(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 688 458**

(51) Int. CI.:

C12Q 1/6886 (2008.01)
C12Q 1/683 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **21.10.2011 PCT/US2011/057350**

(87) Fecha y número de publicación internacional: **26.04.2012 WO12054873**

(96) Fecha de presentación y número de la solicitud europea: **21.10.2011 E 11835243 (4)**

(97) Fecha y número de publicación de la concesión europea: **25.04.2018 EP 2630263**

(54) Título: **Recuento varietal de ácidos nucleicos para obtener información del número de copias genómicas**

(30) Prioridad:

22.07.2011 US 201161510579 P
21.10.2011 US 201113278333
22.10.2010 US 406067 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
02.11.2018

(73) Titular/es:

COLD SPRING HARBOR LABORATORY (100.0%)
1 Bungtown Road
Cold Spring Harbor, NY 11724, US

(72) Inventor/es:

HICKS, JAMES;
NAVIN, NICHOLAS;
TROGE, JENNIFER;
WANG, ZIHUA y
WIGLER, MICHAEL

(74) Agente/Representante:

ARIAS SANZ, Juan

DESCRIPCIÓN

Recuento varietal de ácidos nucleicos para obtener información del número de copias genómicas

Esta solicitud reivindica la prioridad del documento estadounidense con n.º de serie 13/278.333, presentado el 21 de octubre de 2011, y las solicitudes provisionales estadounidenses n.os 61/510.579, presentada el 22 de julio de 2011, y 61/406.067, presentada el 22 de octubre de 2010.

La invención dada a conocer en el presente documento recibió el apoyo del gobierno con la subvención n.º W81XWH-09-1-0591 del Departamento de defensa. Por consiguiente, el gobierno de los Estados Unidos tiene ciertos derechos en esta invención.

A lo largo de toda esta solicitud, se hace referencia a diversas publicaciones mediante números entre paréntesis. Las citas completas de estas referencias pueden encontrarse al final de la memoria descriptiva, inmediatamente antes de las reivindicaciones.

Antecedentes de la invención

Habitualmente se obtiene información del número de copias genómicas usando amplificación del genoma completo (WGA). El problema generalizado con el método de WGA es la toma de muestras excesiva de determinadas regiones, proporcionando una amplificación no uniforme del genoma (1). Los métodos de WGA comienzan con la etapa que inicia el procedimiento, una polimerasa (Phi 29) forma una cadena a partir de ADN genómico usando un cebador al azar acoplado a un adaptador para una PCR posterior (figura 1). Si las cadenas de ADN de entrada se denominan “derivadas de orden 0”, y la primera cadena sintetizada se denomina “primera derivada”, las cadenas posteriores se denominan derivadas de orden (n+1) si su molde era una derivada de orden n. Sólo las cadenas que son 2as derivadas o más se amplifican en la etapa de PCR, dando como resultado un “apilamiento” sobre las regiones “elegidas” por la polimerasa para la primera derivada.

La cobertura del genoma mediante secuenciación de WGA de ADN de una célula individual está limitada por el fenómeno de apilamiento (figura 2). Por tanto, es difícil obtener mediciones de una célula individual, particularmente cuando se basan en WGA, debido a distorsiones que se originan a partir de etapas de toma de muestras y amplificación estocásticas. Además, el método actual de WGA es una caja negra, con los reactivos sin especificar adquiridos de un proveedor, lo cual dificulta la optimización. Además, el método de WGA no se extiende para dar un método que pueda usarse para determinar el perfil de ARN de una célula individual.

La PCR mediada por ligación se desarrolló en un intento por resolver los problemas anteriormente identificados inherentes en WGA. En este método, se ligan adaptadores a un producto de digestión con endonucleasa de restricción MseI de ADN genómico a partir de una célula individual, seguido por amplificación mediante PCR usando cebadores complementarios a los adaptadores. Después se usa el ADN amplificado para secuenciación de ADN o CGH (2, 3). Sin embargo, al igual que WGA, el método todavía requiere una etapa de amplificación.

Parameswaran *et al*. (2007) y patente estadounidense n.º 7.622.281 describen métodos de marcaje de moléculas de ácido nucleico con códigos de barras con el fin de identificar el origen de las moléculas de ácido nucleico, permitiendo así una secuenciación de alto rendimiento de múltiples muestras (4, 5).

Eid *et al*. (2009) describen un método de secuenciación de una molécula individual en el que se obtienen datos de secuenciación en tiempo real de una molécula individual a partir de una ADN polimerasa que realiza una síntesis dirigida por molde sin interrupciones usando cuatro dNT marcados de manera fluorescente distinguibles (6). Sin embargo, estos métodos no proporcionan información genómica no afectada por distorsión por amplificación.

Miner *et al*. (2004) describen un método de aplicación de códigos de barras moleculares para marcar ADN de molde antes de la amplificación mediante PCR, y notifican que el método permite la identificación de secuencias contaminantes y redundantes contando únicamente secuencias marcadas con etiqueta de manera diferenciada (22). La patente estadounidense n.º 7.537.897 describe métodos para el recuento molecular mediante marcaje de moléculas de una muestra de entrada con etiquetas de oligonucleótidos únicas y posteriormente amplificación y recuento del número de etiquetas diferentes (23). Miner *et al*. y la patente estadounidense n.º 7.537.897 describen ambos el marcaje de moléculas de ácido nucleico de entrada mediante ligación, lo cual se ha encontrado que es una reacción ineficaz.

McCloskey *et al*. (2007) describen un método de codificación molecular que no usa ligación, sino que en su lugar usa cebadores específicos de molde para marcar con código de barras moléculas de ADN de molde antes de la amplificación mediante PCR (24). Sin embargo, un método de este tipo requiere que se preparen cebadores específicos de molde para cada especie de molécula de ADN de molde estudiada.

La solicitud estadounidense n.º 2007/172873 describe un método de estimación de un número de polinucleótidos diana en una mezcla, comprendiendo el método las etapas de: marcar mediante toma de muestras cada polinucleótido diana en la mezcla de modo que sustancialmente todos los polinucleótidos diana tienen una etiqueta de oligonucleótido única; amplificar las etiquetas de oligonucleótidos de los polinucleótidos diana marcados; y

determinar el número de etiquetas de oligonucleótidos diferentes en una muestra de etiquetas de oligonucleótidos amplificadas mediante determinación de secuencias de nucleótidos de las mismas. Sin embargo, este método está limitado al análisis de un locus específico y no está previsto para aplicaciones en todo el genoma.

