de una o más redes 193, que pueden incluir conmutación de paquetes u otros tipos de dispositivos de infraestructura de red (por ejemplo, enrutadores, conmutadores, etc.) que están configurados para facilitar el intercambio de información entre sistemas remotos. En el caso ilustrado en la Figura 5B, el sistema 190 de secuenciación es un dispositivo de secuenciación en el que los diversos subsistemas (por ejemplo, tales como la máquina o máquinas 191 de secuenciación, el sistema o sistemas 197 informáticos y posiblemente un depósito 195 de datos) son componentes que están comunicativa y/o operativamente acoplados e integrados dentro del dispositivo de secuenciación.

En algunos contextos operativos, el depósito 195 de datos y/o el sistema o sistemas 197 informáticos de los casos ilustrados en las Figuras 5A y 5B pueden configurarse dentro de un entorno 196 informático en la nube. En un entorno informático en la nube, los dispositivos de almacenamiento que comprenden un depósito de datos y/o los dispositivos informáticos que comprenden un sistema informático pueden asignarse e instanciarse para su uso como utilidad y bajo demanda; por lo tanto, el entorno informático en la nube proporciona como servicios la infraestructura (por ejemplo, máquinas físicas y virtuales, almacenamiento en bruto/en bloque, cortafuegos, equilibradores de carga, agregadores, redes, grupos de almacenamiento, etc.), las plataformas (por ejemplo, un dispositivo informático y/o una pila de soluciones que puede incluir un sistema operativo, un entorno de ejecución de lenguaje de programación, un servidor de base de datos, un servidor web, un servidor de aplicaciones, etc.) y el software (por ejemplo, aplicaciones, interfaces de programación de aplicaciones o API, etc.) necesarios para realizar cualquier tarea relacionada con el almacenamiento y/o informática.

Se observa que en varios casos, las técnicas descritas en el presente documento pueden realizarse mediante diversos sistemas y dispositivos que incluyen algunos o todos los subsistemas y componentes anteriores (por ejemplo, tales como máquinas de secuenciación, sistemas informáticos y depósitos de datos) en diversas configuraciones y factores de forma; así, los ejemplos de casos y configuraciones ilustrados en las Figuras 5A y 5B deben considerarse en un sentido ilustrativo más que restrictivo.

La máquina 191 de secuenciación está configurada y funciona para recibir ácidos nucleicos 192 objetivo derivados de fragmentos de una muestra biológica, y para realizar la secuenciación en los ácidos nucleicos objetivo. Se puede usar cualquier máquina adecuada que pueda realizar la secuenciación, en la que dicha máquina puede usar varias técnicas de secuenciación que incluyen, sin limitación, secuenciación por hibridación, secuenciación por ligadura, secuenciación por síntesis, secuenciación de molécula única, detección de secuencia óptica, detección electromagnética de secuencia, detección de secuencia por cambio de voltaje y cualquier otra técnica conocida o desarrollada posteriormente que sea adecuada para generar lecturas de secuenciación a partir de ADN. En varios casos, una máquina de secuenciación puede secuenciar los ácidos nucleicos objetivo y puede generar lecturas de secuenciación que pueden incluir o no huecos y que pueden o no ser lecturas de pares de empalme (o extremos apareados). Como se ilustra en las Figuras 5A y 5B, la máquina 191 de secuenciación secuencia ácidos nucleicos 192 objetivo y obtiene lecturas 194 de secuenciación, que se transmiten para el almacenamiento (temporal y/o persistente) a uno o más depósitos 195 de datos y/o para el procesamiento por uno o más sistemas197 informáticos.

El depósito 195 de datos puede implementarse en uno o más dispositivos de almacenamiento (por ejemplo, unidades de disco duro, discos ópticos, unidades de estado sólido, etc.) que pueden configurarse como una matriz de discos (por ejemplo, tal como una matriz SCSI), un grupo de almacenamiento o cualquier otra organización de dispositivos de almacenamiento adecuada. El o los dispositivos de almacenamiento de un depósito de datos se pueden configurar como componentes internos/integrales del sistema 190 o como componentes externos (por ejemplo, discos duros externos o matrices de discos) conectables al sistema 190 (por ejemplo, como se ilustra en la Figura 5B) y/o pueden estar interconectados comunicativamente de una manera adecuada, tal como, por ejemplo, una red, un grupo de

almacenamiento, una red de área de almacenamiento (SAN) y/o un almacenamiento conectado a la red (NAS) (por ejemplo, como se ilustra en la Figura 5A). En varios casos e implementaciones, se puede implementar un depósito de datos en los dispositivos de almacenamiento como uno o más sistemas de archivos que almacenan información como archivos, tal como una o más bases de datos que almacenan información en registros de datos, y/o cualquier otra organización de almacenamiento de datos adecuada.

El sistema 197 informático puede incluir uno o más dispositivos informáticos que comprenden procesadores de propósito general (por ejemplo, unidades centrales de procesamiento o CPU), memoria y lógica 199 informática que, junto con los datos de configuración y/o el software del sistema operativo (OS), puede realizar algunas o todas las técnicas y métodos descritos en este documento, y/o puede controlar el funcionamiento de la máquina de secuenciación 191. Por ejemplo, cualquiera de los métodos descritos en este documento (por ejemplo, para corrección de errores, separación por fases del haplotipo, etc.) puede ser realizado total o parcialmente por un dispositivo informático que incluye un procesador que se puede configurar para ejecutar la lógica 199 para realizar varias etapas de los métodos. Además, aunque las etapas del método pueden presentarse como etapas numeradas, se entiende que las etapas de los métodos descritos en el presente documento se pueden realizar al mismo tiempo (por ejemplo, en paralelo por un grupo de dispositivos informáticos) o en un orden diferente. Las funcionalidades de la lógica 199 informática pueden implementarse como un único módulo integrado (por ejemplo, en una lógica integrada) o pueden combinarse en dos o más módulos de software que pueden proporcionar algunas funcionalidades adicionales.

En algunos casos, el sistema 197 informático puede ser un único dispositivo informático. En otros casos, el sistema 197 informático puede comprender múltiples dispositivos informáticos que pueden estar interconectados de forma comunicativa y/u operativa en una cuadrícula, un grupo o en un entorno informático en la nube. Dichos dispositivos informáticos múltiples pueden configurarse en diferentes factores de forma, tales como nodos informáticos, blades o cualquier otra configuración de hardware adecuada. Por estas razones, el sistema 197 informático en las Figuras 5A y 5B debe considerarse en un sentido ilustrativo más que restrictivo.

La Figura 6 es un diagrama de bloques de un ejemplo de dispositivo 200 informático que puede configurarse para ejecutar instrucciones para realizar diversas funciones de procesamiento y/o control de datos como parte de una máquina o máquinas o un sistema o sistemas de secuenciación.

En la Figura 6, el dispositivo 200 informático comprende varios componentes que están interconectados directa o indirectamente a través de uno o más buses del sistema tales como el bus 275. Dichos componentes pueden incluir, entre otros, un teclado 278, un dispositivo o dispositivos 279 de almacenamiento persistente (por ejemplo, tal como discos fijos, discos de estado sólido, discos ópticos y similares) y el adaptador 282 de pantalla al que uno o más dispositivos de pantalla (por ejemplo, tal como monitores LCD, monitores de panel plano, pantallas de plasma y similares) pueden estar acoplados. Los dispositivos periféricos y de entrada/salida (I/O), que se acoplan al controlador 271 de I/O, se pueden conectar al dispositivo 200 informático por cualquier medio conocido en la técnica, incluidos, entre otros, uno o más puertos en serie, uno o más puertos en paralelo y uno o más buses seriales universales (USB). La interfaz o interfaces 281 externas (que pueden incluir una tarjeta de interfaz de red y/o puertos seriales) pueden usarse para conectar el dispositivo 200 informático a una red (por ejemplo, tal como la Internet o una red de área local (LAN)). Las interfaces 281 externas también pueden incluir varias interfaces de entrada que pueden recibir información de varios dispositivos externos, tales como, por ejemplo, una máquina de secuenciación o cualquier componente de la misma. La interconexión a través del bus 275 del sistema permite que uno o más procesadores (por ejemplo, CPU) 273 se comuniquen con cada componente conectado y ejecuten (y/o controlen la ejecución de) instrucciones desde la memoria 272 del sistema y/o desde un dispositivo o dispositivos 279 de almacenamiento, así como el intercambio de información entre varios componentes. La memoria 272 del sistema y/o el dispositivo o dispositivos 279 de almacenamiento se pueden incorporar como uno o más medios de almacenamiento no transitorios legibles por ordenador que almacenan las secuencias de instrucciones ejecutadas por el procesador o procesadores 273, así como otros datos. Dichos medios de almacenamiento no transitorios legibles por ordenador incluyen, entre otros, memoria de acceso aleatorio (RAM), memoria de solo lectura (ROM), un medio electromagnético (por ejemplo, tal como una unidad de disco duro, una unidad de estado sólido, memoria USB, disquete, etc.), un medio óptico tal como un disco compacto (CD) o un disco digital versátil (DVD), memoria flash y similares. Se pueden enviar varios valores de datos y otra información estructurada o no estructurada de un componente o subsistema a otro componente o subsistema, se pueden presentar a un usuario a través del adaptador 282 de pantalla y un dispositivo de pantalla adecuado, se puede enviar a través de una interfaz o interfaces 281 externas a través de una red a un dispositivo remoto o un depósito de datos remoto, o puede almacenarse (temporal y/o permanentemente) en el dispositivo o dispositivos 279 de almacenamiento.

Cualquiera de los métodos y funciones realizados por el dispositivo 200 informático puede implementarse en forma de lógica usando hardware y/o software informáticos de manera modular o integrada. Como se usa en este documento, "lógica" se refiere a un conjunto de instrucciones que, cuando son ejecutadas por uno o más procesadores (por ejemplo, CPU) de uno o más dispositivos informáticos, son operables para realizar una o más funciones y/o devolver datos en la forma de uno o más resultados o datos que utilizan otros elementos lógicos. En varias casos e implementaciones, cualquier lógica dada puede implementarse como uno o más componentes de software que son ejecutables por uno o más procesadores (por ejemplo, CPU), como uno o más componentes de hardware, tales como circuitos integrados para aplicaciones específicas (ASIC) y/o arreglos de compuertas lógicas programables en sitio (FPGA), o como cualquier combinación de uno o más componentes de software y uno o más componentes de

hardware. El o los componentes de software de cualquier lógica particular pueden implementarse, sin limitación, como una aplicación de software independiente, como un cliente en un sistema cliente-servidor, como servidor en un sistema cliente-servidor, como uno o más módulos de software, como una o más bibliotecas de funciones, y como una o más bibliotecas estáticas y/o vinculadas dinámicamente. Durante la ejecución, las instrucciones de cualquier lógica particular pueden incorporarse como uno o más procesos informáticos, subprocesos, fibras y cualquier otra entidad de tiempo de ejecución adecuada que se pueda instanciar en el hardware de uno o más dispositivos informáticos y se les puedan asignar recursos informáticos que pueden incluir, sin limitación, memoria, tiempo de CPU, espacio de almacenamiento y ancho de banda de red.

Técnicas y algoritmos para el proceso de MT

Llamadas de base

En algunas realizaciones, la extracción de datos se basará en dos tipos de datos de imagen: imágenes de campo claro para delimitar las posiciones de todos los DNB en una superficie, y conjuntos de imágenes de fluorescencia adquiridas durante cada ciclo de secuenciación. El software de extracción de datos puede usarse para identificar todos los objetos con las imágenes de campo claro y luego, para cada uno de esos objetos, el software puede usarse para calcular un valor promedio de fluorescencia para cada ciclo de secuenciación. Para cualquier ciclo dado, hay cuatro puntos de datos, correspondientes a las cuatro imágenes tomadas a diferentes longitudes de onda para consultar si esa base es A, G, C o T. Estos puntos de datos sin procesar (también denominados en este documento "llamadas de base") se consolidan, produciendo una lectura de secuenciación discontinua para cada DNB.

Un dispositivo informático puede ensamblar la población de bases identificadas para proporcionar información de secuencia para el ácido nucleico objetivo y/o identificar la presencia de secuencias particulares en el ácido nucleico objetivo. Por ejemplo, el dispositivo informático puede ensamblar la población de bases identificadas de acuerdo con las técnicas y algoritmos descritos en este documento ejecutando diversas lógicas; un ejemplo de dicha lógica es el código de software escrito en cualquier lenguaje de programación adecuado, tal como Java, C++, Perl, Python y cualquier otro lenguaje de programación convencional u orientado a un objetivo adecuado. Cuando se ejecuta en forma de uno o más procesos informáticos, dicha lógica puede leer, escribir y/o o bien procesar datos estructurados y no estructurados que pueden almacenarse en varias estructuras en almacenamiento persistente y/o en una memoria volátil; ejemplos de tales estructuras de almacenamiento incluyen, sin limitación, archivos, tablas, registros de bases de datos, matrices, listas, vectores, variables, registros de memoria y/o procesador, objetos de datos persistentes y/o de memoria instanciados de clases orientadas a objetos, y cualesquier otras estructuras adecuadas de datos. En algunas realizaciones, las bases identificadas se ensamblan en una secuencia completa a través de la alineación de secuencias superpuestas obtenidas de múltiples ciclos de secuenciación realizados en múltiples DNB. Como se usa en el presente documento, el término "secuencia completa" se refiere a la secuencia de genomas parciales o completos, así como de ácidos nucleicos objetivo parciales o completos. En realizaciones adicionales, los métodos de ensamblaje realizados por uno o más dispositivos informáticos o la lógica informática de los mismos utilizan algoritmos que pueden usarse para "juntar" secuencias superpuestas para proporcionar una secuencia completa. En otras realizaciones adicionales, se usan tablas de referencia para ayudar a ensamblar las secuencias identificadas en una secuencia completa. Se puede compilar una tabla de referencia utilizando los datos de secuencia existentes en el organismo de elección. Por ejemplo, se puede acceder a los datos del genoma humano a través del Centro Nacional de Información Biotecnológica en ftp.ncbi.nih.gov/refseq/release, o a través del Instituto J. Craig Venter. Toda o una parte de la información del genoma humano puede usarse para crear una tabla de referencia para consultas de secuenciación particulares. Además, se pueden construir tablas de referencia específicas a partir de datos empíricos derivados de poblaciones específicas, incluida la secuencia genética de humanos con etnias específicas, herencia geográfica, poblaciones religiosas o culturalmente definidas, ya que la variación dentro del genoma humano puede sesgar los datos de referencia dependiendo del origen de la información contenida allí.

En cualquiera de las realizaciones de la invención discutidas en el presente documento, una población de plantillas de ácido nucleico y/o DNB puede comprender una serie de ácidos nucleicos objetivo para cubrir sustancialmente un genoma completo o un polinucleótido objetivo completo. Como se usa en este documento, "cubre sustancialmente" significa que la cantidad de nucleótidos (es decir, secuencias objetivo) analizadas contiene un equivalente de al menos dos copias del polinucleótido objetivo, o en otro aspecto, al menos diez copias, o en otro aspecto, al menos veinte copias, o en otro aspecto, al menos 100 copias. Los polinucleótidos objetivo pueden incluir fragmentos de ADN, incluidos fragmentos de ADN genómico. La orientación para la etapa de la reconstrucción de secuencias de polinucleótidos objetivo se puede encontrar en las siguientes referencias: Lander et al., Genomics, 2: 231-239 (1988); Vingron et al., J. Mol. Biol., 235: 1-12 (1994); y referencias similares.

En algunas realizaciones, se generan cuatro imágenes, una para cada tinte de color, para cada posición consultada de un nucleótido complejo que se secuencia. La posición de cada punto en una imagen y las intensidades resultantes para cada uno de los cuatro colores se determinan ajustando la interferencia entre los tintes y la intensidad de fondo. Un modelo cuantitativo puede ajustarse al conjunto de datos resultante de cuatro dimensiones. Se llama una base para un punto determinado, con un puntaje de calidad que refleja qué tan bien se ajustan las cuatro intensidades al modelo.

La llamada de base de las cuatro imágenes para cada campo se puede realizar en varias etapas mediante uno o más

dispositivos informáticos o la lógica informática de los mismos. Primero, las intensidades de imagen se corrigen para el fondo utilizando la operación de "apertura de imagen" morfológica modificada. Dado que las ubicaciones de los DNB se alinean con las ubicaciones de los píxeles de la cámara, la extracción de intensidad se realiza como una simple lectura de las intensidades de píxeles de las imágenes con corrección de fondo. Estas intensidades se corrigen para varias fuentes tanto de interferencias de señales ópticas y biológicas, como se describe a continuación. Las intensidades corregidas se pasan luego a un modelo probabilístico que finalmente produce para cada DNB un conjunto de cuatro probabilidades de los cuatro posibles resultados de la llamada de base. Luego, se combinan varias métricas para calcular el puntaje de la llamada de base mediante regresión logística previamente ajustada.

Corrección de intensidad

Varias fuentes de interferencias cruzadas biológicas y ópticas se corrigen utilizando un modelo de regresión lineal implementado como lógica informática que se ejecuta por uno o más dispositivos informáticos. Se prefirió la regresión lineal sobre los métodos de deconvolución que son desde el punto de vista informático más costosos y producen resultados con una calidad similar. Las fuentes de las interferencias cruzadas ópticas incluyen superposiciones de bandas de filtro entre los cuatro espectros de tinte fluorescente y las interferencias cruzadas laterales entre los DNB vecinos debido a la difracción de la luz en sus proximidades cercanas. Las fuentes biológicas de las interferencias incluyen lavado incompleto del ciclo anterior, errores de síntesis de la sonda y señales contaminantes de "deslizamiento" de la sonda de posiciones vecinas, extensión incompleta del ancla al interrogar bases "externas" (más distantes) de las anclas. La regresión lineal se usa para determinar la parte de las intensidades de DNB que se puede predecir utilizando intensidades de DNB vecinas o intensidades del ciclo anterior u otras posiciones de DNB. La parte de las intensidades que pueden explicarse por estas fuentes de interferencia se resta de las intensidades extraídas originales. Para determinar los coeficientes de regresión, las intensidades en el lado izquierdo del modelo de regresión lineal deben estar compuestas principalmente de sólo intensidades de "fondo", es decir, intensidades de DNB que no llamarían la base dada para la cual se realiza la regresión. Esto requiere una etapa de llamada previa que se realiza utilizando las intensidades originales. Una vez que se seleccionan los DNB que no tienen una llamada de base particular (con una confianza razonable), un dispositivo informático o la lógica informática realiza una regresión simultánea de las fuentes de interferencia:

$$I_{fondo}^{Base} \approx I_{DNBvecino1}^{Base} + \ldots + I_{DNBvecinoN}^{Base} + I_{DNB}^{Base2} + I_{DNB}^{Base3} + I_{DNB}^{Base4} + I_{DNBcicloanterior}^{Base} + I_{DNBotraposición1}^{Base} + \ldots + I_{DNBotraposiciónN}^{Base} + \varepsilon$$

La interferencia de DNB vecino se corrige utilizando la regresión anterior. Además, cada DNB se corrige para su vecindario particular utilizando un modelo lineal que involucra a todos los vecinos sobre todas las posiciones disponibles de DNB.