Tal como se describe en el presente documento, la obtención de información del número de copias genómicas precisa mediante secuenciación de alto rendimiento de ADN genómico preparado mediante métodos de WGA se ve dificultada por las distorsiones en el número de copias introducidas mediante amplificación no uniforme de ADN genómico. Por tanto, existe una necesidad de un método que permita la determinación del número de copias libre de distorsiones provocadas por etapas de amplificación y que permita una determinación del número de copias precisa y eficaz de muestras complejas. Un método de este tipo debe ser robusto usando metodologías existentes para la secuenciación masiva en paralelo en alto volumen.

Sumario de la invención

Se proporciona un método para obtener, a partir de material genómico, información del número de copias genómicas no afectada por distorsión por amplificación, que comprende:

a) obtener segmentos del material genómico;

b) etiquetar los segmentos con etiquetas de ácido nucleico para generar moléculas de ácido nucleico con etiqueta únicas, de manera que cada una de las moléculas de ácido nucleico con etiqueta únicas comprende un segmento del material genómico de la etapa (a) y una etiqueta;

c) someter las moléculas de ácido nucleico con etiqueta a amplificación mediante reacción en cadena de la polimerasa (PCR);

d) generar lecturas de secuencias asociadas a etiquetas mediante secuenciación del producto de la etapa (c);

e) asignar cada molécula de ácido nucleico con etiqueta a una ubicación en un genoma asociado con el material genómico mediante mapeo de la subsecuencia de cada lectura de secuencia asociada a etiqueta correspondiente a un segmento del material genómico a una ubicación en el genoma; y

f) contar el número de moléculas de ácido nucleico con etiqueta que tienen una etiqueta diferente que se han asignado a la misma ubicación en el genoma,

obteniendo así información del número de copias genómicas no afectada por distorsión por amplificación.

También se da a conocer un método para obtener, a partir de transcritos de ARNm, información del número de copias de ARNm no afectada por distorsión por amplificación, que comprende:

a) obtener ADNc a partir de los transcritos de ARNm;

b) opcionalmente obtener segmentos del ADNc;

c) etiquetar el ADNc o los segmentos de ADNc con etiquetas sustancialmente únicas para generar moléculas de ácido nucleico con etiqueta, de manera que cada molécula de ácido nucleico con etiqueta comprende

i) un ADNc de la etapa (a) o un segmento del ADNc de la etapa (b), y

ii) una etiqueta;

d) someter las moléculas de ácido nucleico con etiqueta a amplificación mediante reacción en cadena de la polimerasa (PCR);

e) generar lecturas de secuencias asociadas a etiquetas mediante secuenciación del producto de la etapa (d);

f) asignar cada molécula de ácido nucleico con etiqueta a una ubicación en una biblioteca de ADNc asociada con los transcritos de ARNm mediante mapeo de la subsecuencia de cada lectura de secuencia asociada a etiqueta correspondiente a un ADNc o un segmento de ADNc a una ubicación en la biblioteca de ADNc; y

g) contar el número de moléculas de ácido nucleico con etiqueta que tienen una etiqueta diferente que se han asignado a la misma ubicación en la biblioteca de ADNc,

obteniendo así información del número de copias de ARNm no afectada por distorsión por amplificación.

También se proporciona un método para obtener, a partir de transcritos de ARNm, información del número de copias de ARNm no afectada por distorsión por amplificación, que comprende:

a) generar moléculas de ácido nucleico con etiqueta únicas, que comprende:

i) someter los transcritos de ARNm a una reacción de polimerasa en condiciones que fomentan la formación de tan sólo un complemento, generando así cadenas derivadas de primer orden;

ii) añadir una cola de polinucleótido a las cadenas derivadas de primer orden; y

iii) someter las cadenas derivadas de primer orden a una reacción de polimerasa en presencia de cebadores que pueden hibridarse con la cola de polinucleótido añadida en la etapa (ii) en condiciones que fomentan la formación de tan sólo un complemento, generando así cadenas derivadas de segundo orden,

en el que los cebadores de al menos una de las etapas (i) y (iii) comprenden etiquetas de ácido nucleico, de manera que cada molécula de ácido nucleico con etiqueta es única generando así moléculas de ácido nucleico con etiqueta únicas;

b) someter las moléculas de ácido nucleico con etiqueta a amplificación mediante reacción en cadena de la polimerasa (PCR);

c) generar lecturas de secuencias asociadas a etiquetas mediante secuenciación del producto de la etapa (b);

d) asignar cada molécula de ácido nucleico con etiqueta a una ubicación en una biblioteca de ADNc asociada con los transcritos de ARNm mediante mapeo de la subsecuencia de cada lectura de secuencia asociada a etiqueta correspondiente a un transcrito de ARNm a una ubicación en la biblioteca de ADNc; y

e) contar el número de moléculas de ácido nucleico con etiqueta que tienen una etiqueta diferente que se han asignado a la misma ubicación en la biblioteca de ADNc,

obteniendo así información del número de copias de ARNm no afectada por distorsión por amplificación.

También se da a conocer un método para obtener, a partir de material genómico, información de metilación de ADN no afectada por distorsión por amplificación, que comprende:

a) obtener segmentos del material genómico;

b) añadir una cola de polinucleótido a los extremos de los segmentos del material genómico para generar cadenas derivadas de orden cero;

c) someter las cadenas derivadas de orden cero a una reacción de polimerasa en presencia de cebadores esencialmente únicos, en el que los cebadores esencialmente únicos pueden hibridarse con la cola de polinucleótido de las cadenas de orden cero, en condiciones que fomentan la formación de tan sólo un complemento, generando así moléculas de ácido nucleico con etiqueta;

d) separar las moléculas de ácido nucleico con etiqueta para dar un grupo que consiste en moléculas de ácido nucleico con etiqueta semimetiladas y un grupo que consiste en moléculas de ácido nucleico con etiqueta sin metilar;

e) someter cada grupo de la etapa (d) a amplificación mediante reacción en cadena de la polimerasa (PCR);

f) generar lecturas de secuencias asociadas a etiquetas mediante secuenciación del producto de la etapa (e);

g) asignar cada molécula de ácido nucleico con etiqueta a una ubicación en un genoma asociado con el material genómico mediante mapeo de la subsecuencia de cada lectura de secuencia asociada a etiqueta correspondiente a un segmento del material genómico a una ubicación en el genoma; y

h) contar el número de moléculas de ácido nucleico con etiqueta que tienen una etiqueta diferente que se han asignado a la misma ubicación en el genoma,

obteniendo así información de metilación de ADN no afectada por distorsión por amplificación.

También se da a conocer una composición de materia derivada de material genómico que comprende moléculas de ácido nucleico con etiqueta, produciéndose dichas moléculas de ácido nucleico con etiqueta mediante un procedimiento que comprende:

a) obtener segmentos del material genómico;

b) añadir una cola de polinucleótido a los extremos de los segmentos del material genómico para generar cadenas derivadas de orden cero;

c) someter las cadenas derivadas de orden cero de la etapa (b) a una reacción de polimerasa en presencia de cebadores que pueden hibridarse con la cola de polinucleótido de las cadenas de orden cero en condiciones que fomentan la formación de tan sólo un complemento, generando así cadenas derivadas de primer orden;

d) añadir una cola de polinucleótido a las cadenas derivadas de primer orden;

e) someter las cadenas derivadas de primer orden a una reacción de polimerasa en presencia de cebadores que pueden hibridarse con la cola de polinucleótido de las cadenas derivadas de primer orden en condiciones que fomentan la formación de tan sólo un complemento, generando así cadenas derivadas de segundo orden,

en el que los cebadores de al menos una de las etapas (c) y (e) comprenden cebadores sustancialmente únicos, generando así moléculas de ácido nucleico con etiqueta.

También se da a conocer una composición de materia derivada de material genómico que comprende moléculas de ácido nucleico con etiqueta, produciéndose dichas moléculas de ácido nucleico con etiqueta mediante un procedimiento que comprende:

a) obtener segmentos de material genómico;

b) añadir una cola de polinucleótido a los extremos de los segmentos de material genómico para generar cadenas derivadas de orden cero,

c) someter las cadenas derivadas de orden cero de la etapa (b) a una reacción de ligación en presencia de cebadores que pueden hibridarse con la cola de polinucleótido de las cadenas de orden cero añadidas en la etapa (b) en condiciones que fomentan la ligación de un cebador a los extremos 5’ de las cadenas derivadas de orden cero,

d) someter el producto de la etapa (c) a una reacción de polimerasa en presencia de cebadores que pueden hibridarse con la cola de polinucleótido añadida en la etapa (b) en condiciones que fomentan la formación de tan sólo un complemento, en el que los cebadores de la etapa (d) tienen secuencias de nucleótidos diferentes de los cebadores de la etapa (c), y en el que la polimerasa tiene actividad de corrección de lectura 3’-5’,

en el que los cebadores de al menos una de las etapas (c) y (d) comprenden cebadores sustancialmente únicos, generando así moléculas de ácido nucleico con etiqueta.

También se da a conocer una composición de materia derivada de transcritos de ARNm que comprenden moléculas de ácido nucleico con etiqueta, produciéndose dichas moléculas de ácido nucleico con etiqueta mediante un procedimiento que comprende:

a) obtener ADNc a partir de los transcritos de ARNm;

b) opcionalmente obtener segmentos del ADNc; y

c) etiquetar el ADNc o los segmentos de ADNc con etiquetas sustancialmente únicas para generar moléculas de ácido nucleico con etiqueta, en el que cada molécula de ácido nucleico con etiqueta comprende

i) un ADNc de la etapa (a) o un segmento del ADNc de la etapa (b) y

ii) una etiqueta.

También se da a conocer una composición de materia derivada de transcritos de ARNm que comprenden moléculas de ácido nucleico con etiqueta, produciéndose dichas moléculas de ácido nucleico con etiqueta mediante un procedimiento que comprende:

a) obtener transcritos de ARNm;

b) someter los transcritos de ARNm a una reacción de polimerasa en presencia de cebadores que pueden hibridarse con la cola de poliA de los transcritos de ARNm en condiciones que fomentan la formación de tan sólo un complemento, generando así cadenas derivadas de primer orden;

c) añadir una cola de polinucleótido a las cadenas derivadas de primer orden;

d) someter las cadenas derivadas de primer orden a una reacción de polimerasa en presencia de cebadores que pueden hibridarse con la cola de polinucleótido añadida en la etapa (c) en condiciones que fomentan la formación de tan sólo un complemento, generando así cadenas derivadas de segundo orden,

en el que los cebadores de al menos una de las etapas (b) y (d) comprenden cebadores sustancialmente únicos, generando así moléculas de ácido nucleico con etiqueta.

También se da a conocer un kit para determinar información del número de copias de ácido nucleico no afectada por distorsión por amplificación que comprende:

a) una transferasa terminal, y

b) una pluralidad de cebadores sustancialmente únicos,

en el que los cebadores sustancialmente únicos comprenden etiquetas sustancialmente únicas, y en el que los cebadores sustancialmente únicos pueden hibridarse con la cola de polinucleótido de una molécula de ácido nucleico que tiene una cola de polinucleótido añadida mediante la transferasa terminal.

Breve descripción de las figuras

Figura 1. Amplificación de genoma completo (WGA).

Figura 2. Comparación de cobertura de genoma.

Comparación de cobertura genómica de 7 carriles de secuenciación a partir de una biblioteca de una célula individual frente a 7 carriles de una biblioteca de un millón de células. El apilamiento provoca que las lecturas de la célula individual se concentren sobre las lecturas de la primera derivada a partir de WGA. Más lecturas proporcionan rendimientos decrecientes.

Figura 3. Recuentos de lectura de secuenciación para dinucleótidos CpG dentro de islas de CpG seleccionadas.

Línea celular normal SKN-1 (A, C, E) y línea de cáncer de mama MDA-MB-231 (B, D, F). Naranja indica residuos de T (sin metilar). Azul indica residuos de C (metilados). Los recuentos de la cadena positiva están por encima del eje X y los recuentos en la cadena negativa están por debajo.

Figura 4. Comparación de CGH en matriz con “recuento de secuencias” para analizar datos del número de copias.

El gráfico muestra una porción del cromosoma 17 a partir de una muestra de FFPE, JZ33, que incluye la razón sin procesar (gris) y datos segmentados (verde) a partir de CGH en matriz, y datos de recuento de secuencias sin procesar (naranja) y su segmentación (azul).