Probabilidades de llamadas de base

Las bases de llamada que usan la intensidad máxima no tienen en cuenta las diferentes formas de distribución de intensidad de fondo de las cuatro bases. Para abordar estas posibles diferencias, se desarrolló un modelo probabilístico basado en distribuciones de probabilidad empírica de las intensidades de fondo. Una vez que se corrigen las intensidades, un dispositivo informático o la lógica informática de las mismas llaman previamente algunos DNB usando intensidades máximas (DNB que pasan un cierto umbral de confianza) y utiliza estos DNB previamente llamados para derivar las distribuciones de intensidad de fondo (distribuciones de intensidades de DNB que no llaman una base dada). Al obtener tales distribuciones, el dispositivo informático puede calcular para cada DNB una probabilidad de cola bajo esa distribución que describe la probabilidad empírica de que la intensidad sea la intensidad de fondo. Por lo tanto, para cada DNB y cada una de las cuatro intensidades, el dispositivo informático o la lógica del mismo pueden obtener y almacenar sus probabilidades de ser de fondo (p_{BG}^{A}, p_{BG}^{C}, p_{BG}^{G}, p_{BG}^{T}). Luego, el dispositivo informático puede calcular las probabilidades de todos los posibles resultados de la llamada de base utilizando estas probabilidades. Los posibles resultados de la llamada de base deben describir también los puntos que pueden ser ocupados en forma dobles o en general múltiple o no ocupados por un DNB. La combinación de las probabilidades calculadas con sus probabilidades previas (previamente menor para puntos ocupados en forma múltiple o vacíos) da lugar a las probabilidades de los 16 resultados posibles:

$$p^{A} = \frac{!p_{BG}^{A} + p_{BG}^{C} + p_{BG}^{G} + p_{BG}^{T}}{\sum p} * p_{\substack{previamente \\ Una\ sola\ base}}$$

$$p^{AC} = \frac{!p_{BG}^{A} + !p_{BG}^{C} + p_{BG}^{G} + p_{BG}^{T}}{\sum p} * p_{\substack{previamente \\ Doblemente\ ocupado}}$$

$$p^{ACG} = \frac{!\,p^{A}_{BG} + !\,p^{C}_{BG} + !\,p^{G}_{BG} + p^{T}_{BG}}{\sum p} * p^{\text{previamente}}_{\text{Ocupación triple}}$$

$$p^{ACGT} = \frac{!\,p^{A}_{BG} + !\,p^{C}_{BG} + !\,p^{G}_{BG} + !\,p^{T}_{BG}}{\sum p} * p^{\text{previamente}}_{\text{Ocupación cuádruple}}$$

$$p^{N} = \frac{p^{A}_{BG} + p^{C}_{BG} + p^{G}_{BG} + p^{T}_{BG}}{\sum p} * p^{\text{previamente}}_{\text{Punto vacío}}$$

Estas 16 probabilidades se pueden combinar para obtener un conjunto reducido de cuatro probabilidades para las cuatro posibles llamadas de base. Es decir:

$$p^{A}_{4base} = p^{A} + \tfrac{1}{2}\left(p^{AC} + p^{AG} + p^{AT}\right) + \tfrac{1}{3}\left(p^{ACG} + p^{ACT} + p^{AGT}\right) + \tfrac{1}{4}\,p^{ACGT} + \tfrac{1}{4}\,p^{N}$$

Cálculo de puntaje

Se usó regresión logística para derivar la fórmula de cálculo de el puntaje. Un dispositivo informático o la lógica informática del mismo ajustaban la regresión logística a los resultados del mapeo de la llamada de base utilizando varias métricas como entradas. Las métricas incluyeron la razón de probabilidad entre la base llamada y la siguiente base más alta, llamada intensidad de base, variable indicadora de la identidad de la llamada de base y métricas que describen la calidad del agrupamiento general del campo. Todas las métricas se transformaron para ser colineales con la relación logarítmica de probabilidades entre llamadas concordantes y discordantes. El modelo fue refinado mediante validación cruzada. La función logit con los coeficientes de regresión logística final se utilizó para calcular los puntajes en la producción.

Mapeo y ensamblaje

En realizaciones adicionales, los datos leídos se codifican en un formato binario compacto e incluyen tanto una base llamada como un puntaje de calidad. El puntaje de calidad se correlaciona con la precisión de la base. La lógica del software de análisis, incluido el software de ensamblaje de secuencias, puede usar el puntaje para determinar la contribución de la evidencia de bases individuales con una lectura.

Las lecturas pueden estar "separadas" debido a la estructura DNB. Los tamaños de los huecos varían (generalmente +/- 1 base) debido a la variabilidad inherente a la digestión enzimática. Debido a la naturaleza de acceso aleatorio de cPAL, las lecturas pueden ocasionalmente tener una base no leída ("sin llamada") en un DNB de alta calidad. Los pares de lectura están emparejados.

La lógica de software de mapeo capaz de alinear datos leídos con una secuencia de referencia se puede usar para mapear datos generados por los métodos de secuenciación descritos en este documento. Cuando se ejecuta por uno o más dispositivos informáticos, dicha lógica de mapeo generalmente tolerará pequeñas variaciones de una secuencia de referencia, tal como las causadas por variaciones genómicas individuales, errores de lectura o bases no leídas. Esta propiedad a menudo permite la reconstrucción directa de los SNP. Para admitir el ensamblaje de variaciones más grandes, incluidos cambios estructurales a gran escala o regiones de variación densa, cada brazo de un DNB se puede mapear por separado, con restricciones de emparejamiento por pares aplicadas después de la alineación.

Como se usa en el presente documento, el término "variante de secuencia" o simplemente "variante" incluye cualquier variante, que incluye pero no se limita a una sustitución o reemplazo de una o más bases; una inserción o eliminación de una o más bases (también denominado "indel"); inversión; conversión; duplicación o variación de número de copia (CNV); expansión de repetición de trinucleótidos; variación estructural (SV; por ejemplo, reordenamiento intracromosómico o intercromosómico, por ejemplo, una translocación); etc. En un genoma diploide, una "heterocigosidad" o "het" son dos alelos diferentes de un gen particular en un par de genes. Los dos alelos pueden ser mutantes diferentes o un alelo de tipo silvestre emparejado con un mutante. Los métodos actuales también pueden usarse en el análisis de organismos no diploides, ya sea que dichos organismos sean haploides/monoploides (N = 1, en el que N = número haploide de cromosomas), poliploides o aneuploides.

El ensamblaje de las lecturas de secuencia puede en algunas realizaciones utilizar lógica de software que soporta la estructura de lectura de DNB (lecturas acopladas, separadas con bases no llamadas) para generar un ensamblaje del genoma diploide que en algunas realizaciones puede aprovecharse de la información de secuencia generando métodos de MT de la presente invención para la fase de sitios heterocigóticos.

Los métodos de la presente invención se pueden usar para reconstruir segmentos nuevos que no están presentes en una secuencia de referencia. Algoritmos que utilizan una combinación de razonamiento evidencial (bayesiano) y algoritmos basados en gráficos de Bruijin pueden usarse en algunas realizaciones. En algunas realizaciones, se pueden usar modelos estadísticos calibrados empíricamente para cada conjunto de datos, permitiendo que todos los datos leídos se usen sin prefiltrado o recorte de datos. Las variaciones estructurales a gran escala (incluidas, entre otras, las eliminaciones, las translocaciones y similares) y las variaciones del número de copias también se pueden detectar aprovechando las lecturas acopladas.

Separación por fases de los datos de MT

La Figura 7 describe las etapas principales en la separación por fases de datos de MT. Estas etapas son los siguientes:

(1) Construcción del gráfico utilizando datos de MT: uno o más dispositivos informáticos o la lógica informática de los mismos genera un gráfico no dirigido, en el que los vértices representan los SNP heterocigotos y los bordes representan la conexión entre esos SNP heterocigotos. El borde se compone de la orientación y la fuerza de la conexión. Los uno o más dispositivos informáticos pueden almacenar dicho gráfico en estructuras de almacenamiento que incluyen, sin limitación, archivos, tablas, registros de bases de datos, matrices, listas, vectores, variables, registros de memoria y/o procesador, objetos de datos persistentes y/o de memoria instanciados a partir de clase de objetos orientados y cualquier otra estructura de datos temporal y/o persistente adecuada.

(2) Construcción del gráfico usando datos de emparejamiento por pares: la etapa 2 es similar a la etapa 1, en la que las conexiones se realizan con base en los datos emparejados por pares, en oposición a los datos de MT. Para que se realice una conexión, se puede encontrar un DNB con los dos SNP heterocigotos de interés en la misma lectura (mismo brazo o brazo pareado).

(3) Combinación de gráficos: un dispositivo informático o la lógica informática del mismo representa cada uno de los gráficos anteriores mediante una matriz dispersa de N x N, en la que N es el número de los SNP heterocigóticos candidatos en ese cromosoma. Dos nodos solo pueden tener una conexión en cada uno de los métodos anteriores. Cuando se combinan los dos métodos, puede haber hasta dos conexiones para dos nodos. Por lo tanto, el dispositivo informático o la lógica informática del mismo pueden usar un algoritmo de selección para seleccionar una conexión tal como la conexión de elección. La calidad de los datos de emparejamiento por pares es significativamente inferior a la de los datos de MT. Por lo tanto, solo se utilizan las conexiones derivadas de MT.

(4) Recorte de gráficos: un dispositivo informático ideó y aplicó una serie de heurísticas a los datos de gráficos almacenados para eliminar algunas de las conexiones erróneas. Más precisamente, un nodo puede satisfacer la condición de al menos dos conexiones en una dirección y una conexión en la otra dirección; de lo contrario, se elimina.

(5) Optimización del gráfico: un dispositivo informático o la lógica informática del mismo optimizó el gráfico al generar el árbol de expansión mínima (MST). La función de energía se estableció como - |fuerza|. Durante este proceso, cuando sea posible, se eliminan los bordes de menor fuerza, debido a la competencia con las rutas más fuertes. Por lo tanto, MST proporciona una selección natural para las conexiones más fuertes y confiables.

(6) Construcción del cóntigo: una vez que el árbol de expansión mínima se genera y/o almacena en un medio legible por ordenador, un dispositivo informático o lógica del mismo puede reorientar todos los nodos tomando un nodo (en el presente documento, el primer nodo) constante. Este primer nodo es el nodo de anclaje. Para cada uno de los nodos, el dispositivo informático encuentra la ruta al nodo de anclaje. La orientación del nodo de prueba es el agregado de las orientaciones de los bordes en la ruta.

(7) Separación universal por fases: después de las etapas anteriores, un dispositivo informático o la lógica del mismo separa por fases cada uno de los cóntigos que se construyeron en la etapa o etapas anteriores. En el presente documento, los resultados de esta parte se denominan separación previa por fases, en lugar de separación por fases, lo que indica que esta no es la separación final por fases. Dado que el primer nodo se eligió arbitrariamente como el nodo de anclaje, la separación por fases de todo el cóntigo no está necesariamente en línea con los cromosomas parentales. Para la separación universal por fases, se utilizan algunos SNP heterocigotos en el cóntigo para los cuales se dispone de información del trío. Estos trio heterocigotos SNP se utilizan para identificar la alineación del cóntigo. Al final de la etapa de separación universal por fases, todos los cóntigos han sido marcados correctamente y, por lo tanto, pueden considerarse como un cóntigo de todo el cromosoma.

Elaboración de cóntigos

Con el fin de elaborar cóntigos, para cada par SNP heterocigoto, un dispositivo informático o la lógica informática del mismo prueba dos hipótesis: la orientación hacia adelante y la orientación inversa. Una orientación hacia adelante significa que los dos SNP heterocigotos están conectados de la misma manera en que se enumeraron originalmente (inicialmente alfabéticamente). Una orientación inversa significa que los dos SNP heterocigotos están conectados en orden inverso a su listado original. La Figura 8 representa el análisis por pares de SNP heterocigotos cercanos que implican la asignación de orientaciones directas e inversas a un par SNP heterocigótico.

Cada orientación tendrá un soporte numérico, que muestra la validez de la hipótesis correspondiente. Este soporte es

una función de las 16 células de la matriz de conectividad que se muestra en la Figura 9, que muestra un ejemplo de la selección de una hipótesis y la asignación de un puntaje para ella. Para simplificar la función, las 16 variables se reducen a 3: Energía 1, Energía 2 e Impureza. La Energía 1 y la Energía2 son dos células de mayor valor correspondientes a cada hipótesis. La impureza es la relación de la suma de todas las otras células (que las dos correspondientes a la hipótesis) con respecto a la suma total de las células en la matriz. La selección entre las dos hipótesis se realiza en función de la suma de las células correspondientes. La hipótesis con la suma más alta es la hipótesis ganadora. Los siguientes cálculos solo se utilizan para asignar la fuerza de esa hipótesis. Una hipótesis fuerte es la que tiene un alto valor para Energía 1 y Energía 2, y un valor bajo para Impureza.

Las tres métricas Energía 1, Energía 2 e Impureza se introducen en un sistema de inferencia difusa (Figura 10), para reducir sus efectos en un solo valor -puntaje - entre (e incluyendo) 0 y 1. El sistema de interferencia difusa (FIS) se implementa como una lógica informática que puede ser ejecutada por uno o más dispositivos informáticos.

La operación de conectividad se realiza para cada par SNP heterocigoto que está dentro de una distancia razonable hasta la longitud de cóntigo esperada (por ejemplo, 20-50 Kb). La Figura 6 muestra la construcción del gráfico, que representa algunos ejemplos de conectividades y fuerzas para tres SNP heterocigotos cercanos.

Las reglas del motor de inferencia difusa se definen de la siguiente manera:

(1) Si Energía 1 es pequeña y Energía 2 es pequeña, entonces el Puntaje es muy pequeño.

(2) Si Energía 1 es mediano y Energía 2 es pequeño, entonces el Puntaje es pequeño.

(3) Si Energía 1 es medio y Energía 2 es mediano, entonces el Puntaje es medio.

(4) Si Energía 1 es grande y Energía 2 es pequeño, entonces el Puntaje es medio.

(5) Si Energía 1 es grande y Energía 2 es mediano, entonces el Puntaje es grande.

(6) Si Energía 1 es grande y Energía 2 es grande, entonces el Puntaje es muy grande.

(7) Si la impureza es pequeña, entonces el Puntaje es grande.

(8) Si la impureza es media, entonces el Puntaje es pequeño.

(9) Si la impureza es grande, entonces el Puntaje es muy pequeño.

Para cada variable, la definición de Pequeño, Mediano y Grande es diferente, y se rige por sus funciones específicas de membresía.

Después de exponer el sistema de inferencia difusa (FIS) a cada conjunto de variables, la contribución del conjunto de entrada en las reglas se propaga a través del sistema de lógica difusa, y se genera un único número (no difuso) en la salida - puntaje. Este puntaje está limitado entre 0 y 1, en el que 1 muestra la calidad más alta

Después de la aplicación del FIS a cada par de nodos, un dispositivo informático o la lógica informática del mismo construye un gráfico completo. La Figura 11 muestra un ejemplo de dicho gráfico. Los nodos se colorean de acuerdo con la orientación de la hipótesis ganadora. La fuerza de cada conexión se deriva de la aplicación del FIS en el par de interés SNP heterocigoto. Una vez que se construye el gráfico preliminar (el gráfico superior de La Figura 11), el dispositivo informático o la lógica del mismo optimiza el gráfico (el gráfico inferior de La Figura 11) y lo reduce a un árbol. Este proceso de optimización se realiza creando un árbol de expansión mínima (MST) a partir del gráfico original. El MST garantiza una ruta única desde cada nodo a cualquier otro nodo.

La Figura 11 muestra la optimización del gráfico. En esta aplicación, el primer nodo en cada cóntigo se usa como nodo de anclaje, y todos los demás nodos están orientados a ese nodo. Dependiendo de la orientación, cada acierto tendría que voltearse o no, para que coincida con la orientación del nodo de anclaje. La Figura 12 muestra el proceso de alineación del cóntigo para el ejemplo dado. Al final de este proceso, se pone a disposición un cóntigo separado por fases.

En este punto del proceso de separación por fases, se separan los dos haplotipos. Aunque se sabe que uno de estos haplotipos proviene de la mamá y el otro del papá, no se sabe exactamente cuál proviene de qué padre. En la siguiente etapa de separación por fases, un dispositivo informático o la lógica informática del mismo intenta asignar la etiqueta parental correcta (mamá/papá) a cada haplotipo. Este proceso se conoce como la separación universal por fases. Para hacerlo, es necesario conocer la asociación de al menos algunos de los SNP heterocigotos (en el cóntigo) a los padres. Esta información se puede obtener haciendo una separación por fases tripartita (mamá-papá-hijo). Usando los genomas secuenciados del trío, se identifican algunos loci con asociaciones parentales conocidas, más específicamente cuando al menos uno de los padres es homocigoto. Luego, el dispositivo informático o la lógica informática del mismo utilizan estas asociaciones para asignar la etiqueta parental correcta (mamá/papá) a todo el cóntigo, es decir, para realizar una separación universal por fases asistida por los padres (Figura 13).