Figura 5. CGH de ROMA de sKBR3 comparado con perfiles de 3 aislados de células individuales independientes derivados de secuenciación de ADN escasa.

Figura 6. Secuenciación a través de los puntos de rotura de una deleción homocigota en célula en el cromosoma 5 en la línea T47D.

Figura 7. Diagrama de flujo para determinación de perfil genómico mediante recuento de secuencias de representación.

Figura 8. Esquema para recuento varietal. Se muestran etiquetas varietales como secuencias de nucleótidos al azar (N_{10}).

Diagrama de flujo de una realización de la invención.

Figura 9. Protocolo de ligación por mellas.

La secuencia de cebador A incluye cebador universal A (PrA), etiqueta varietal (B) y oligómero dT+CATG, la secuencia de cebador B incluye cebador universal B (PrB) y oligómero dT+CATG, PrA y PrB son cebadores universales, y cPrA, cPrB y cB son complementarios a PrA, PrB y B, respectivamente.

Figura 10. Esquema para secuenciación dirigida de marcadores regionales.

UfBPrA es cebador directo, B es etiqueta, UrPrR es cebador inverso, PrB es cebador inverso de gen, Uf es cebador directo universal, PrA es cebador directo específico de gen, Ur es cebador inverso universal.

Figura 11. Comparación de recuento varietal con recuento de secuencias.

La figura 11A muestra datos del número de copias en todo el genoma. La figura 11B muestra datos del número de copias de genoma parcial.

Descripción detallada de la invención

Se proporciona un método para obtener, a partir de material genómico, información del número de copias genómicas no afectada por distorsión por amplificación, que comprende:

a) obtener segmentos del material genómico;

b) etiquetar los segmentos con etiquetas de ácido nucleico para generar moléculas de ácido nucleico con etiqueta únicas, de manera que cada una de las moléculas de ácido nucleico con etiqueta únicas comprende un segmento del material genómico de la etapa (a) y una etiqueta;

c) someter las moléculas de ácido nucleico con etiqueta a amplificación mediante reacción en cadena de la polimerasa (PCR);

d) generar lecturas de secuencias asociadas a etiquetas mediante secuenciación del producto de la etapa (c);

e) asignar cada molécula de ácido nucleico con etiqueta a una ubicación en un genoma asociado con el material genómico mediante mapeo de la subsecuencia de cada lectura de secuencia asociada a etiqueta correspondiente a un segmento del material genómico a una ubicación en el genoma; y

f) contar el número de moléculas de ácido nucleico con etiqueta que tienen una etiqueta diferente que se han asignado a la misma ubicación en el genoma,

obteniendo así información del número de copias genómicas no afectada por distorsión por amplificación.

En una realización, el método comprende además estimar un número de copias genómicas de una región del genoma que comprende más de una ubicación en el genoma asignando como número de copias de la región el mayor recuento obtenido en la etapa (f) para las ubicaciones dentro de la región.

En una realización, el método comprende además comparar un recuento obtenido en la etapa (f) para una ubicación en el genoma con un recuento para la misma ubicación obtenido a partir de una muestra de referencia, estimando así un número de copias genómicas relativo de la ubicación.

En una realización, el método comprende además

a) sumar los recuentos obtenidos en la etapa (f) para ubicaciones en el genoma que comprenden una primera región del genoma, en el que la primera región comprende más de una ubicación;

b) sumar los recuentos obtenidos en la etapa (f) para ubicaciones en el genoma que comprenden una segunda región del genoma, en el que la segunda región está compuesta por un número de ubicaciones que es comparable al número de ubicaciones de la primera región;

c) comparar el valor obtenido en la etapa (a) con el valor obtenido en la etapa (b),

estimando así el número de copias genómicas relativo de la primera región del genoma con respecto al número de copias genómicas de la segunda región del genoma.

La etapa (b) de la realización anterior puede comprender además

i) sumar los recuentos obtenidos en la etapa (f) para ubicaciones en el genoma que comprenden una tercera región del genoma, en el que la tercera región está compuesta por un número de ubicaciones que es comparable al número de ubicaciones de la primera región; y

ii) obtener un promedio de la suma de los recuentos obtenidos en la etapa (f) para ubicaciones que comprenden la segunda región y la suma de los recuentos obtenidos en la etapa (f) para ubicaciones que comprenden la tercera región.

En una realización del método, la segunda región del genoma comprende un centrómero.

En una realización, el método comprende además sumar los recuentos obtenidos en la etapa (f) para ubicaciones que comprenden una región del genoma, y comparar la suma con una suma obtenida a partir de una muestra de referencia para la misma región del genoma, estimando así un número de copias genómicas relativo de la región del genoma.

También se da a conocer un método para obtener, a partir de transcritos de ARNm, información del número de copias de ARNm no afectada por distorsión por amplificación, que comprende:

a) obtener ADNc a partir de los transcritos de ARNm;

b) opcionalmente obtener segmentos del ADNc;

c) etiquetar el ADNc o los segmentos de ADNc con etiquetas sustancialmente únicas para generar moléculas de ácido nucleico con etiqueta, de manera que cada molécula de ácido nucleico con etiqueta comprende

i) un ADNc de la etapa (a) o un segmento del ADNc de la etapa (b), y

ii) una etiqueta;

d) someter las moléculas de ácido nucleico con etiqueta a amplificación mediante reacción en cadena de la polimerasa (PCR);

e) generar lecturas de secuencias asociadas a etiquetas mediante secuenciación del producto de la etapa (d);

f) asignar cada molécula de ácido nucleico con etiqueta a una ubicación en una biblioteca de ADNc asociada con los transcritos de ARNm mediante mapeo de la subsecuencia de cada lectura de secuencia asociada a etiqueta correspondiente a un ADNc o un segmento de ADNc a una ubicación en la biblioteca de ADNc; y

g) contar el número de moléculas de ácido nucleico con etiqueta que tienen una etiqueta diferente que se han asignado a la misma ubicación en la biblioteca de ADNc,

obteniendo así información del número de copias de ARNm no afectada por distorsión por amplificación.