Para garantizar una alta precisión, se puede realizar lo siguiente: (1) cuando sea posible (por ejemplo, en el caso de NA19240), obtener la información del trío de múltiples fuentes y usar una combinación de tales fuentes; (2) requerir que los cóntigos incluyan al menos dos loci conocidos separando el trio por fases; (3) eliminar los cóntigos que tienen una serie de desajustes del trio en una fila (lo que indica un error segmentario); y (4) eliminar los cóntigos que tienen una sola falta de coincidencia del trío al final de los loci del trio (lo que indica un posible error segmentario).

La Figura 14 muestra separaciones de cóntigos naturales. Ya sea que se usen o no los datos de los padres, los cóntigos a menudo no continúan naturalmente más allá de cierto punto. Las razones para la separación de cóntigos son: (1) fragmentación de ADN más de lo habitual o falta de amplificación en ciertas áreas, (2) baja densidad de SNP heterocigoto, (3) secuencia de poli-N en el genoma de referencia y (4) regiones de repetición de ADN (propensas a falta de mapeo).

La Figura 15 muestra la separación universal por fases. Una de las principales ventajas de la separación universal por fases es la capacidad de obtener los "cóntigos" cromosómicos completos. Esto es posible porque cada cóntigo (después de la separación universal por fases) porta haplotipos con las etiquetas parentales correctas. Por lo tanto, todos los cóntigos que portan la etiqueta mamá se pueden poner en el mismo haplotipo; y se puede hacer una operación similar para los cóntigos de papá.

Otra de las principales ventajas del proceso de MT es la capacidad de aumentar dramáticamente la precisión de las llamadas de SNP heterocigotos. La Figura 16 muestra dos ejemplos de detección de errores resultantes del uso del proceso de MT. El primer ejemplo se muestra en La Figura 16 (izquierda), en la que la matriz de conectividad no admite ninguna de las hipótesis esperadas. Esto es una indicación de que uno de los SNP heterocigotos no es realmente un SNP heterocigótico. En este ejemplo, el SNP heterocigótico A/C es en realidad un locus homocigoto (A/A), que se marcó erróneamente como un locus heterocigoto por el ensamblador. Este error se puede identificar y eliminar o (en este caso) corregir. El segundo ejemplo se muestra en La Figura 17 (derecha), en la cual la matriz de conectividad para este caso admite ambas hipótesis al mismo tiempo. Esta es una señal de que las llamadas SNP heterocigoto no son reales.

Una matriz "sana" de conexión de SNP heterocigótico es aquella que tiene solo dos células altas (en las posiciones esperadas de SNP heterocigoto, es decir, no en una línea recta). Todas las demás posibilidades apuntan a problemas potenciales y pueden eliminarse o utilizarse para realizar llamadas de base alternativas para los lugares de interés.

Otra ventaja del proceso de MT es la capacidad de llamar SNP heterocigotos con soportes débiles (por ejemplo, donde era difícil mapear los DNB debido a la tasa de sesgo o falta de coincidencia). Dado que el proceso de MT requiere una restricción adicional en los SNP heterocigotos, se podría reducir el umbral que requiere una llamada de SNP heterocigótico en un ensamblador que no sea MT. La Figura 17 muestra un ejemplo de este caso en el que se puede realizar una llamada segura de SNP heterocigótico a pesar de un pequeño número de lecturas. En La Figura 17 (derecha), en un escenario normal, el bajo número de lecturas de soporte habría impedido que cualquier ensamblador llamara con confianza a los SNP heterocigotos correspondientes. Sin embargo, dado que la matriz de conectividad es "limpia", se podría asignar con mayor confianza llamadas de SNP heterocigóticos a estos loci.

Anotación de SNP en sitios de empalme

Los intrones en los ARN transcritos necesitan ser empalmados antes de que se conviertan en ARNm. La información para empalmar está incrustada dentro de la secuencia de estos ARN, y se basa en el consenso. Las mutaciones en la secuencia de consenso del sitio de empalme son causas de muchas enfermedades humanas (Faustino y Cooper, Genes Dev. 17: 419-437, 2011). La mayoría de los sitios de empalme se ajustan a un consenso simple en posiciones fijas alrededor de un exón. En este sentido, se desarrolló un programa para anotar mutaciones en el sirio de empalme. En este programa, se utilizaron modelos de posición de empalme de consenso (accesible desde el sitio web de la Escuela de Informática, Matemáticas y Ciencias Naturales, de acuerdo con Steve Mount). Se realiza una búsqueda de un patrón: CAG|G en la región del extremo 5’ de un exón ("|" indica el comienzo del exón) y MAG|GTRAG en la región del extremo 3' del mismo exón ("|" indica el final del exón). Aquí M = {A, C}, R = {A, G}. Además, las posiciones de consenso de empalme se clasifican en dos tipos: tipo I, en el que se requiere el consenso del modelo al 100%; y tipo II, en el que se preserva el consenso con el modelo en más del 50% de los casos. Presumiblemente, una mutación de SNP en una posición de tipo I hará que el empalme se pierda, mientras que un SNP en una posición de tipo II solo disminuirá la eficiencia del evento de empalme.

La lógica del programa para anotar mutaciones en el sitio de empalme comprende dos partes. En la parte I, se genera un archivo que contiene secuencias de posiciones del modelo del genoma de referencia de entrada. En la parte 2, los SNP de un proyecto de secuenciación se comparan con estas secuencias de posiciones del modelo e informan cualquier tipo de mutaciones de tipo I y tipo II. La lógica del programa está centrada en el exón en lugar de centrada en el intrón (por conveniencia para analizar el genoma). Para un exón dado, en su extremo 5’ se busca el consenso "cAGg" (para las posiciones -3, -2, -1, 0. 0 significa el inicio del exón). Las letras mayúsculas significan posiciones de tipo I, y las letras minúsculas significan posiciones de tipo II). En el extremo 3’ del exón, se realiza una búsqueda del consenso "magGTrag" (para la secuencia de posición -3, -2, -1, 0, 1, 2, 3, 4). Los exones de la liberación del genoma que no confirman estos requisitos simplemente se ignoran (-5% de todos los casos). Estos exones caen en otras clases menores de consenso de sitio de empalme y no son investigados por la lógica del programa. Cualquier SNP

del genoma secuenciado se compara con la secuencia modelo en estas posiciones genómicas. Cualquier falta de coincidencia en el tipo I será reportado. La falta de coincidencia en las posiciones de tipo II se informan si la mutación se aleja del consenso.

La lógica del programa anterior detecta la mayoría de las mutaciones malas del sitio de empalme. Los SNP defectuosos que se informan son definitivamente problemáticos. Pero hay muchos otros SNP defectuosos que causan problemas de empalme que este programa no detecta. Por ejemplo, hay muchos intrones dentro del genoma humano que no confirman el consenso mencionado anteriormente. Además, las mutaciones en los puntos de bifurcación en el medio del intrón también pueden causar problemas de empalme. Estas mutaciones en el sitio de empalme no se informan.

Anotación de los SNP que afectan a los sitios de unión del factor de transcripción (TFBS).

Los modelos JASPAR se usan para encontrar TFBS a partir de las secuencias liberadas del genoma humano (ya sea la construcción 36 o la construcción 37). JASPAR Core es una colección de 130 datos de frecuencia posicional de TFBS para vertebrados, modelados como matrices (Bryne et al., Nucl. Acids Res. 36: D102-D106, 2008; Sandelin et al., Nucl. Acids Res. 23: D91-D94, 2004). Estos modelos se descargan del sitio web de JASPAR (http://jaspar.genereg.nef/cgi-bin/jaspar_db.pl?Rm=browse&db=core&tax_group=vertebrates). Estos modelos se convierten en matrices ponderadas de posición (PWM) utilizando la siguiente fórmula: wi = log2 [(fi + p Ni1/2)/(Ni + Ni1/2)/p], en la que: fi es la frecuencia observada para la base específica en la posición I; Ni es el total de observaciones en la posición; y p la frecuencia de fondo para el nucleótido actual, que está predeterminado como 0,25 (Wasserman y Sandelin, Nature Reviews, Genetics 5: P276-287, 2004). Se utiliza un programa específico, mast (meme.sdsc.edu/meme/mast-intro.html), para buscar segmentos de secuencia dentro del genoma para sitios TFBS. Se ejecutó un programa para extraer sitios TFBS en el genoma de referencia. El esquema de las etapas es el siguiente: (i) Para cada gen con ARNm, se extraen [-5000, 1000] regiones que contienen TFBS putativo del genoma, siendo 0 la ubicación inicial del ARNm. (ii) Ejecutar la búsqueda por mast de todos los modelos de PWM para las secuencias que contienen TFBS putativo. (iii) Seleccionar esos aciertos por encima de un umbral dado. (iv) Para regiones con aciertos múltiples o de superposición, seleccionar solo 1 acierto, el que tenga el puntaje más alto de búsqueda por mast.

Con los aciertos del modelo de TFBS del genoma de referencia generado y/o almacenado en un medio adecuado legible por ordenador, un dispositivo informático o la lógica informática del mismo pueden identificar los SNP que se encuentran dentro de la región del acierto. Estos SNP tendrán un impacto en el modelo y un cambio en el puntaje de acierto. Se escribió un segundo programa para calcular dichos cambios en el puntaje de aciertos, ya que el segmento que contiene el SNP se ejecuta dos veces en el modelo de PWM, una vez para la referencia, y la segunda vez para el que tiene la sustitución de SNP. Un SNP que causa que el puntaje de acierto del segmento caiga más de 3 se identifica como un SNP defectuoso.

Selección de genes con dos SNP defectuosos. Los genes con SNP defectuosos se clasifican en dos categorías: (1) los que afectan la secuencia de AA transcrita; y (2) aquellos que afectan el sitio de unión de la transcripción. Para la secuencia de AA que afecta, las siguientes subcategorías de SNP se incluyen:

(1) Variaciones sin sentido o sin parada. Estas mutaciones causan una proteína truncada o una proteína extendida. En cualquier situación, la función del producto proteico se pierde por completo o es menos eficiente.

(2) Variaciones del sitio de empalme. Estas mutaciones provocan que el sitio de empalme para un intrón se destruya (para que las posiciones requieran el 100% de un determinado nucleótido de acuerdo con el modelo) o disminuya severamente (para que el sitio requiera más del 50% para un determinado nucleótido por el modelo). El SNP hace que el nucleótido del sitio de empalme mute a otro nucleótido (que esta por debajo del 50% del consenso de acuerdo con lo predicho por el modelo de secuencia de consenso del sitio de empalme). Es probable que estas mutaciones produzcan proteínas truncadas, que falten exones o que disminuyan severamente la cantidad de proteína.

(3) Anotación Polyphen2 de variaciones de AA. Para los SNP que causan cambios en la secuencia de aminoácidos de una proteína, pero no su longitud, se usó Polyphen2 (Adzhubei et al., Nat. Methods 7: 248-249, 2010) como la herramienta principal de anotación. Polyphen2 anota el SNP con "benigno", "desconocido", "posiblemente dañino" y "probablemente dañino". Tanto "posiblemente dañino" como "probablemente dañino" fueron identificados como SNP defectuosos. Estas asignaciones de categoría de Polyphen2 se basan en predicciones estructurales del software Polyphen2.

Para las mutaciones en el sitio de unión a la transcripción, se usó el 75% de maxScore de los modelos con base en el genoma de referencia como un cribado para los sitios de unión a TFBS. Se elimina cualquier acierto del modelo en la región que sea <= 75% de maxScore. Para los restantes, si un SNP hace que el puntaje de acierto caiga 3 o más, se considera como un SNP perjudicial.

Se informan dos clases de genes. Los genes de Clase 1 son aquellos que tenían al menos 2 mutaciones que afectan a AA defectuosos. Estas mutaciones pueden estar todas en un solo alelo (Clase 1.1), o extenderse en 2 alelos distintos (Clase 1.2). Los genes de Clase 2 son un superconjunto del conjunto de Clase 1. Los genes de Clase 2 son genes que contienen al menos 2 SNP defectuosos, independientemente de que afecte a AA o al sitio TFBS. Pero un requisito

es que al menos 1 SNP afecte a AA. Los genes de Clase 2 son aquellos en la Clase 1 o aquellos que tienen 1 mutación AA perjudicial y 1 o más variaciones perjudiciales que afectan a TFBS. La Clase 2.1 significa que todas estas mutaciones perjudiciales son de un solo alelo, mientras que la Clase 2.2 significa que los SNP perjudiciales provienen de dos alelos distintos.

Las técnicas y algoritmos anteriores son aplicables a los métodos para secuenciar ácidos nucleicos complejos, opcionalmente junto con el procesamiento de MT antes de la secuenciación (MT en combinación con la secuenciación puede denominarse "secuenciación de MT"), que se describen en detalle como sigue. Dichos métodos para secuenciar ácidos nucleicos complejos pueden ser realizados por uno o más dispositivos informáticos que ejecutan la lógica informática. Un ejemplo de dicha lógica es el código de software escrito en cualquier lenguaje de programación adecuado, tal como Java, C++, Perl, Python y cualquier otro lenguaje de programación convencional adecuado y/u orientado al objeto. Cuando se ejecuta en forma de uno o más procesos informáticos, dicha lógica puede leer, escribir y/o bien procesar datos estructurados y no estructurados que pueden almacenarse en varias estructuras en almacenamiento persistente y/o en memoria volátil; ejemplos de tales estructuras de almacenamiento incluyen, sin limitación, archivos, tablas, registros de bases de datos, matrices, listas, vectores, variables, registros de memoria y/o procesador, objetos de datos persistentes y/o de memoria instanciados de clases orientadas a objetos, y cualquier otra estructura adecuada de datos.

Mejora de la precisión en la secuenciación de lectura larga

En la secuenciación de ADN usando ciertas tecnologías de lectura larga (por ejemplo, secuenciación de nanoporos), están disponibles longitudes de lectura largas (por ejemplo, 10-100 kb) pero generalmente tienen altas tasas de falsos negativos y falsos positivos. La precisión final de la secuencia de tales tecnologías de lectura larga puede mejorarse significativamente utilizando información de haplotipo (separación por fases completa o parcial) de acuerdo con el siguiente proceso general.

Primero, un dispositivo informático o de lógica informática del mismo alinean las lecturas entre sí. Se espera que exista una gran cantidad de llamadas heterocigotas en la superposición. Por ejemplo, si dos a cinco fragmentos de 100 kb se superponen mínimo en un 10%, esto da como resultado una superposición de más de 10 kb, lo que podría traducirse aproximadamente en 10 loci heterocigotos. Alternativamente, cada lectura larga se alinea con un genoma de referencia, mediante el cual se obtendría implícitamente una alineación múltiple de las lecturas.

Una vez que se han logrado las alineaciones de lectura múltiple, se puede considerar la región de superposición. El hecho de que la superposición podría incluir un gran número (por ejemplo, N = 10) de los loci de het puede aprovecharse para considerar combinaciones de hets. Esta modalidad combinatoria da como resultado un gran espacio (4^N o 4^N; si N = 10, entonces 4^N = ~1 millón) de posibilidades para los haplotipos. De todos estos 4^N puntos en el espacio N dimensional, solo se espera que dos puntos contengan información biológicamente viable, es decir, los correspondientes a los dos haplotipos. En otras palabras, hay una relación de supresión de ruido de $4^N/2$ (aquí $10^6/2$ o ~500.000). En realidad, gran parte de este espacio 4^N está degenerado, particularmente porque las secuencias ya están alineadas (y, por lo tanto, se parecen), y también porque cada locus generalmente no porta más de dos bases posibles (si esta es una het real). En consecuencia, un límite inferior para este espacio es en realidad 2^N (si N = 10, entonces 2N = ~1000). Por lo tanto, la relación de supresión de ruido solo podría ser $2^N/2$ (aquí 1000/2 = 500), lo que sigue siendo bastante impresionante. A medida que aumenta el número de falsos positivos y falsos negativos, el tamaño del espacio se expande de 2^N a 4^N, lo que a su vez resulta en una mayor relación de supresión de ruido. En otras palabras, a medida que crece el ruido, se suprimirá automáticamente. Por lo tanto, se espera que los productos de salida retengan solo una cantidad muy pequeña (y bastante constante) de ruido, casi independientemente del ruido de entrada. (La compensación es la pérdida de rendimiento en las condiciones más ruidosas). Por supuesto, estas relaciones de supresión se alteran si (1) los errores son sistemáticos (u otras idiosincrasias de datos), (2) los algoritmos no son óptimos, (3) las secciones de superposición son más cortas o (4) la redundancia de cobertura es menor. N es cualquier número entero mayor que uno, tal como 2, 3, 5, 10 o más.

La siguiente metodología es útil para aumentar la precisión de los métodos de secuenciación de lectura larga, que podrían tener una gran tasa de error inicial.

Primero, un dispositivo informático o la lógica informática del mismo alinea algunas lecturas, por ejemplo 5 lecturas. Suponiendo que las lecturas son ~100 kb, y la superposición compartida es del 10%, esto da como resultado una superposición de 10 kb en las 5 lecturas o más, tal como 10-20 lecturas. También se supone que hay un het en cada 1 Kb. Por lo tanto, habría un total de 10 hets en esta región común.

A continuación, el dispositivo informático o la lógica informática del mismo se completa en una parte (por ejemplo, solo elementos distintos de cero) o toda la matriz de posibilidades $alfa^{10}$ (en la que alfa está entre 2 y 4) para los 10 hets candidatos anteriores. En una implementación, solo 2 de las células $alfa^{10}$ de esta matriz tienen alta densidad (por ejemplo, medida por un umbral, que puede ser predeterminado o dinámico). Estas son las células que corresponden a los hets reales. Estas dos células pueden considerarse centros sustancialmente libres de ruido. El resto contendrá principalmente membresías 0 y ocasionalmente 1, especialmente si los errores no son sistemáticos. Si los errores son sistemáticos, puede haber un evento de agrupación (por ejemplo, una tercera célula que tiene más que solo 0 o 1), lo que dificulta la tarea. Sin embargo, incluso en este caso, la membresía del grupo para el grupo falso debería ser

significativamente más débil (por ejemplo, medida por una cantidad absoluta o relativa) que aquella de los dos grupos esperados. La compensación en este caso es que el punto de partida incluiría más secuencias múltiples alineadas, lo que se relaciona directamente con tener lecturas más largas o una mayor redundancia de cobertura.