También se proporciona un método para obtener, a partir de transcritos de ARNm, información del número de copias de ARNm no afectada por distorsión por amplificación, que comprende:

a) generar moléculas de ácido nucleico con etiqueta, que comprende:

i) someter los transcritos de ARNm a una reacción de polimerasa en condiciones que fomentan la formación de tan sólo un complemento, generando así cadenas derivadas de primer orden;

ii) añadir una cola de polinucleótido a las cadenas derivadas de primer orden; y

iii) someter las cadenas derivadas de primer orden a una reacción de polimerasa en presencia de cebadores que pueden hibridarse con la cola de polinucleótido añadida en la etapa (ii) en condiciones que fomentan la formación de tan sólo un complemento, generando así cadenas derivadas de segundo orden,

en el que los cebadores de al menos una de las etapas (i) y (iii) comprenden cebadores sustancialmente únicos, generando así moléculas de ácido nucleico con etiqueta;

a) someter las moléculas de ácido nucleico con etiqueta a amplificación mediante reacción en cadena de la polimerasa (PCR);

b) generar lecturas de secuencias asociadas a etiquetas mediante secuenciación del producto de la etapa (b);

c) asignar cada molécula de ácido nucleico con etiqueta a una ubicación en una biblioteca de ADNc asociada con los transcritos de ARNm mediante mapeo de la subsecuencia de cada lectura de secuencia asociada a etiqueta correspondiente a un transcrito de ARNm a una ubicación en la biblioteca de ADNc; y

d) contar el número de moléculas de ácido nucleico con etiqueta que tienen una etiqueta diferente que se han asignado a la misma ubicación en la biblioteca de ADNc,

obteniendo así información del número de copias de ARNm no afectada por distorsión por amplificación.

También se da a conocer un método para obtener, a partir de material genómico, información de metilación de ADN no afectada por distorsión por amplificación, que comprende:

a) obtener segmentos del material genómico;

b) añadir una cola de polinucleótido a los extremos de los segmentos del material genómico para generar cadenas derivadas de orden cero;

c) someter las cadenas derivadas de orden cero a una reacción de polimerasa en presencia de cebadores esencialmente únicos, en el que los cebadores esencialmente únicos pueden hibridarse con la cola de polinucleótido de las cadenas de orden cero, en condiciones que fomentan la formación de tan sólo un complemento, generando así moléculas de ácido nucleico con etiqueta;

d) separar las moléculas de ácido nucleico con etiqueta para dar un grupo que consiste en moléculas de ácido nucleico con etiqueta semimetiladas y un grupo que consiste en moléculas de ácido nucleico con etiqueta sin metilar;

e) someter cada grupo de la etapa (d) a amplificación mediante reacción en cadena de la polimerasa (PCR);

f) generar lecturas de secuencias asociadas a etiquetas mediante secuenciación del producto de la etapa (e);

g) asignar cada molécula de ácido nucleico con etiqueta a una ubicación en un genoma asociado con el material genómico mediante mapeo de la subsecuencia de cada lectura de secuencia asociada a etiqueta correspondiente a un segmento del material genómico a una ubicación en el genoma; y

h) contar el número de moléculas de ácido nucleico con etiqueta que tienen una etiqueta diferente que se han asignado a la misma ubicación en el genoma,

obteniendo así información de metilación de ADN no afectada por distorsión por amplificación.

En una realización de los métodos, etiquetar los segmentos para generar moléculas de ácido nucleico con etiqueta comprende:

a) añadir una cola de polinucleótido a los extremos de los segmentos del material genómico para generar cadenas derivadas de orden cero;

b) someter las cadenas derivadas de orden cero de la etapa (a) a una reacción de polimerasa en presencia de cebadores que pueden hibridarse con la cola de polinucleótido de las cadenas de orden cero en condiciones que fomentan la formación de tan sólo un complemento, generando así cadenas derivadas de primer orden;

c) añadir una cola de polinucleótido a las cadenas derivadas de primer orden;

d) someter las cadenas derivadas de primer orden a una reacción de polimerasa en presencia de cebadores que pueden hibridarse con la cola de polinucleótido de las cadenas derivadas de primer orden en condiciones que fomentan la formación de tan sólo un complemento, generando así cadenas derivadas de segundo orden,

en el que los cebadores de al menos una de las etapas (b) y (d) comprenden cebadores sustancialmente únicos, generando así moléculas de ácido nucleico con etiqueta.

En una realización de los métodos, etiquetar los segmentos para generar moléculas de ácido nucleico con etiqueta comprende:

a) añadir una cola de polinucleótido a los extremos de los segmentos de material genómico para generar cadenas derivadas de orden cero,

b) someter las cadenas derivadas de orden cero de la etapa (a) a una reacción de ligación en presencia de cebadores que pueden hibridarse con la cola de polinucleótido de las cadenas de orden cero añadidas en la etapa (a) en condiciones que fomentan la ligación de un cebador a los extremos 5’ de las cadenas derivadas de orden cero,

c) someter el producto de la etapa (b) a una reacción de polimerasa en presencia de cebadores que pueden hibridarse con la cola de polinucleótido añadida en la etapa (a) en condiciones que fomentan la formación de tan sólo un complemento, en el que los cebadores de la etapa (c) tienen secuencias de nucleótidos diferentes de los cebadores de la etapa (b), y en el que la polimerasa tiene actividad de corrección de lectura 3’-5’,

en el que los cebadores de al menos una de las etapas (b) y (c) comprenden cebadores sustancialmente únicos, generando así moléculas de ácido nucleico con etiqueta.

En una realización de los métodos, añadir una cola de polinucleótido comprende el uso de una transferasa terminal.

En una realización de los métodos, etiquetar los segmentos para generar moléculas de ácido nucleico con etiqueta comprende la ligación de adaptadores que comprenden las etiquetas a al menos un extremo de los segmentos del material genómico.

En una realización de los métodos, los adaptadores que comprenden las etiquetas sólo se ligan a un extremo de los segmentos del material genómico.

En una realización de los métodos, las etiquetas comprenden una secuencia que ayuda en la amplificación mediante PCR.

En una realización de los métodos, cada molécula de ácido nucleico con etiqueta comprende una etiqueta.

En una realización de los métodos, cada molécula de ácido nucleico con etiqueta comprende más de una etiqueta.

En una realización de los métodos, se producen segmentos del material genómico mediante digestión con endonucleasa de restricción, cizallamiento mecánico, calentamiento o sonicación.

En una realización de los métodos, se producen segmentos del ADNc mediante digestión con endonucleasa de restricción, cizallamiento mecánico, calentamiento o sonicación.

En una realización de los métodos, el número de copias máximo de una ubicación en una biblioteca de ADNc no es menor que el número de moléculas de ácido nucleico con etiqueta que tienen una etiqueta diferente que se han asignado a la misma ubicación en la biblioteca de ADNc.

Una realización de los métodos anteriores comprende además analizar el número de copias de ARNm.

En un aspecto de la divulgación, separar las moléculas de ácido nucleico con etiqueta para dar un grupo que consiste en moléculas de ácido nucleico con etiqueta semimetiladas y un grupo que consiste en moléculas de ácido

nucleico con etiqueta sin metilar es mediante escisión con enzimas de restricción sensibles a la metilación, reparto con anticuerpos o reparto con proteínas de unión a metil-C dirigidas a citosina metilada o hidroximetilada.

En una realización de los métodos, las moléculas de ácido nucleico con etiqueta se someten a captura de híbridos antes de PCR o antes de la secuenciación.

En una realización de los métodos, cada molécula de ácido nucleico con etiqueta difiere en más de un nucleótido.

En una realización de los métodos, las secuencias de etiqueta comprenden además una etiqueta de muestra.

En una realización de los métodos, las moléculas de ácido nucleico con etiqueta se combinan con una pluralidad de moléculas de ácido nucleico con etiqueta que tienen una etiqueta de muestra diferente antes de la amplificación mediante PCR o antes de la secuenciación.

Una realización de los métodos comprende además realizar la deconvolución de las lecturas de secuencias asociadas a etiquetas agrupando las lecturas de secuencias asociadas a etiquetas según la etiqueta de muestra.

En una realización de los métodos, las moléculas de ácido nucleico con etiqueta se generan a partir de una especie individual.

En una realización de los métodos, las moléculas de ácido nucleico con etiqueta se generan a partir de un organismo individual.

En una realización de los métodos, las moléculas de ácido nucleico con etiqueta se generan a partir de una célula individual.

En una realización, la célula individual procede de un aspirado con aguja de lesiones de cáncer sospechadas.

En una realización, la célula individual procede de una biopsia con aguja gruesa de lesiones de cáncer sospechadas.

En una realización de los métodos, las moléculas de ácido nucleico con etiqueta se generan a partir de dos o más organismos.

En una realización, las moléculas de ácido nucleico con etiqueta se generan a partir de una célula individual de cada organismo.

En una realización de los métodos, la especie individual es ser humano.

En una realización de los métodos, las moléculas de ácido nucleico con etiqueta se generan a partir de dos o más especies.

En una realización, las moléculas de ácido nucleico con etiqueta se generan a partir de una población de microbios.

Una realización adicional comprende comparar la información del número de copias genómicas obtenida para especies diferentes de la población para determinar el recuento relativo de esas especies diferentes en la población.

También se da a conocer una composición de materia derivada de material genómico que comprende moléculas de ácido nucleico con etiqueta, produciéndose dichas moléculas de ácido nucleico con etiqueta mediante un procedimiento que comprende:

a) obtener segmentos del material genómico;

b) añadir una cola de polinucleótido a los extremos de los segmentos del material genómico para generar cadenas derivadas de orden cero;

c) someter las cadenas derivadas de orden cero de la etapa (b) a una reacción de polimerasa en presencia de cebadores que pueden hibridarse con la cola de polinucleótido de las cadenas de orden cero en condiciones que fomentan la formación de tan sólo un complemento, generando así cadenas derivadas de primer orden;

d) añadir una cola de polinucleótido a las cadenas derivadas de primer orden;

e) someter las cadenas derivadas de primer orden a una reacción de polimerasa en presencia de cebadores que pueden hibridarse con la cola de polinucleótido de las cadenas derivadas de primer orden en condiciones que fomentan la formación de tan sólo un complemento, generando así cadenas derivadas de segundo orden,

en el que los cebadores de al menos una de las etapas (c) y (e) comprenden cebadores sustancialmente únicos, generando así moléculas de ácido nucleico con etiqueta.

También se da a conocer una composición de materia derivada de material genómico que comprende moléculas de

ácido nucleico con etiqueta, produciéndose dichas moléculas de ácido nucleico con etiqueta mediante un procedimiento que comprende:

a) obtener segmentos de material genómico;

b) añadir una cola de polinucleótido a los extremos de los segmentos de material genómico para generar cadenas derivadas de orden cero,

c) someter las cadenas derivadas de orden cero de la etapa (b) a una reacción de ligación en presencia de cebadores que pueden hibridarse con la cola de polinucleótido de las cadenas de orden cero añadidas en la etapa (b) en condiciones que fomentan la ligación de un cebador a los extremos 5' de las cadenas derivadas de orden cero,

d) someter el producto de la etapa (c) a una reacción de polimerasa en presencia de cebadores que pueden hibridarse con la cola de polinucleótido añadida en la etapa (b) en condiciones que fomentan la formación de tan sólo un complemento, en el que los cebadores de la etapa (d) tienen secuencias de nucleótidos diferentes de los cebadores de la etapa (c), y en el que la polimerasa tiene actividad de corrección de lectura 3'-5',

en el que los cebadores de al menos una de las etapas (c) y (d) comprenden cebadores sustancialmente únicos, generando así moléculas de ácido nucleico con etiqueta.

También se da a conocer una composición de materia derivada de transcritos de ARNm que comprenden moléculas de ácido nucleico con etiqueta, produciéndose dichas moléculas de ácido nucleico con etiqueta mediante un procedimiento que comprende:

a) obtener ADNc a partir de los transcritos de ARNm;

b) opcionalmente obtener segmentos del ADNc; y

c) etiquetar el ADNc o los segmentos de ADNc con etiquetas sustancialmente únicas para generar moléculas de ácido nucleico con etiqueta, en el que cada molécula de ácido nucleico con etiqueta comprende

i) un ADNc de la etapa (a) o un segmento del ADNc de la etapa (b) y

ii) una etiqueta.

También se da a conocer una composición de materia derivada de transcritos de ARNm que comprenden moléculas de ácido nucleico con etiqueta, produciéndose dichas moléculas de ácido nucleico con etiqueta mediante un procedimiento que comprende:

a) obtener transcritos de ARNm;

b) someter los transcritos de ARNm a una reacción de polimerasa en presencia de cebadores que pueden hibridarse con la cola de poliA de los transcritos de ARNm en condiciones que fomentan la formación de tan sólo un complemento, generando así cadenas derivadas de primer orden;

c) añadir una cola de polinucleótido a las cadenas derivadas de primer orden;

d) someter las cadenas derivadas de primer orden a una reacción de polimerasa en presencia de cebadores que pueden hibridarse con la cola de polinucleótido añadida en la etapa (c) en condiciones que fomentan la formación de tan sólo un complemento, generando así cadenas derivadas de segundo orden,

en el que los cebadores de al menos una de las etapas (b) y (d) comprenden cebadores sustancialmente únicos, generando así moléculas de ácido nucleico con etiqueta.