La etapa anterior supone que los dos grupos viables se observan entre las lecturas superpuestas. Para una gran cantidad de falsos positivos, este no sería el caso. Si este es el caso, en el espacio alfa-dimensional, los dos grupos esperados serán borrosos, es decir, en lugar de ser puntos únicos con alta densidad, serán grupos borrosos de puntos M alrededor de las células de interés, en las que estas células de interés son los centros libres de ruido que se encuentran en el centro del grupo. Esto permite que los métodos de agrupación capturen la localización de los puntos esperados, a pesar del hecho de que la secuencia exacta no está representada en cada lectura. También puede ocurrir un evento de agrupación cuando las agrupaciones están borrosas (es decir, podría haber más de dos centros), pero de manera similar a la descrita anteriormente, se puede usar un puntaje (por ejemplo, los recuentos totales para las células de un grupo) para distinguir un grupo más débil de los dos grupos reales, para un organismo diploide. Los dos grupos reales pueden usarse para crear cóntigos, como se describe en el presente documento, para diversas regiones, y los cóntigos pueden combinarse en dos grupos para formar haplotipos para una gran región del ácido nucleico complejo.

Finalmente, en el dispositivo informático o la lógica informática del mismo, los haplotipos basados en la población (conocidos) pueden usarse para aumentar la confianza y/o proporcionar una guía adicional para encontrar los grupos reales. Una forma de habilitar este método es proporcionar a cada haplotipo observado un peso y proporcionar un valor menor pero no nulo a los haplotipos no observados. Al hacerlo, se logra un sesgo hacia los haplotipos naturales que se han observado en la población de interés.

Convertir lecturas largas a MT virtual

Los algoritmos que están diseñados para MT (incluido el algoritmo de separación por fases) se pueden usar para lecturas largas asignando una etiqueta virtual aleatoria (con distribución uniforme) a cada uno de los fragmentos largos. La etiqueta virtual tiene la ventaja de permitir una verdadera distribución uniforme para cada código. MT no puede alcanzar este nivel de uniformidad debido a la diferencia en la agrupación de los códigos y la diferencia en la eficiencia de decodificación de los códigos. Se puede observar fácilmente una relación de 3:1 (y hasta de 10:1) en la representación de cualquiera de los dos códigos en MT. Sin embargo, el proceso de MT virtual da como resultado una verdadera relación 1:1 entre dos códigos cualesquiera.

En vista de la descripción anterior, de acuerdo con un aspecto, se proporcionan métodos para determinar una secuencia de un ácido nucleico complejo (por ejemplo, un genoma completo) de uno o más organismos, es decir, un organismo individual o una población de organismos. Dichos métodos comprenden: (a) recibir en uno o más dispositivos informáticos una pluralidad de lecturas del ácido nucleico complejo; y (b) producir, con los dispositivos informáticos, una secuencia ensamblada del ácido nucleico complejo a partir de las lecturas, comprendiendo la secuencia ensamblada menos de 1,0; 0,8; 0,7; 0,6; 0,5; 0,4; 0,3; 0,2; 0,1; 0,08; 0,07; 0,06; 0,05 o 0,04 variantes falsas de un solo nucleótido por megabase a una tasa de llamada de 70, 75, 80, 85, 90 o 95 por ciento o más, en la que los métodos son realizados por uno o más dispositivos informáticos. En algunos aspectos, un medio de almacenamiento no transitorio legible por ordenador almacena una o más secuencias de instrucciones que comprenden instrucciones que, cuando son ejecutadas por uno o más dispositivos informáticos, hacen que uno o más dispositivos informáticos realicen las etapas de dichos métodos.

De acuerdo con una realización, en la que dichos métodos implican la separación por fases del haplotipo, el método comprende además identificar una pluralidad de variantes de secuencia en la secuencia ensamblada y la separación por fases de las variantes de secuencia (por ejemplo, 70, 75, 80, 85, 90, 95 por ciento o más de las variantes de secuencia) para producir una secuencia separada por fases, es decir, una secuencia en la que las variantes de secuencia son separadas por fases. Dicha información de separación por fases puede usarse en el contexto de la corrección de errores. Por ejemplo, de acuerdo con una realización, dichos métodos comprenden identificar como error una variante de secuencia que es inconsistente con la separación de fases de al menos dos (o tres o más) variantes de secuencia separadas por fases.

De acuerdo con otra realización de este tipo, en tales métodos, la etapa de recibir la pluralidad de lecturas del ácido nucleico complejo comprende un dispositivo informático y/o una lógica informática del mismo que recibe una pluralidad de lecturas de cada uno de una pluralidad de fragmentos largos del ácido nucleico complejo. La información con respecto a dichos fragmentos es útil para corregir errores o para llamar a una base que de otro modo hubiera sido una "no llamada". De acuerdo con una de tales realizaciones, dichos métodos comprenden un dispositivo informático y/o una lógica informática del mismo que llama a una base en una posición de dicha secuencia ensamblada sobre la base de llamadas de base preliminares para la posición de dos o más fragmentos largos. Por ejemplo, los métodos pueden comprender llamar a una base en una posición de dicha secuencia ensamblada sobre la base de llamadas de base preliminares de al menos dos, al menos tres, al menos cuatro o más de cuatro fragmentos largos. En algunas realizaciones, dichos métodos pueden comprender identificar una llamada de base como verdadera si están presente al menos dos, al menos tres, al menos cuatro fragmentos largos o más de cuatro fragmentos largos. En algunas realizaciones, tales métodos pueden comprender identificar una llamada de base como verdadera si está presente al menos una mayoría (o al menos 60%, al menos 75% o al menos 80%) de los fragmentos para los que se hace una

llamada de base preliminar para esa posición en la secuencia ensamblada. De acuerdo con otra realización de este tipo, dichos métodos comprenden un dispositivo informático y/o una lógica informática del mismo que identifica una llamada de base como verdadera si está presente tres o más veces en lecturas de dos o más fragmentos largos.

De acuerdo con otra realización de este tipo, el fragmento largo del que se originan las lecturas se determina identificando una etiqueta (o patrón único de etiquetas) que está asociada con el fragmento. Dichas etiquetas comprenden opcionalmente un código de corrección o detección de errores (por ejemplo, un código de corrección de errores de Reed-Solomon). De acuerdo con una realización de la invención, al secuenciar un fragmento y una etiqueta, la lectura resultante comprende datos de secuencia de etiqueta y datos de secuencia de fragmento.

De acuerdo con otra realización, dichos métodos comprenden además: un dispositivo informático y/o una lógica informática del mismo que proporciona una primera secuencia separada por fases de una región del ácido nucleico complejo en la región que comprende una repetición en tándem corta; un dispositivo informático y/o una lógica informática del mismo que compara lecturas (por ejemplo, lecturas regulares o de emparejamiento por pares) de la primera secuencia separada por fases de la región con lecturas de una segunda secuencia separada por fases de la región (por ejemplo, utilizando cobertura de secuencia); y un dispositivo informático y/o una lógica informática del mismo que identifica una expansión de la repetición en tándem corta en una de la primera secuencia separada por fases o la segunda secuencia separada por fases con base en la comparación.

De acuerdo con otra realización, el método comprende además un dispositivo informático y/o una lógica informática del mismo que obtiene datos de genotipo de al menos uno de los padres del organismo y produce una secuencia ensamblada del ácido nucleico complejo a partir de las lecturas y los datos del genotipo .

De acuerdo con otra realización, el método comprende además un dispositivo informático y/o una lógica informática del mismo que realiza etapas que comprenden: alinear una pluralidad de las lecturas para una primera región del ácido nucleico complejo, creando así una superposición entre las lecturas alineadas; identificar N hets candidatos dentro de la superposición; agrupar el espacio de 2^N a 4^N posibilidades o un subespacio seleccionado del mismo, creando así una pluralidad de grupos; identificar dos grupos con la densidad más alta, cada grupo identificado comprende un centro sustancialmente libre de ruido; y repetir las etapas anteriores para una o más regiones adicionales del ácido nucleico complejo.

De acuerdo con otra realización, dichos métodos comprenden además proporcionar una cantidad del ácido nucleico complejo y secuenciar el ácido nucleico complejo para producir las lecturas.

De acuerdo con otra realización, en tales métodos, el ácido nucleico complejo se selecciona del grupo que consiste en un genoma, un exoma, un transcriptoma, un metiloma, una mezcla de genomas de diferentes organismos y una mezcla de genomas de diferentes tipos células de un organismo.

De acuerdo con otro aspecto, se proporciona una secuencia del genoma humano ensamblada que se produce por cualquiera de los métodos anteriores. Por ejemplo, uno o más medios de almacenamiento no transitorios legibles por ordenador almacenan una secuencia del genoma humano ensamblado que se produce por cualquiera de los métodos anteriores. De acuerdo con otro aspecto, un medio de almacenamiento no transitorio legible por ordenador almacena una o más secuencias de instrucciones que comprenden instrucciones que, cuando son ejecutadas por uno o más dispositivos informáticos, hacen que uno o más dispositivos informáticos realicen alguno, algunos o todos de los métodos anteriores.

De acuerdo con otro aspecto, se proporcionan métodos para determinar una secuencia completa del genoma humano, comprendiendo tales métodos: (a) recibir, en uno o más dispositivos informáticos, una pluralidad de lecturas del genoma; y (b) producir, con uno o más dispositivos informáticos, una secuencia ensamblada del genoma a partir de las lecturas que comprenden menos de 600 variantes de nucleótidos individuales heterocigóticos falsos por gigabase a una tasa de llamada del genoma del 70% o más. De acuerdo con una realización, la secuencia ensamblada del genoma tiene una tasa de llamada del genoma del 70% o más y una tasa de llamada del exoma del 70% o más. En algunos aspectos, un medio de almacenamiento no transitorio legible por ordenador almacena una o más secuencias de instrucciones que comprenden instrucciones que, cuando son ejecutadas por uno o más dispositivos informáticos, hacen que uno o más dispositivos informáticos realicen cualquiera de los métodos descritos en el presente documento.

De acuerdo con otro aspecto, se proporcionan métodos para determinar una secuencia completa del genoma humano, comprendiendo tales métodos: (a) recibir, en uno o más dispositivos informáticos, una pluralidad de lecturas de cada uno de una pluralidad de fragmentos largos, comprendiendo cada fragmento largo uno o más fragmentos del genoma; y (b) producir, con uno o más dispositivos informáticos, una secuencia ensamblada separada por fases del genoma a partir de las lecturas que comprenden menos de 1000 variantes de nucleótidos individuales falsos por gigabase a una tasa de llamada del genoma del 70% o más. En algunos aspectos, un medio de almacenamiento no transitorio legible por ordenador almacena una o más secuencias de instrucciones que comprenden instrucciones que, cuando son ejecutadas por uno o más dispositivos informáticos, hacen que uno o más dispositivos informáticos realicen dichos métodos.

Kits

Los kits útiles para la práctica de MT pueden incluir uno, dos, tres o más de los siguientes componentes:

A. Bibliotecas

i) Una biblioteca de códigos de barras flanqueada por extremos de transposones (es decir, una biblioteca de transposones etiquetados). En algunos casos, los extremos de transposones son repeticiones terminales invertidas. En algunos casos, los extremos de transposones tienen una longitud de 9-40 bases. En algunos casos, los códigos de barras tienen una longitud de 6-20 bases. En algunos casos, los transposones etiquetados también comprenden sitios de unión del cebador de amplificación (por ejemplo, en los que la mayor parte o la biblioteca completa tiene los mismos sitios de unión del cebador). En algunos casos, los transposones etiquetados comprenden al menos dos sitios de unión del cebador de amplificación. En algunos casos, los dos sitios de unión del cebador de amplificación hibridan con la misma secuencia del cebador. En algunos casos, el kit comprende un cebador o cebadores de amplificación que se hibridan con secuencias de unión a cebadores de los transposones etiquetados.

ii) Una biblioteca de códigos de barras clonales que comprende una pluralidad de 10^4 o más fuentes distintas de códigos de barras clonales. En algunos casos, los códigos de barras clonales son transposones etiquetados como se describe en (i). En algunos casos, los códigos de barras clonales se inmovilizan en un portador o soporte, tal como un polímero, perla, dendrímero o partícula magnética. En algunos casos, las fuentes de los códigos de barras clonales se crean mediante PCR en emulsión. En algunos casos, las fuentes de los códigos de barras clonales se crean utilizando una síntesis combinatoria de mezclar y dividir. En algunos casos, los códigos de barras clonales se unen al soporte con un enlazador (por ejemplo, en el que la mayor parte o la biblioteca completa tiene el mismo enlazador). En algunos casos, el enlazador es escindible de modo que la secuencia del código de barras puede liberarse del soporte mediante tratamiento con un agente de escisión. En algunos casos, el agente de escisión es una endonucleasa de restricción o nickasa.

iii) Una biblioteca de concatámeros que comprende monómeros, en la que los monómeros comprenden códigos de barras. En algunos casos, los monómeros comprenden sitios de unión de cebadores y/o secuencias finales de transposón y/o sitios de reconocimiento de endonucleasas de restricción (por ejemplo, en los que la mayor parte o la biblioteca completa comparte los mismos sitios o secuencias). En algunos casos, los monómeros comprenden transposones etiquetados como se describe en (i).

iv) Una biblioteca de plantillas adecuadas para la amplificación por círculo rodante, en la que las plantillas comprenden un monómero como se describe en (iii). En algunos casos, el kit contiene una enzima (por ejemplo, polimerasa phi29) adecuada para convertir las plantillas en concatámeros.

v) Una biblioteca de oligonucleótidos en horquilla o en bucle de tallo, en la que la biblioteca comprende una pluralidad de al menos aproximadamente 10^4 códigos de barras, comprendiendo cada oligonucleótido dos copias de una secuencia de código de barras (que puede estar en la porción de bucle del oligonucleótido). En algunos casos, cada oligonucleótido comprende dos sitios de unión del cebador de amplificación colocados entre las copias de la secuencia del código de barras. En algunos casos, los oligonucleótidos comprenden secuencias aleatorias o semi aleatorias en los terminales 5' y 3'. En algunos casos, las secuencias tienen 3-8 bases de longitud o 3-5 bases de longitud.

En algunos casos, las bibliotecas (i)-(v) comprenden al menos aproximadamente 10^4, al menos aproximadamente 10^5, al menos aproximadamente 10^6, o al menos aproximadamente 10^7 códigos de barras diferentes . En algunos casos, las bibliotecas (i)-(iv) comprenden al menos aproximadamente 10^4, al menos aproximadamente 10^5, al menos aproximadamente 10^6 o al menos aproximadamente 10^7 fuentes diferentes de códigos de barras clonales.

B. Enzimas

i) Transposasa, por ejemplo, una transposasa que actúa en la biblioteca de códigos de barras;
ii) ADN polimerasa (por ejemplo, ADN polimerasa I, fragmento Klenow, Taq I);
iii) Polimerasa phi29;
iv) Exonucleasa (por ejemplo, Exonucleasa III);
v) Endonucleasa de restricción;
vi) ADN ligasa;
vii) Fosfatasa alcalina;
viii) Enzimas de corte;
ix) Endonucleasa (por ejemplo, Vvn);
x) Componentes de escisión de ADN con base en uracilo o ribo (por ejemplo, uracilo ADN glicosilasa).

El kit también puede incluir uno o más tubos; un agente limitante de movilidad (por ejemplo, agarosa o PEG) y reactivos para aislar ADN de alto peso molecular de células eucariotas. Los componentes del kit pueden empacarse juntos y el empaque puede contener o ir acompañado de instrucciones impresas para usar el kit.

Composiciones

Una composición (por ejemplo, una mezcla en un solo tubo o recipiente) puede comprender cualquiera de las bibliotecas (i) - (v), descritas anteriormente, y el ADN genómico como se describió anteriormente. El ADN genómico

puede ser, por ejemplo, de un animal, tal como un mamífero (por ejemplo, Humano), una planta, un hongo. La composición puede comprender más de un genoma equivalente de ADN genómico. En varios casos, la mezcla puede comprender al menos 5 equivalentes de genoma, al menos 10 equivalentes de genoma, al menos 25 equivalentes de genoma, al menos 50 equivalentes de genoma, al menos 100 equivalentes de genoma, al menos 500 equivalentes de genoma o al menos 1000 equivalentes de genoma, tal como de 5-20 equivalentes de genoma, tal como de 5-100 equivalentes de genoma, tal como de 50-1000 equivalentes de genoma. En algunos casos, el ADN genómico comprende solo secuencias naturales y no comprende adaptadores o enlazadores. La composición puede comprender una o más enzimas seleccionadas independientemente de una transposasa, una ADN polimerasa, una endonucleasa de restricción, una ADN ligasa y fosfatasa alcalina.

Si bien esta invención se ha descrito con referencia a aspectos y realizaciones específicos, es evidente que otras realizaciones y variaciones de esta invención pueden ser desarrolladas por otros expertos en la materia sin apartarse del alcance de la invención. El alcance de la invención está definido por las reivindicaciones adjuntas.

La cita de publicaciones y documentos de patente no pretende ser una indicación de que dicho documento es una técnica anterior pertinente, ni constituye una admisión en cuanto a su contenido o fecha.

REIVINDICACIONES

1. Un método para preparar ADN genómico para el análisis de secuencias, por reacción homogénea sin el uso de compartimentación física tal como nanogotas, comprendiendo el método:

(a) combinar una pluralidad de fragmentos largos que comprenden secuencias de ADN genómico en una sola mezcla con una población de perlas, en las que (i) los fragmentos largos son de 5 kilobases a 750 kilobases de longitud, (ii) cada perla comprende al menos 1000 copias del mismo oligonucleótido inmovilizado sobre la misma, comprendiendo dicho oligonucleótido una secuencia que contiene una etiqueta, (iii) cada secuencia que contiene una etiqueta comprende una secuencia de etiqueta, y (iv) la población de perlas comprende, en conjunto, al menos 10.000 secuencias de etiquetas diferentes;

(b) producir fragmentos largos etiquetados incorporando en cada uno de una pluralidad de fragmentos largos individuales múltiples copias de una secuencia de etiqueta, en los que en una pluralidad de dichos fragmentos largos individuales dichas copias múltiples son de una sola perla;

(c) producir subfragmentos de los fragmentos largos etiquetados, en los que una pluralidad de subfragmentos del mismo fragmento largo etiquetado comprende secuencias de etiquetas de la misma perla;

en el que la etapa (b) se lleva a cabo bajo condiciones que promueven la interacción de solo una secuencia de etiqueta por fragmento largo

2. El método de la reivindicación 1, en el que la población de perlas comprende, en conjunto, al menos 100.000 secuencias de etiquetas diferentes.