En una realización de las composiciones anteriores, añadir una cola de polinucleótido comprende el uso de una transferasa terminal.

En una realización de las composiciones anteriores, cada molécula de ácido nucleico con etiqueta comprende una etiqueta.

En una realización de las composiciones anteriores, cada molécula de ácido nucleico con etiqueta comprende más de una etiqueta.

En una realización de las composiciones anteriores, se producen segmentos del material genómico mediante digestión con endonucleasa de restricción, cizallamiento mecánico, calentamiento o sonicación.

En una realización de las composiciones anteriores, se producen segmentos del ADNc mediante digestión con endonucleasa de restricción, cizallamiento mecánico, calentamiento o sonicación.

En una realización de las composiciones anteriores, la complejidad de las moléculas de ácido nucleico con etiqueta se reduce mediante captura de híbridos.

En una realización de las composiciones anteriores, cada molécula de ácido nucleico con etiqueta difiere en más de un nucleótido.

En una realización de las composiciones anteriores, las secuencias de etiqueta comprenden además una etiqueta de muestra.

También se da a conocer una composición de materia que comprende al menos dos combinaciones de moléculas de ácido nucleico con etiqueta tal como se describe en el presente documento.

En una realización de las composiciones anteriores, las moléculas de ácido nucleico con etiqueta se generan a partir de una célula individual.

En una realización, la célula individual procede de un aspirado con aguja de lesiones de cáncer sospechadas.

En una realización, la célula individual procede de una biopsia con aguja gruesa de lesiones de cáncer sospechadas.

En una realización de las composiciones anteriores, las moléculas de ácido nucleico con etiqueta se generan a partir de dos o más organismos.

También se da a conocer un kit para determinar información del número de copias de ácido nucleico no afectada por distorsión por amplificación que comprende:

a) una transferasa terminal, y

b) pluralidad de cebadores sustancialmente únicos,

en el que los cebadores sustancialmente únicos comprenden etiquetas sustancialmente únicas, y en el que los cebadores sustancialmente únicos pueden hibridarse con la cola de polinucleótido de una molécula de ácido nucleico que tiene una cola de polinucleótido añadida mediante la transferasa terminal.

En una realización, el kit comprende además una ADN polimerasa que tiene actividad de corrección de lectura 3'-5'.

En una realización de los kits, la pluralidad de cebadores sustancialmente únicos comprende 10^n cebadores, en los que n es un número entero de desde 2 hasta 9.

En una realización, los kits comprenden además una segunda transferasa terminal y un cebador que puede hibridarse con la cola de polinucleótido de una molécula de ácido nucleico que tiene una cola de polinucleótido añadida mediante la segunda transferasa terminal.

En una realización de los kits, las etiquetas tienen seis nucleótidos de longitud.

En una realización de los kits, las etiquetas tienen 15 nucleótidos de longitud.

En una realización, los kits comprenden además etiquetas de muestra.

En una realización, las etiquetas de muestra tienen 2 nucleótidos de longitud.

En una realización, las etiquetas de muestra tienen 4 nucleótidos de longitud.

En una realización, las etiquetas de muestra comprenden un conjunto de etiquetas de muestra.

En una realización de los kits anteriores, las etiquetas sustancialmente únicas comprenden las etiquetas de muestra.

Términos

Para el fin de esta invención, diferentes términos y frases se definen de la siguiente manera:

Tal como se usa en el presente documento, el término "adaptador" se refiere a un oligonucleótido o segmento o fragmento de ácido nucleico que puede ligarse a una molécula de ácido nucleico de interés. Para los fines de esta invención los adaptadores pueden comprender, como opciones, sitios de unión a cebador, sitios de reconocimiento para endonucleasas, secuencias comunes, promotores, secuencias de etiqueta y secuencias de etiqueta de muestra. Preferiblemente, los adaptadores están colocados para estar ubicados a ambos lados de (flanqueando) una molécula de ácido nucleico de interés particular. Según la invención, pueden añadirse adaptadores a moléculas de ácido nucleico de interés mediante técnicas recombinantes convencionales (por ejemplo, ligación y digestión de restricción). Por ejemplo, pueden añadirse adaptadores a una población de moléculas lineales (por ejemplo un ADN genómico que se ha escindido o digerido) para formar una población de moléculas lineales que contienen

adaptadores en uno y preferiblemente ambos extremos terminales de la totalidad, o una parte sustancial, de moléculas. El adaptador puede ser total o sustancialmente bicatenario o totalmente monocatenario. Un adaptador bicatenario puede comprender dos oligonucleótidos que son al menos parcialmente complementarios. El adaptador puede estar fosforilado o no fosforilado en una o ambas cadenas. Pueden usarse adaptadores para secuenciación de ADN. Los adaptadores también pueden incorporar nucleótidos modificados que modifican las propiedades de la secuencia de adaptador. Por ejemplo, pueden sustituirse citosinas por citosinas metiladas. En una realización de esta invención los adaptadores ligados a ADN genómico para permitir la generación de agrupaciones en el secuenciador contienen citosinas que están todas ellas metiladas. Esta modificación protege a tales adaptadores frente a la conversión con bisulfito, y se tiene en cuenta en las aplicaciones posteriores y el análisis de esta invención.

Tal como se usa en el presente documento, el término “amplificar” se refiere al procedimiento de sintetizar moléculas de ácido nucleico que son complementarias a una o ambas cadenas de un ácido nucleico de molde. Amplificar una molécula de ácido nucleico incluye normalmente desnaturalizar el ácido nucleico de molde, aparear cebadores al ácido nucleico de molde a una temperatura que es inferior a las temperaturas de fusión de los cebadores, y realizar la elongación enzimática a partir de los cebadores para generar un producto de amplificación. Las etapas de desnaturalización, apareamiento y elongación pueden realizarse una vez cada una. Sin embargo, generalmente las etapas de desnaturalización, apareamiento y elongación se realizan múltiples veces (por ejemplo, reacción en cadena de la polimerasa (PCR)) de manera que aumenta la cantidad de producto de amplificación, con frecuencia de manera exponencial, aunque los presentes métodos no requieren la amplificación exponencial. La amplificación requiere normalmente la presencia de trifosfatos de desoxirribonuclósido, una enzima ADN polimerasa y un tampón apropiado y/o cofactores para una actividad óptima de la enzima polimerasa. El término “producto de amplificación” se refiere a las secuencias de ácido nucleico que se producen a partir del procedimiento de amplificación tal como se describe en el presente documento.