3. El método de la reivindicación 1 o la reivindicación 2, en el que el ADN genómico es de un organismo diploide y la mezcla única comprende más de una cantidad haploide de ADN genómico.

4. El método de la reivindicación 1 o la reivindicación 2, en el que el ADN es un microbioma que comprende ADN bacteriano de una mezcla de bacterias.

5. El método de cualquiera de las reivindicaciones 1 a 4, en el que producir los subfragmentos en la etapa (c) comprende realizar una reacción de amplificación para producir una pluralidad de amplicones.

6. El método de cualquiera de las reivindicaciones 1 a 4, en el que las secuencias que contienen una etiqueta comprenden extremos de transposones, comprendiendo el método combinar la pluralidad de fragmentos largos con la población de perlas en condiciones adecuadas para la transposición de las secuencias de etiquetas en los fragmentos largos.

7. El método de cualquiera de las reivindicaciones 1 a 4, en el que las etapas (a) y (b) comprenden proporcionar una única mezcla que comprende (i) una transposasa, (ii) la pluralidad de fragmentos largos y (iii) la población de perlas, en el que dichas secuencias que contienen una etiqueta comprenden una secuencia de etiqueta y una secuencia del extremo del transposón, y en el que diferentes perlas comprenden diferentes secuencias de etiquetas, en el único recipiente de reacción en condiciones en las que, para una pluralidad de perlas, la secuencia de etiqueta asociada con una perla individual es transpuesta a múltiples sitios de un fragmento largo individual.

8. El método de la reivindicación 7, en el que la transposasa es una transposasa Tn5.

9. El método de las reivindicaciones 1 a 4, en el que cada fragmento largo etiquetado comprende una pluralidad de secuencias que contienen una etiqueta incorporadas en la secuencia de ADN genómico con un espaciado promedio seleccionado.

10. El método de la reivindicación 9, en el que la separación promedio está en el intervalo de 100 a 5000 bases.

11. El método de cualquier reivindicación precedente que comprende las etapas de

(d) secuenciar los subfragmentos etiquetados para producir una pluralidad de lecturas de secuencia que incluyen la secuencia de la secuencia de ácido nucleico objetivo y al menos una secuencia de etiqueta;
(e) combinar de lecturas de secuencia obtenidas en (d) para producir una secuencia o secuencias ensambladas del fragmento de ADN genómico, en el que la combinación comprende determinar que las lecturas de secuencia obtenidas en (d) se originaron a partir del mismo fragmento largo si dichas lecturas de secuencia comprenden la mismo secuencia de etiqueta

12. El método de la reivindicación 1, en el que los fragmentos largos comprenden secuencias de ADN genómico y secuencias adaptadoras.

13. El método de la reivindicación 12, en el que las secuencias que contienen una etiqueta comprenden una secuencia de etiqueta y una secuencia complementaria a las secuencias adaptadoras.

14. El método de la reivindicación 13, que comprende además hibridar oligonucleótidos que comprenden secuencias que contienen una etiqueta con la secuencia adaptadora, y extender los oligonucleótidos para formar subfragmentos etiquetados.

15. El método de la reivindicación 12, en el que, antes de combinar la pluralidad de fragmentos largos con la población de perlas, los fragmentos largos se preparan mediante corte y apertura, extensión de cebador aleatorio o inserción de transposón.

16. El método de la reivindicación 12, en el que, antes de combinar la pluralidad de fragmentos largos con la población de perlas, se introducen huecos en los fragmentos largos usando un transposón.

17. El método de la reivindicación 12, en el que antes de combinar la pluralidad de fragmentos largos con la población de perlas, los fragmentos largos se preparan cortando en ambas cadenas.

18. El método de la reivindicación 17 que comprende además secuencias de adaptador de ligadura en cada cadena en los cortes.

19. El método de la reivindicación 11, en el que el ADN genómico es de un organismo diploide y la mezcla única comprende más de una cantidad haploide de ADN genómico, que comprende además determinar un haplotipo del genoma.

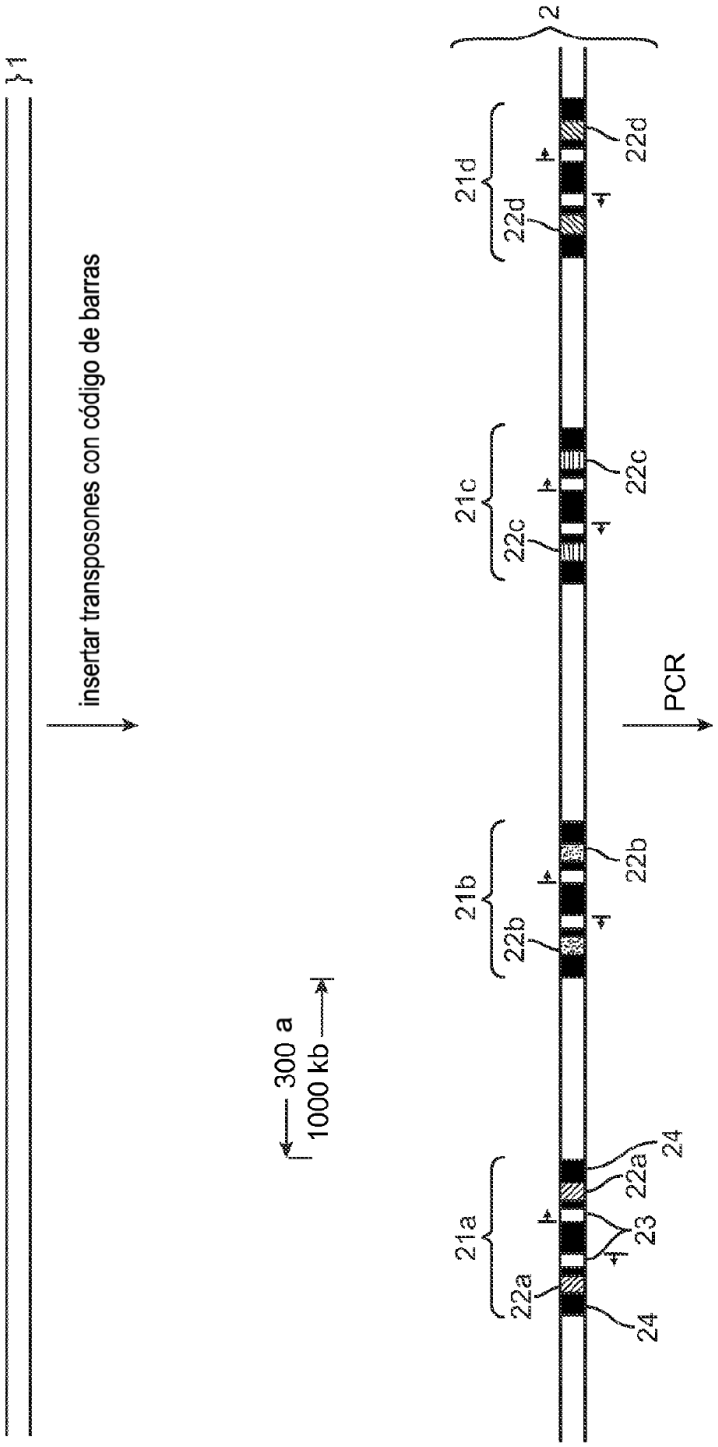

1
insertar transposones con código de barras
300 a
1000 kb
2
21a
22a
24
23
22a
24
21b
22b
22b
21c
22c
22c
21d
22d
22d
PCR

FIG. 1A

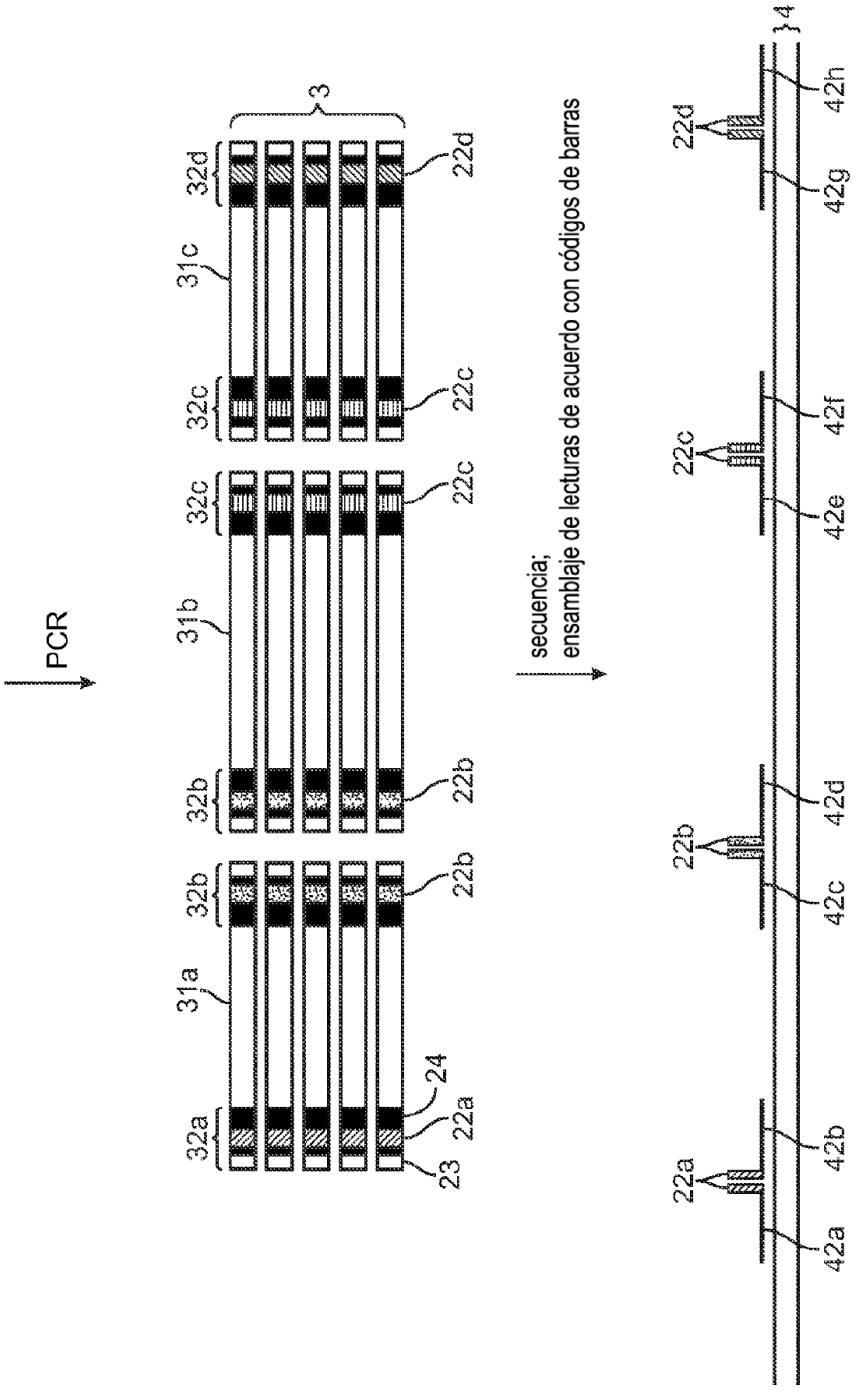

PCR
secuencia;
ensamblaje de lecturas de acuerdo con códigos de barras
3
4
31a
31b
31c
32a
32b
32c
32d
22a
22b
22c
22d
23
24
42a
42b
42c
42d
42e
42f
42g
42h

FIG. 1B

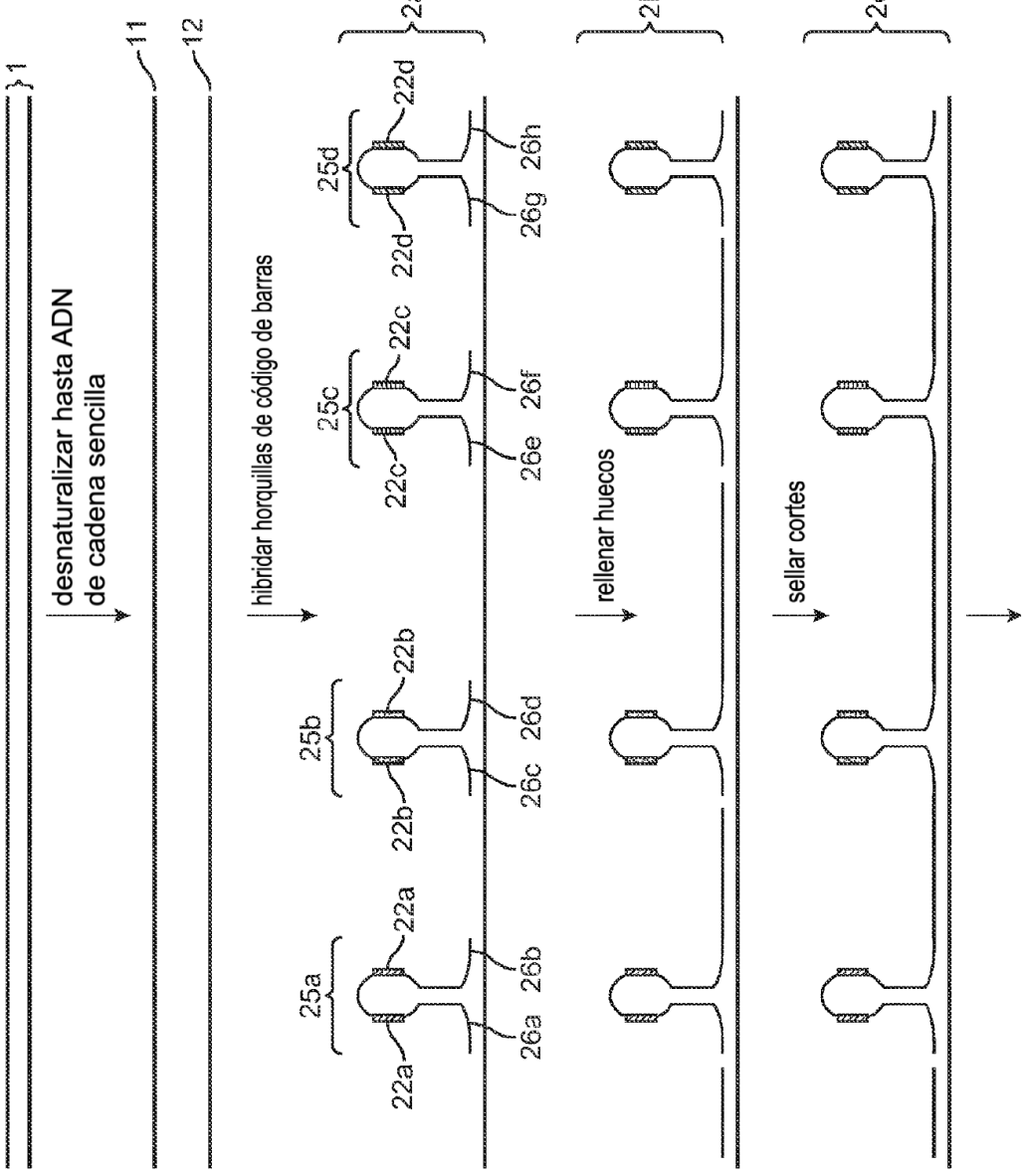

1
desnaturalizar hasta ADN
de cadena sencilla
11
12
hibridar horquillas de código de barras
25a
22a
26a
26b
25b
22b
26c
26d
25c
22c
26e
26f
25d
22d
26g
26h
2a
rellenar huecos
2b
sellar cortes
2c

FIG. 2A

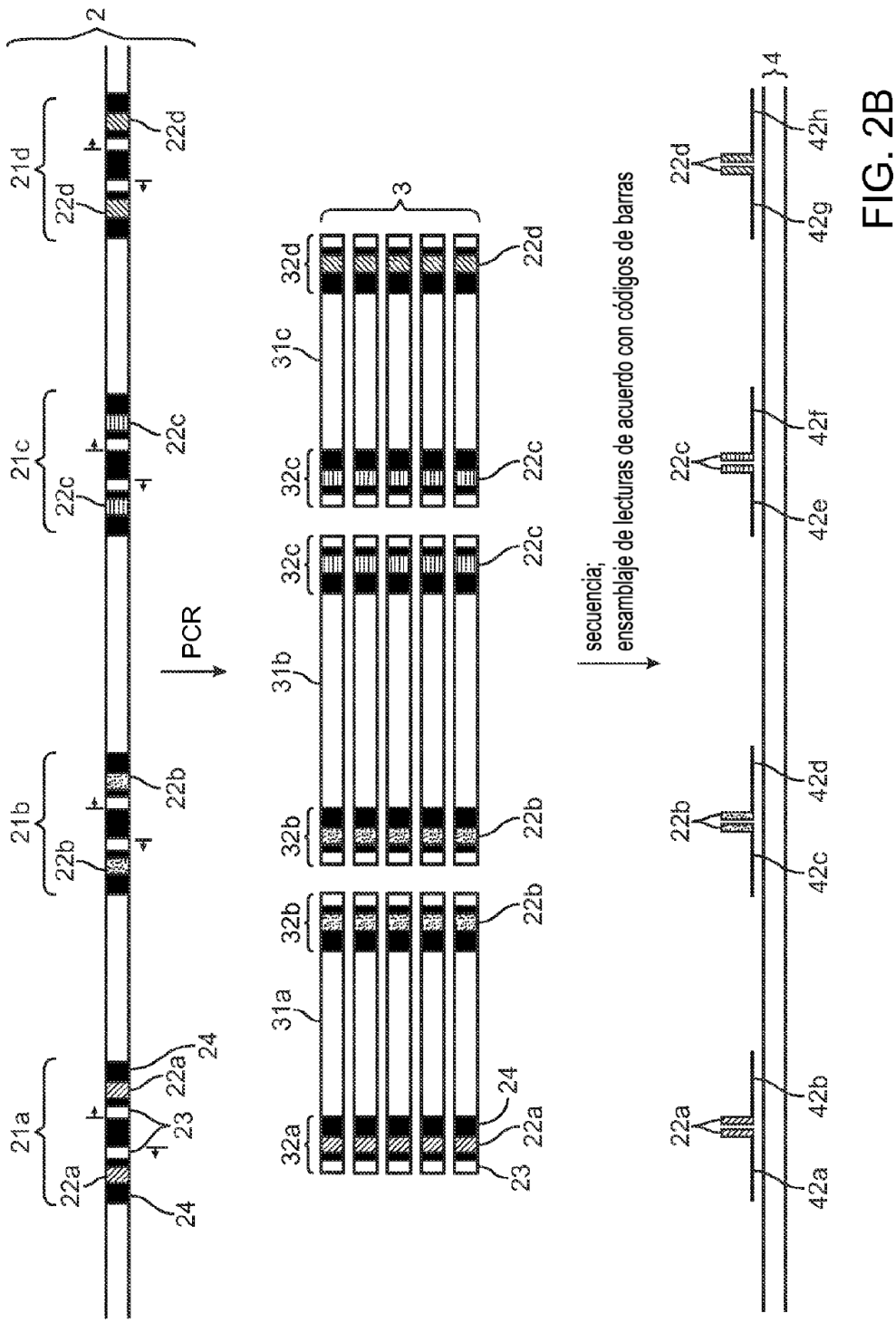

2
21a
21b
21c
21d
22a
22b
22c
22d
23
24
PCR
3
31a
31b
31c
32a
32b
32c
32d
secuencia;
ensamblaje de lecturas de acuerdo con códigos de barras
4
42a
42b
42c
42d
42e
42f
42g
42h

FIG. 2B

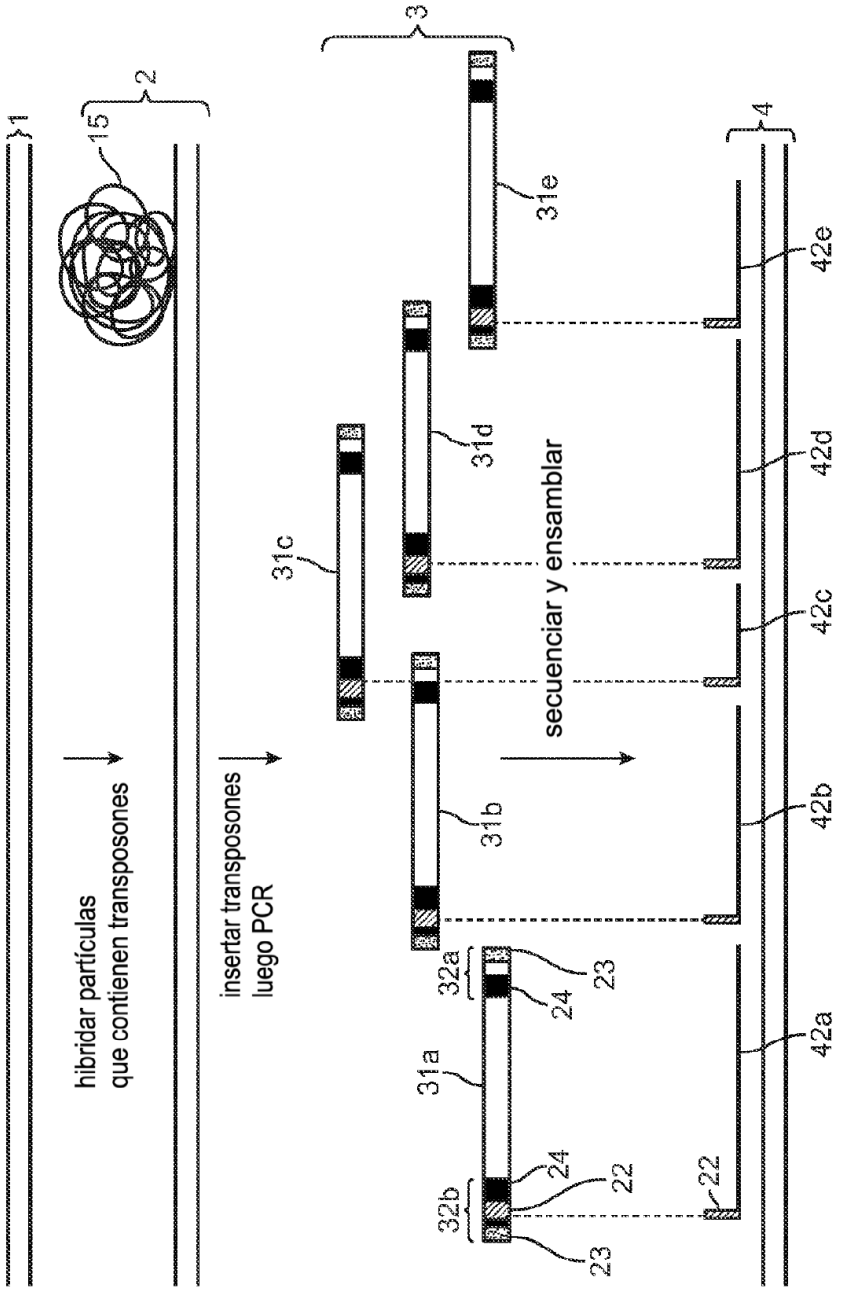

1
2
15
hibridar partículas
que contienen transposones
insertar transposones
luego PCR
3
31c
31d
31e
31b
31a
secuenciar y ensamblar
32a
32b
24
23
22
4
42e
42d
42c
42b
42a

FIG. 3

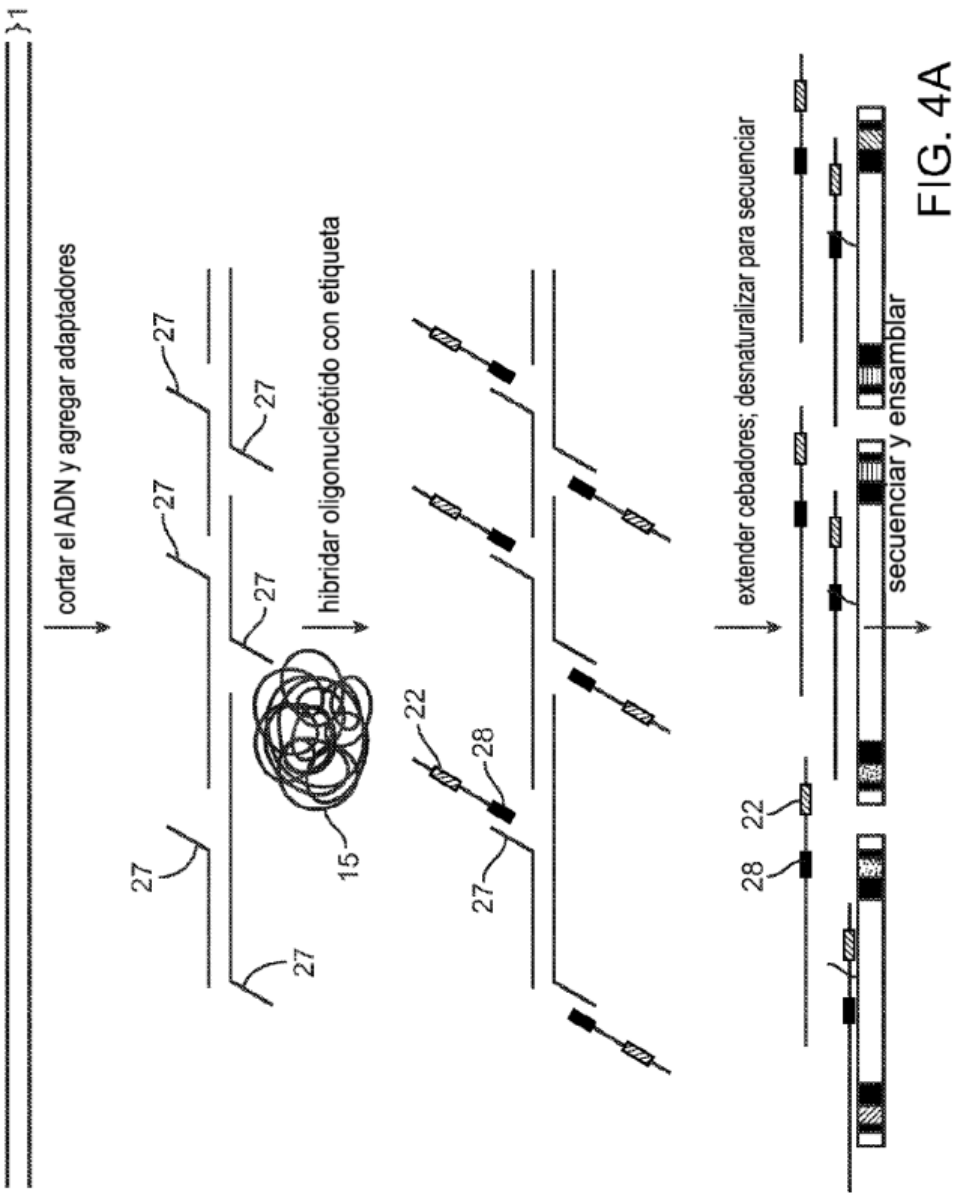

1
cortar el ADN y agregar adaptadores
27
27
27
27
27
hibridar oligonucleótido con etiqueta
15
22
28
27
extender cebadores; desnaturalizar para secuenciar
28
22
secuenciar y ensamblar

FIG. 4A

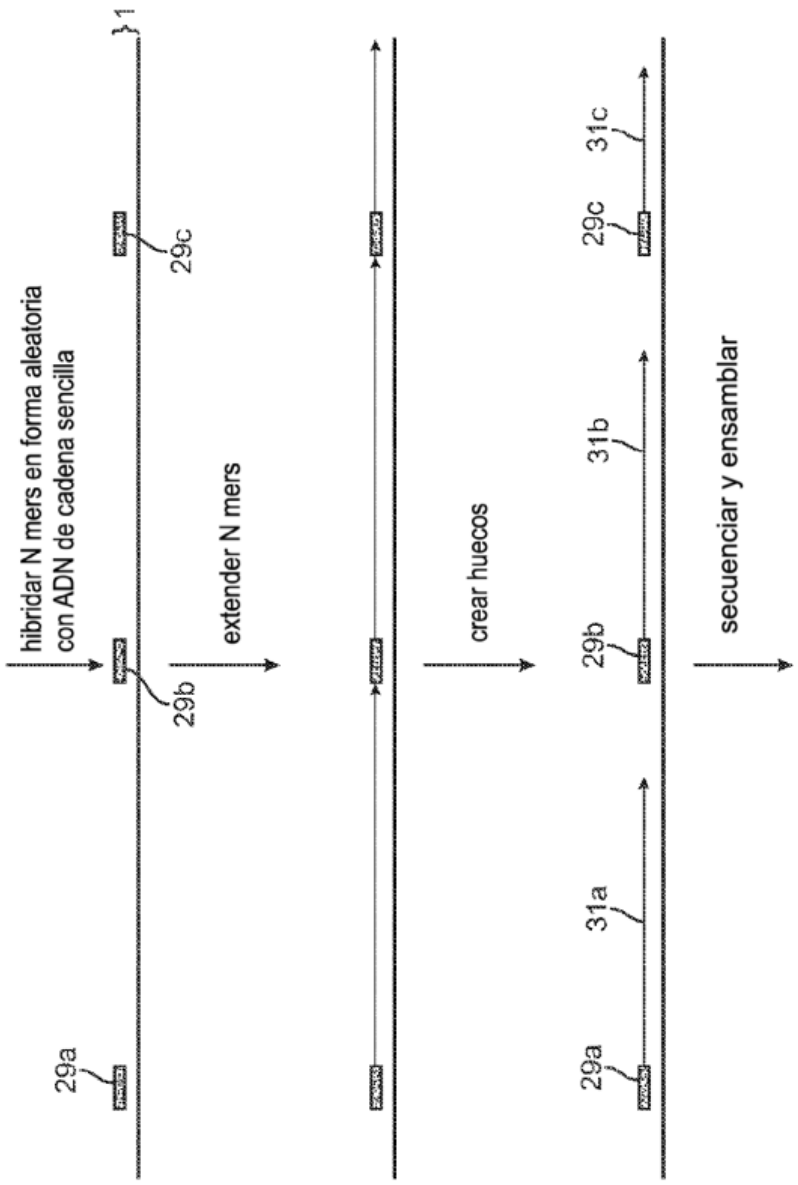

hibridar N mers en forma aleatoria con ADN de cadena sencilla
extender N mers
crear huecos
secuenciar y ensamblar
11
29a
29b
29c
31a
31b
31c

FIG. 4B

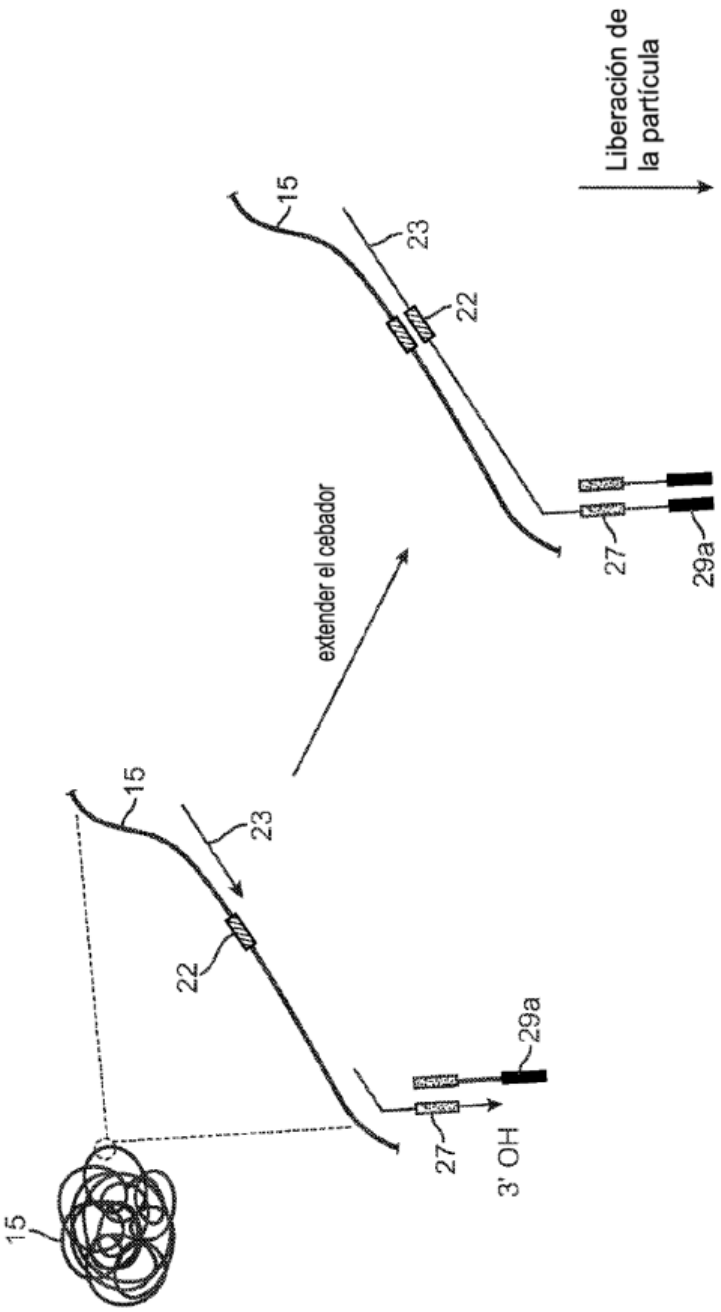

15
23
22
27
29a
3' OH
extender el cebador
Liberación de la partícula

FIG. 4C

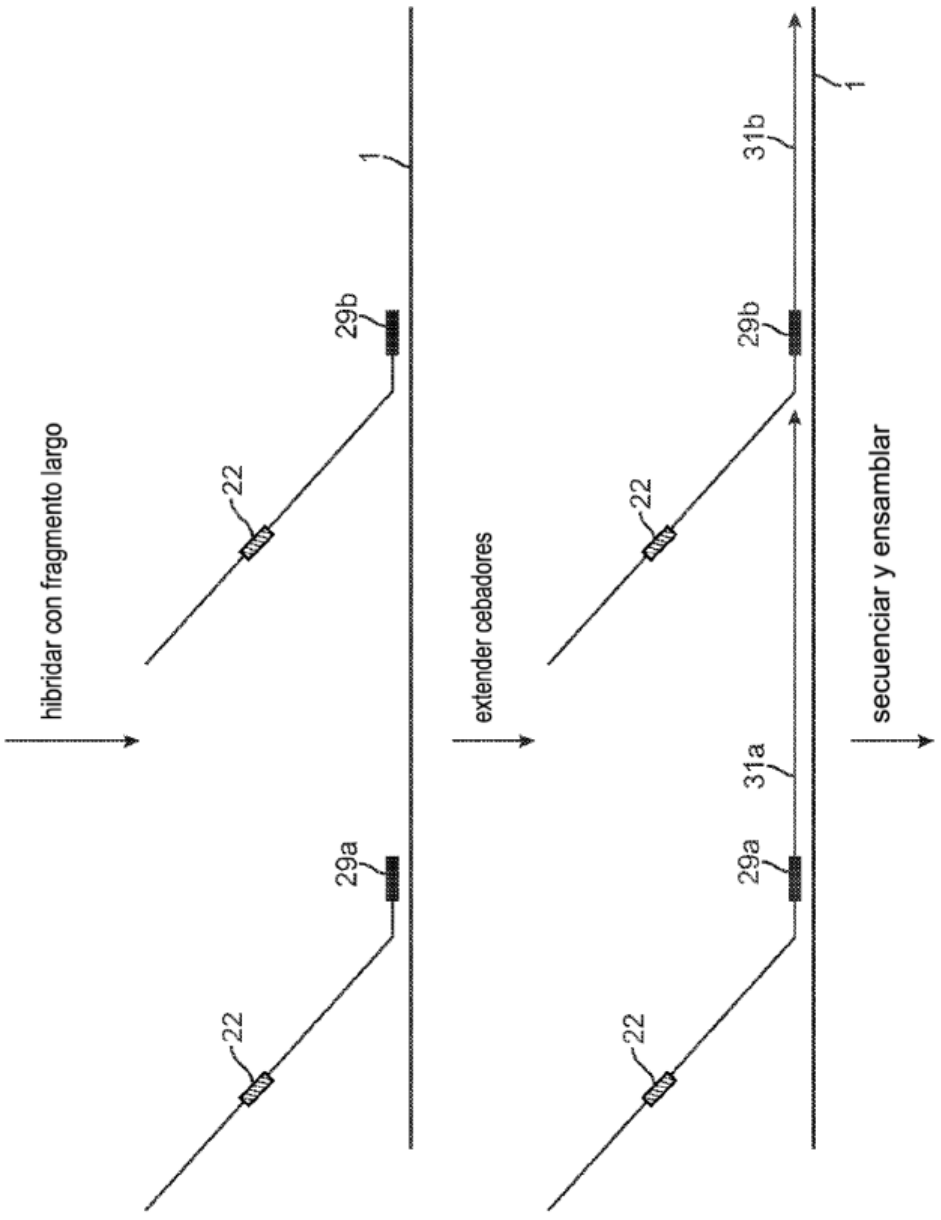

hibridar con fragmento largo
22
29b
1
22
29a
extender cebadores
22
29b
31b
1
22
31a
29a
secuenciar y ensamblar

FIG. 4D

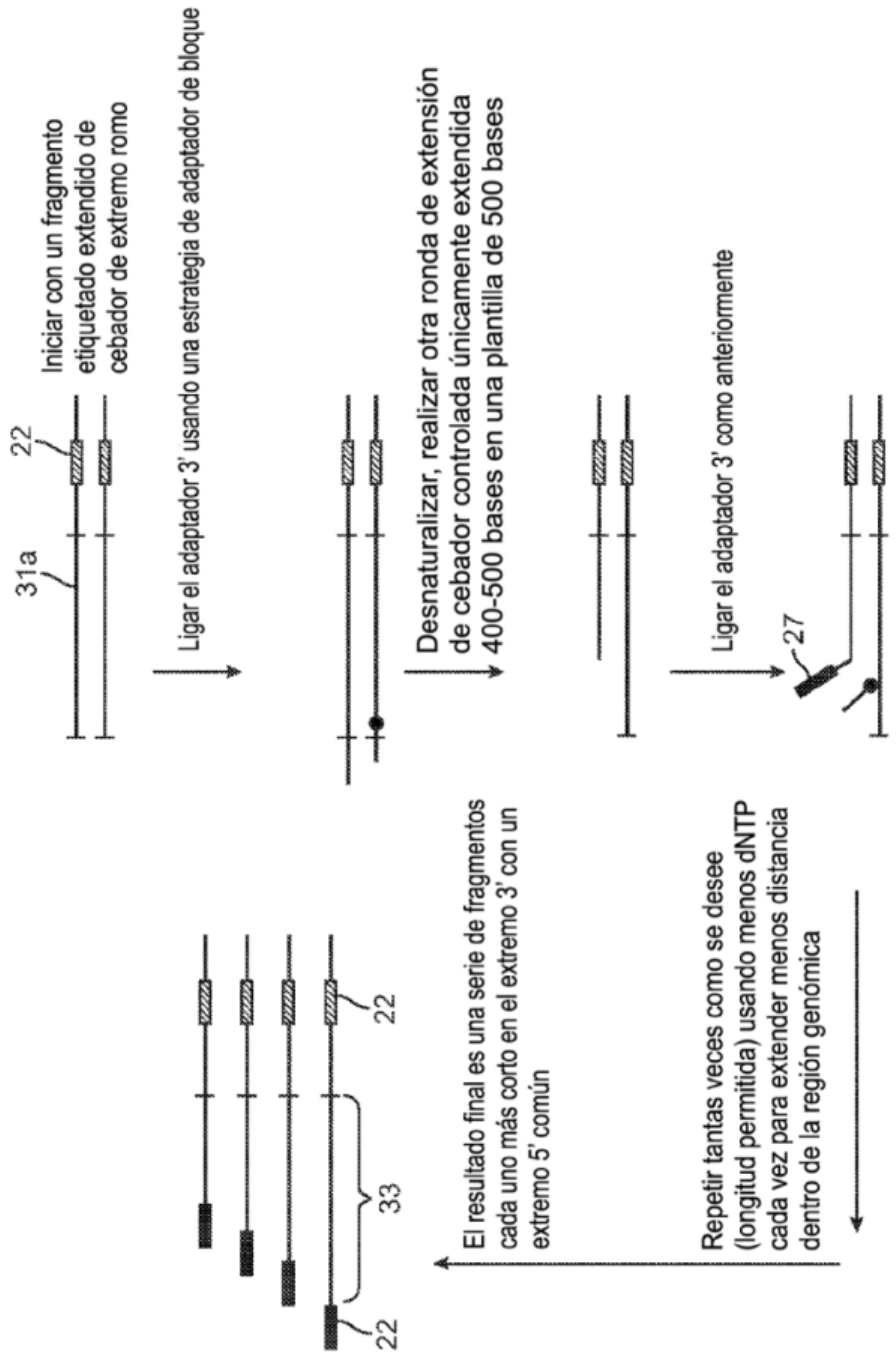

Iniciar con un fragmento etiquetado extendido de cebador de extremo romo
31a
22
Ligar el adaptador 3' usando una estrategia de adaptador de bloque
Desnaturalizar, realizar otra ronda de extensión de cebador controlada únicamente extendida 400-500 bases en una plantilla de 500 bases
Ligar el adaptador 3' como anteriormente
27
El resultado final es una serie de fragmentos cada uno más corto en el extremo 3' con un extremo 5' común
Repetir tantas veces como se desee (longitud permitida) usando menos dNTP cada vez para extender menos distancia dentro de la región genómica
22
33
22

FIG. 4E

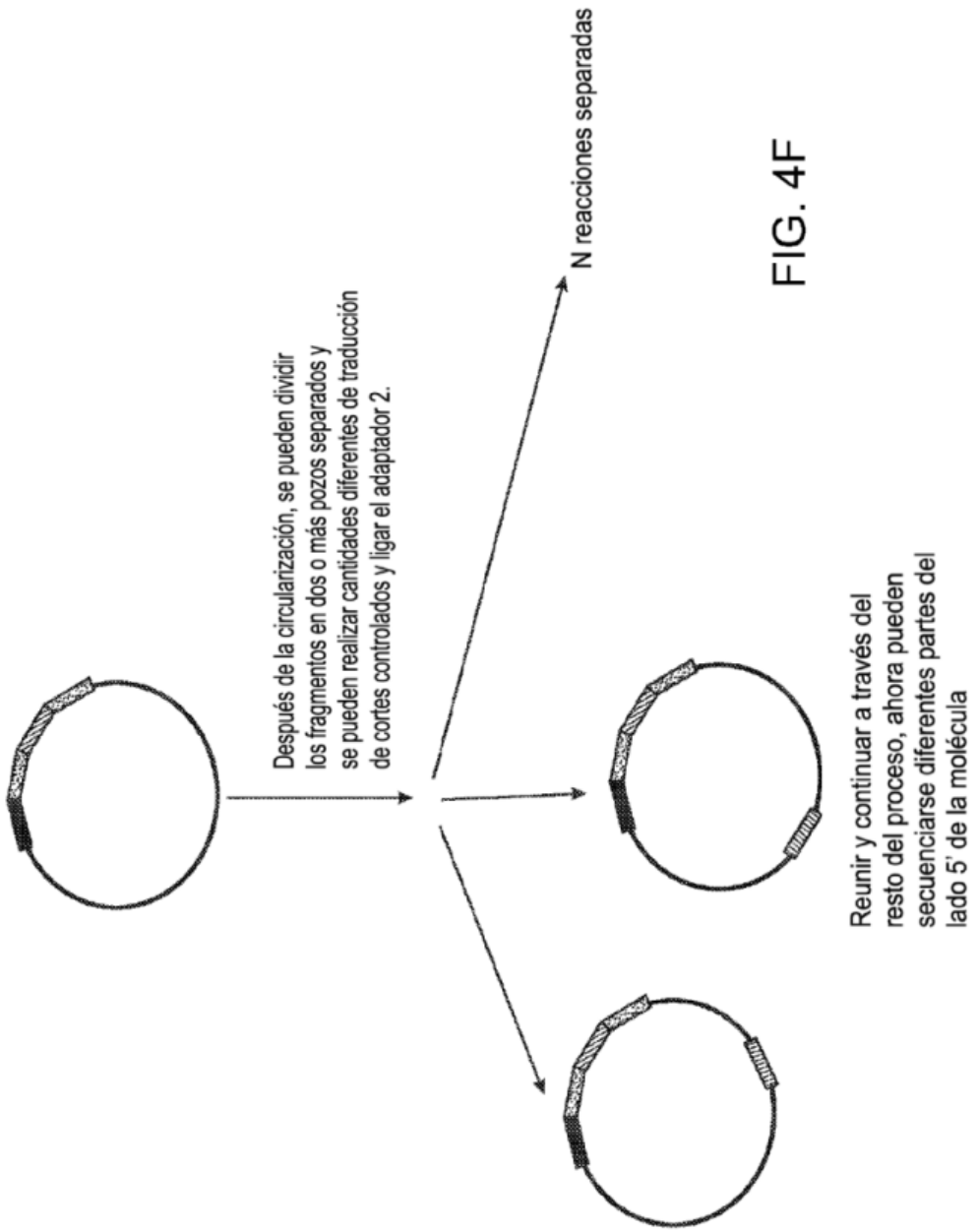

Después de la circularización, se pueden dividir los fragmentos en dos o más pozos separados y se pueden realizar cantidades diferentes de traducción de cortes controlados y ligar el adaptador 2.
N reacciones separadas
Reunir y continuar a través del resto del proceso, ahora pueden secuenciarse diferentes partes del lado 5' de la molécula

FIG. 4F

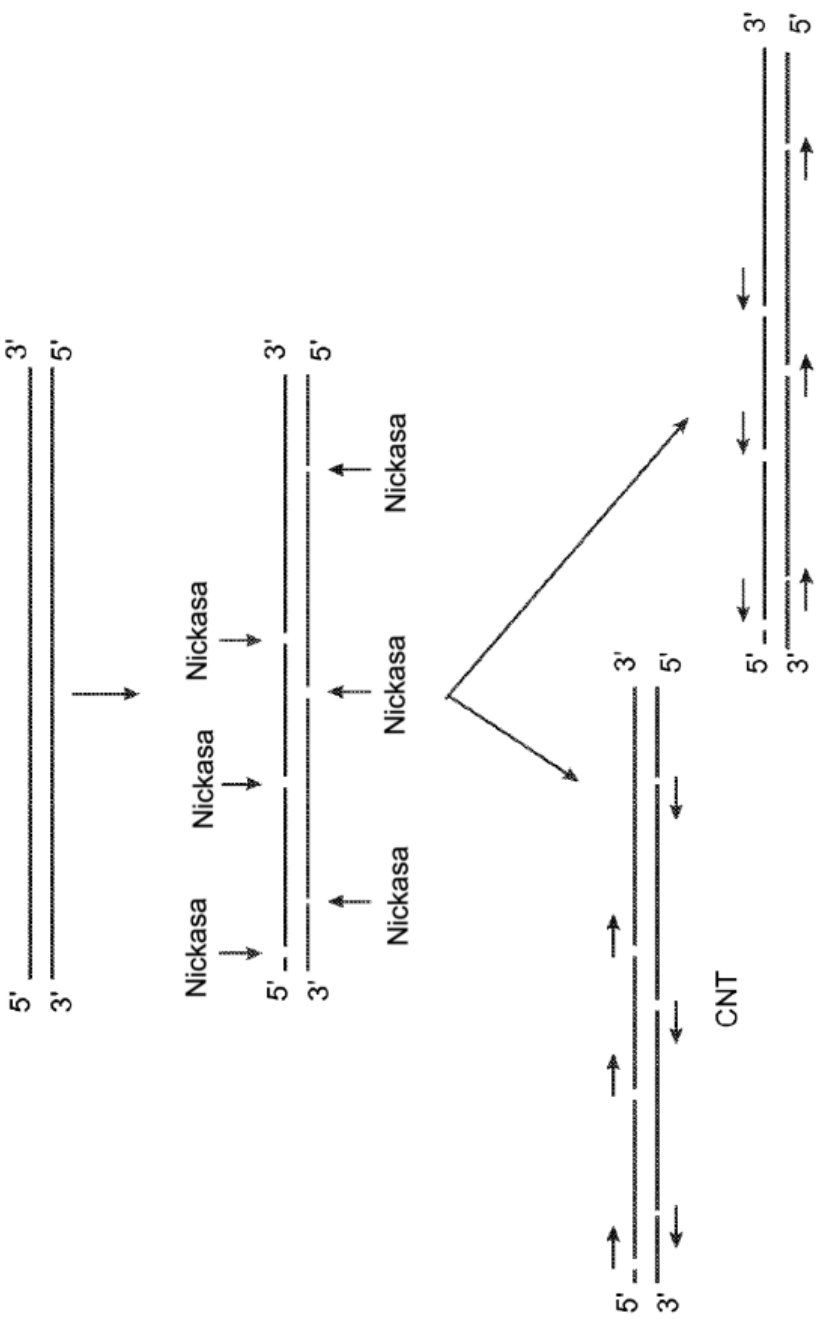

5'
3'
3'
5'
Nickasa
Nickasa
Nickasa
Nickasa
Nickasa
Nickasa
5'
3'
3'
5'
CNT
5'
3'
3'
5'
Klenow (sin dNTP)
5'
3'
3'
5'

FIG. 4G

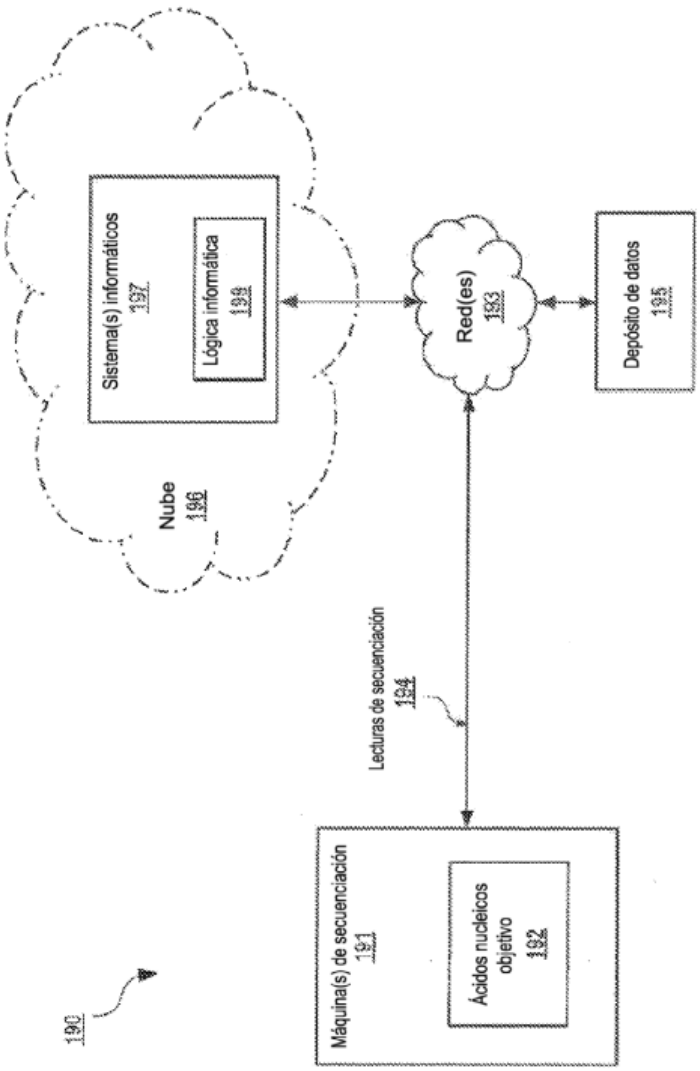

190
Máquina(s) de secuenciación
191
Ácidos nucleicos objetivo
192
Lecturas de secuenciación
194
Nube
196
Sistema(s) informáticos
197
Lógica informática
198
Red(es)
193
Depósito de datos
195

FIG. 5A

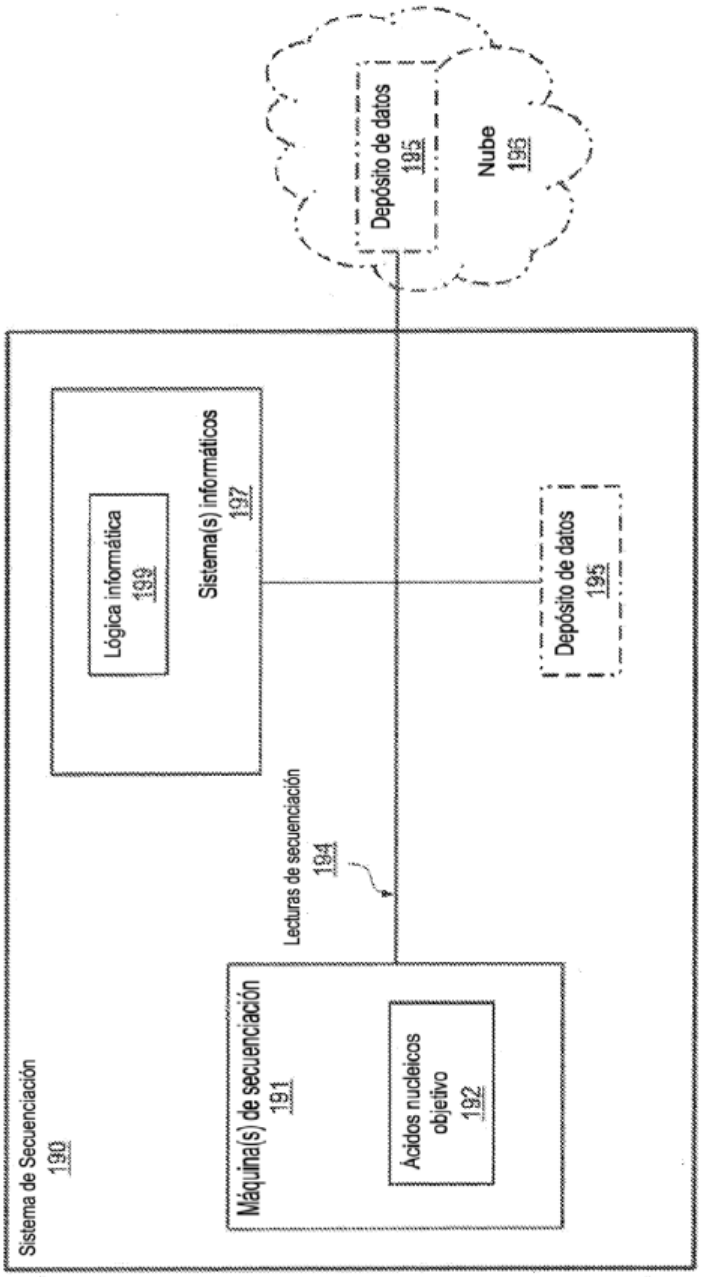

Nube
193
Depósito de datos
195
Sistema de Secuenciación
190
Sistema(s) informáticos
197
Lógica informática
199
Depósito de datos
195
Lecturas de secuenciación
194
Máquina(s) de secuenciación
191
Ácidos nucleicos objetivo
192

FIG. 5B

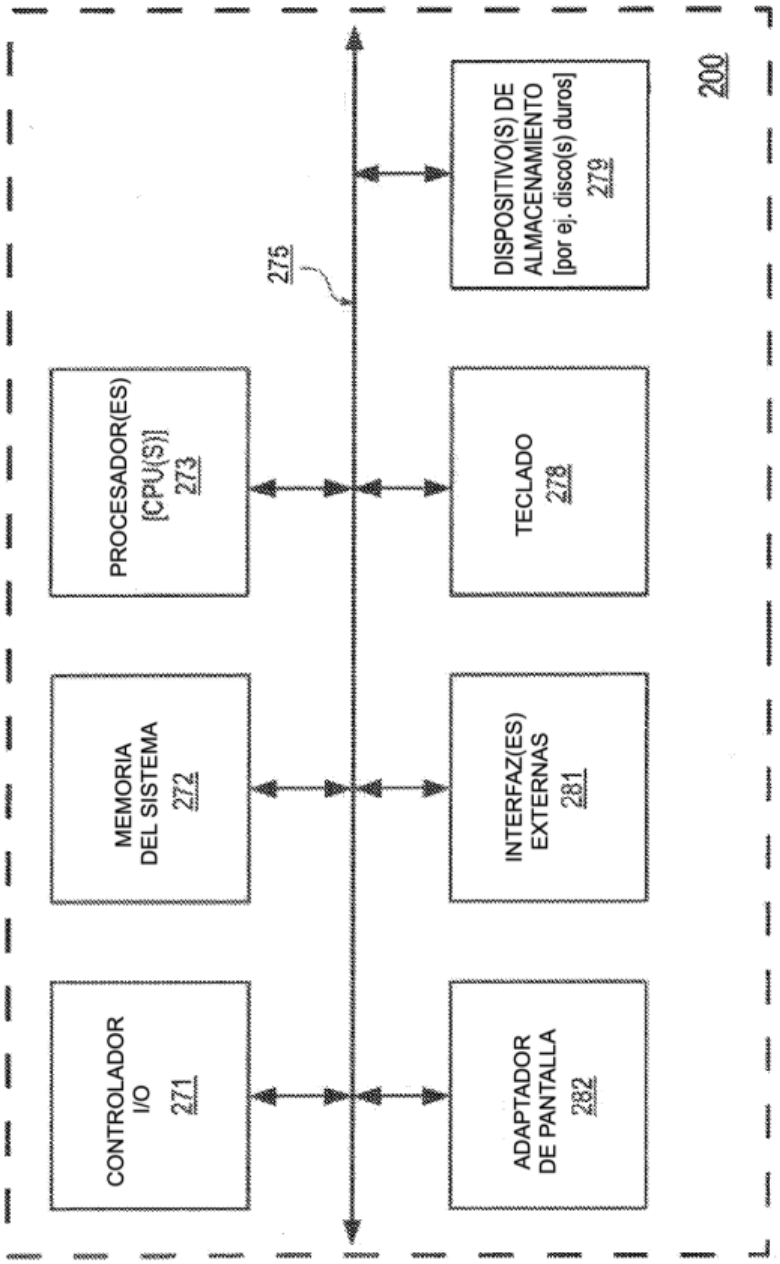

200
275
CONTROLADOR I/O
271
MEMORIA DEL SISTEMA
272
PROCESADOR(ES) [CPU(S)]
273
ADAPTADOR DE PANTALLA
282
INTERFAZ(ES) EXTERNAS
281
TECLADO
278
DISPOSITIVO(S) DE ALMACENAMIENTO [por ej. disco(s) duros]
279

FIG. 6

Construcción de gráfico a través de LFR	• Elaborar un gráfico de todos los pares de het que están dentro de la distancia esperada.
Construcción de gráfico a través de emparejamiento por pares	• Opcionalmente, poblar el gráfico de todos los pares de het que están dentro de la distancia esperada usando datos de emparejamiento por pares.
Combinación de gráficos	• Combinar los gráficos generados por LFR y por emparejamiento por pares.
Recorte del gráfico	• Opcionalmente, recortar el gráfico usando heurística.
Optimización del gráfico	• Optimizar el gráfico elaborando el árbol de expansión mínima.
Construcción del cóntigo	• Ensamblar los cóntigos usando el gráfico optimizado.
Separación universal por fases	• Usar separación por fases del trio para asignar cóntigos a los padres.

FIG. 7

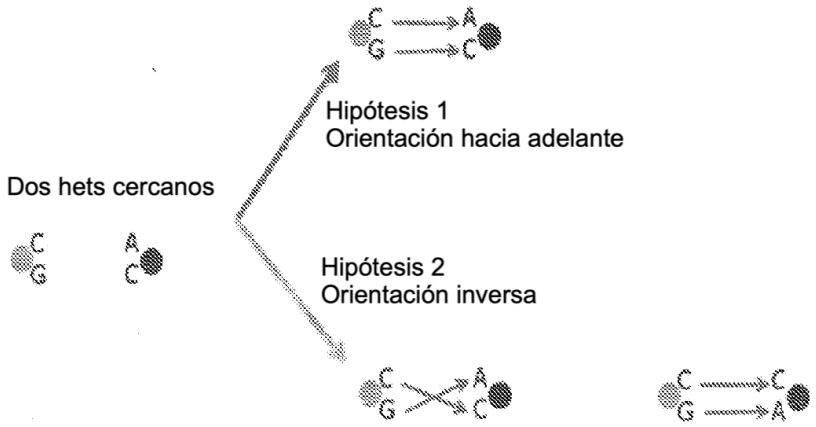
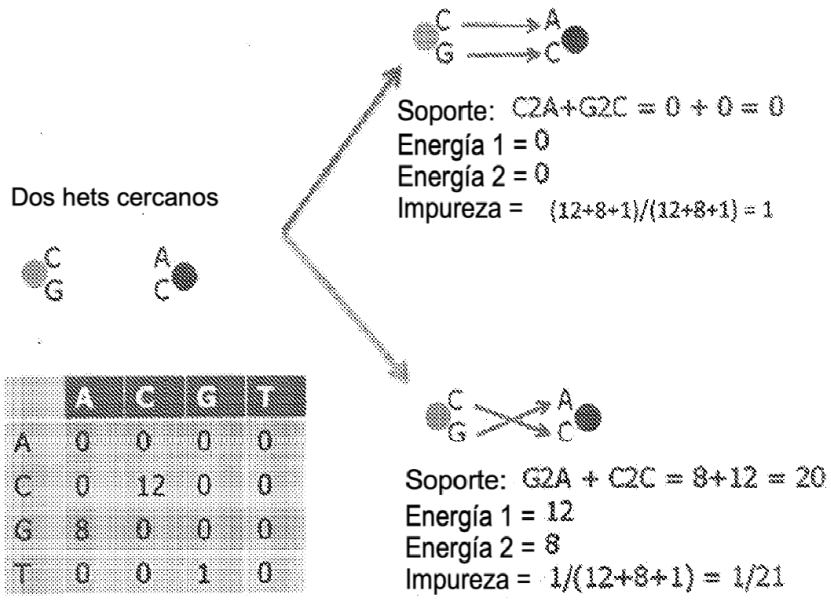
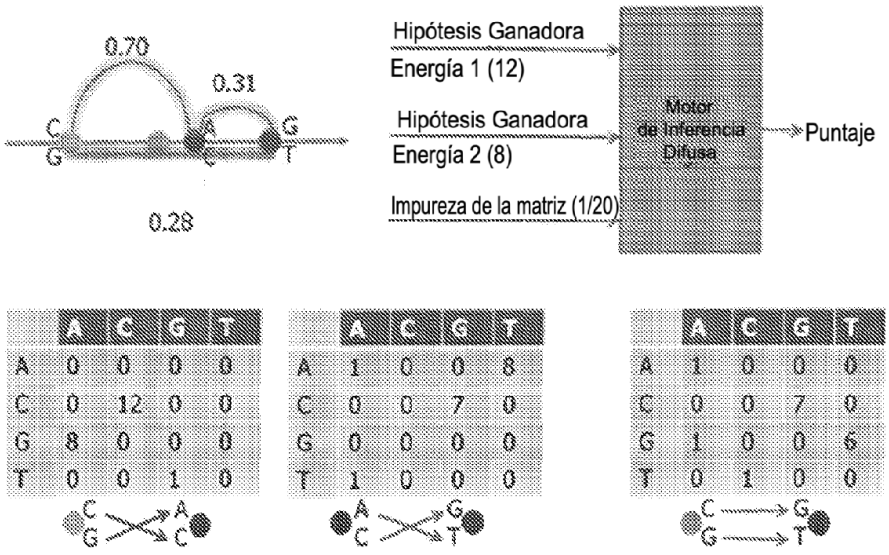


FIG. 8



	A	C	G	T
A	0	0	0	0
C	0	12	0	0
G	8	0	0	0
T	0	0	1	0

FIG. 9



	A	C	G	T
A	0	0	0	0
C	0	12	0	0
G	8	0	0	0
T	0	0	1	0

	A	C	G	T
A	1	0	0	8
C	0	0	7	0
G	0	0	0	0
T	1	0	0	0

	A	C	G	T
A	1	0	0	0
C	0	0	7	0
G	1	0	0	6
T	0	1	0	0

FIG. 10

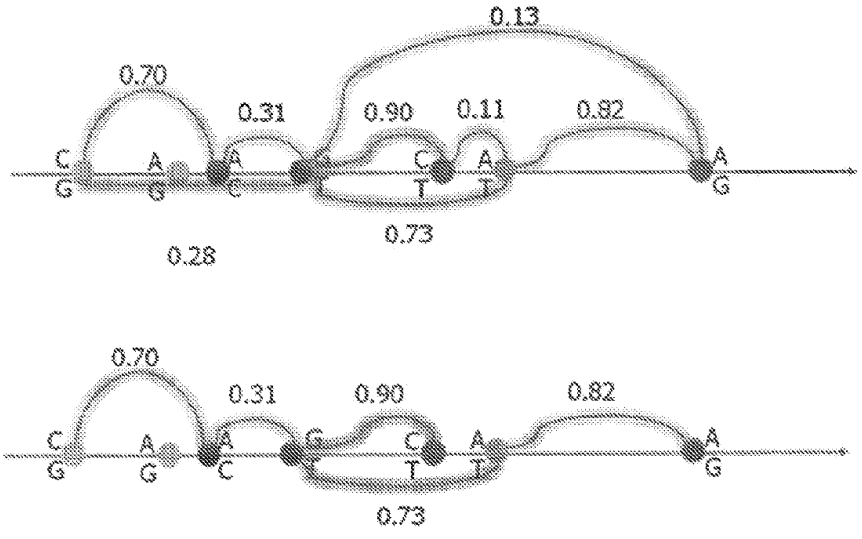
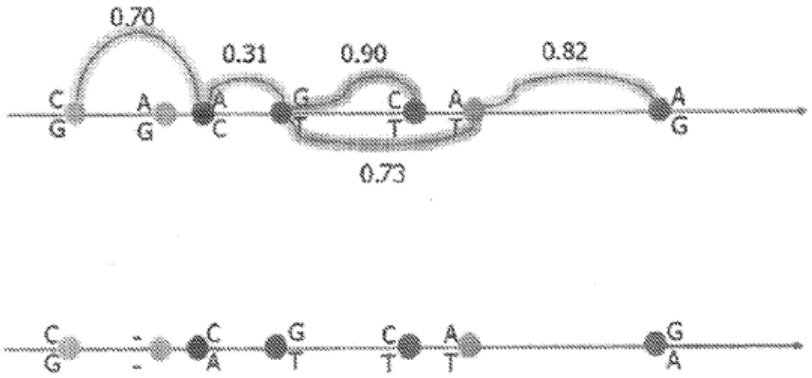


FIG. 11



Padre 1: C-CGCAG
Padre 2: G-ATTTA

FIG. 12

FIG. 13

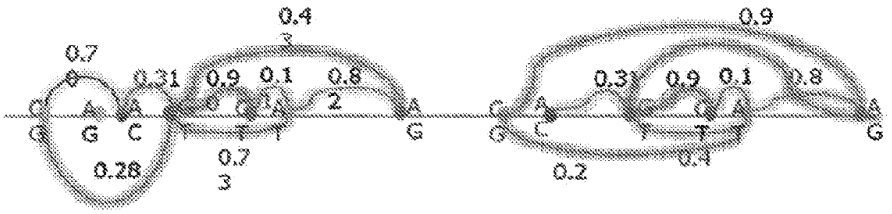

0.4
3
0.7
0
0.31
0.9
0.1
0.8
2
0.28
0.7
3
0.9
0.31
0.9
0.1
0.8
0.2
0.3

FIG. 14

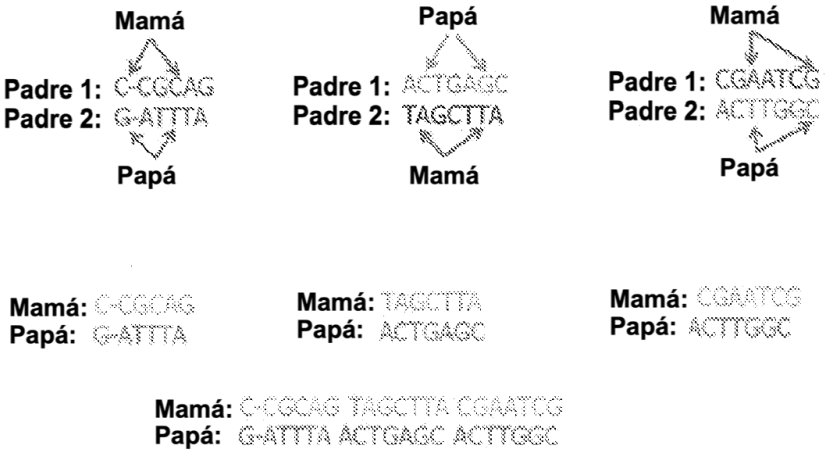

Mamá
Padre 1: C-CGCAG
Padre 2: G-ATTTA
Papá
Papá
Padre 1: ACTGAGC
Padre 2: TAGCTTA
Mamá
Mamá
Padre 1: CGAATCG
Padre 2: ACTTGGC
Papá
Mamá: C-CGCAG
Papá: G-ATTTA
Mamá: TAGCTTA
Papá: ACTGAGC
Mamá: CGAATCG
Papá: ACTTGGC
Mamá: C-CGCAG TAGCTTA CGAATCG
Papá: G-ATTTA ACTGAGC ACTTGGC

FIG. 15

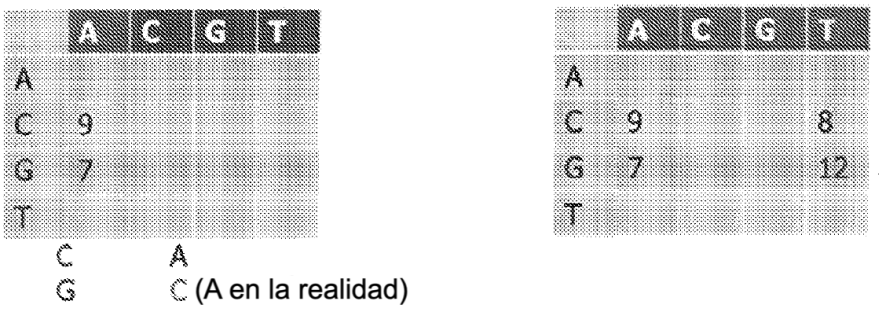

A C G T
A
C 9
G 7
T
C A
G C (A en la realidad)
A C G T
A
C 9 8
G 7 12
T

FIG. 16

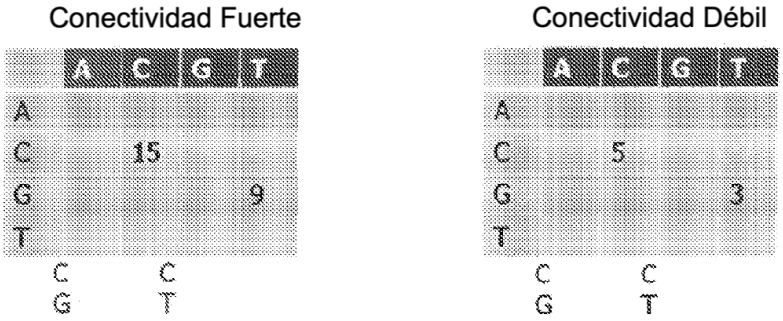

Conectividad Fuerte
A C G T
A
C 15
G 9
T
C C
G T
Conectividad Débil
A C G T
A
C 5
G 3
T
C C
G T

FIG. 17