Tal como se usa en el presente documento, el término “distorsión por amplificación” se refiere a la amplificación no uniforme de moléculas de ácido nucleico de molde.

Tal como se usa en el presente documento, el término “tratamiento con bisulfito” se refiere al tratamiento de ácido nucleico con un reactivo usado para la conversión con bisulfito de citosina para dar uracilo. Los ejemplos de reactivos de conversión con bisulfito incluyen, pero no se limitan a, tratamiento con un compuesto de bisulfito, un disulfito o un hidrogenosulfito.

Tal como se usa en el presente documento, el término “material convertido con bisulfito” se refiere a un ácido nucleico que se ha puesto en contacto con ion bisulfito en una cantidad apropiada para protocolos de conversión con bisulfito conocidos en la técnica. Por tanto, el término “material convertido con bisulfito” incluye ácidos nucleicos que se han puesto en contacto, por ejemplo, con bisulfito de magnesio o bisulfito de sodio, antes del tratamiento con base.

Tal como se usa en el presente documento, el término “captura de secuencias” o “captura de híbridos” se refiere a un procedimiento de hibridar una “sonda de captura” con un ácido nucleico que tiene una secuencia que es complementaria a la secuencia de la sonda de captura. Puede inmovilizarse una sonda de captura en un sustrato en una micromatriz en fase sólida, en el que “sustrato” se refiere a secuencias de ácido nucleico cortas que se conocen. Su ubicación en la micromatriz en fase sólida puede estar predeterminada o no. La sonda de captura que comprende una “secuencia complementaria al sustrato” puede inmovilizarse en la micromatriz en fase sólida mediante hibridación con su “secuencia de sustrato” complementaria.

Tal como se usa en el presente documento, el término “abrazadera” se refiere a una secuencia de nucleótidos de CC, CG, GC o GG que puede estar ubicada en las dos primeras posiciones de nucleótido en el extremo 5’ de un cebador, etiqueta o etiqueta de muestra, o en las dos últimas posiciones de nucleótido en el extremo 3’ de un cebador, etiqueta o etiqueta de muestra.

Tal como se usa en el presente documento, el término “número de ubicaciones comparable” significa que una primera región genómica tiene un número de ubicaciones en el genoma que está dentro del 25% del número de ubicaciones de una segunda región genómica. En una realización, la primera región genómica que tiene un número de ubicaciones comparable a la segunda región genómica tiene un número de ubicaciones que está dentro del 20%, 15%, 10%, 5%, 1%, o es exactamente el mismo que el número de ubicaciones de la segunda ubicación genómica.

Tal como se usa en el presente documento, una subsecuencia de una lectura de secuencia asociada a etiqueta es “correspondiente a” una etiqueta cuando la subsecuencia es idéntica a la secuencia de nucleótidos de la etiqueta.

Tal como se usa en el presente documento, una subsecuencia de una lectura de secuencia asociada a etiqueta es “correspondiente a” una especie de una molécula de ácido nucleico cuando la subsecuencia es sustancialmente idéntica o sustancialmente complementaria a al menos aproximadamente 10, 12, 14, 16, 18, 20 o más nucleótidos de la secuencia de la especie de molécula de ácido nucleico. En una realización, la subsecuencia es idéntica o totalmente complementaria a al menos aproximadamente 10, 12, 14, 16, 18, 20 o más nucleótidos de la secuencia de la especie de molécula de ácido nucleico.

El término “sitio de CpG” se refiere al dinucleótido CpG.

Tal como se usa en el presente documento, el término “isla de CpG” se refiere a una región de ADN con un alto contenido en G+C y una alta frecuencia de dinucleótidos CpG con respecto al genoma completo, tal como se define adicionalmente en el sitio de UCSC Genome Bioinformatics en genome.ucsc.edu/index.html?org=Human&db=hg19&hgsid=171216665, y en particular, mediante anotación en la base de datos del sitio de USCS Genome Bioinformatics (CpG Islands Track o Table: cpgIslandExt).

Tal como se usa en el presente documento, el término “totalmente complementario” se refiere al complemento inverso de una secuencia de ácido nucleico.

Tal como se usa en el presente documento, el término “biblioteca” se refiere a una colección de moléculas de ácido nucleico (circular o lineal). En una realización preferida, una biblioteca es representativa de todo el contenido de ADN de un organismo (tal biblioteca se denomina biblioteca “genómica”), o un conjunto de moléculas de ácido nucleico representativo de todos los genes expresados (tal biblioteca se denomina biblioteca de ADNc) en una célula, tejido, órgano u organismo. Una biblioteca también puede comprender secuencias al azar realizadas mediante síntesis *de novo*, mutagénesis de una o más secuencias y similares. Una biblioteca puede estar contenida en un vector.

Tal como se usa en el presente documento, el término “mapear” se refiere a identificar una ubicación en un genoma o biblioteca de ADNc que tiene una secuencia que es sustancialmente idéntica o complementaria de manera sustancialmente total a la subsecuencia de una lectura de secuencia asociada a etiqueta correspondiente a una especie de molécula de ácido nucleico, y asignar la lectura de secuencia asociada a etiqueta o la molécula de ácido nucleico con etiqueta que genera la lectura de secuencia asociada a etiqueta a la ubicación. La molécula de ácido nucleico puede ser, pero no se limita a, lo siguiente: un segmento de material genómico, un ADNc, un ARNm o un segmento de un ADNc.

Tal como se usa en el presente documento, el término “metilación” se refiere a la unión covalente de un grupo metilo en la posición C5 de la base de nucleótido citosina dentro de los dinucleótidos CpG de la región genómica de interés. El término “estado de metilación” se refiere a la presencia o ausencia de 5-metil-citosina (“5-Me”) en uno o una pluralidad de dinucleótidos CpG dentro de una secuencia de ADN. Un sitio de metilación es una secuencia de nucleótidos unidos de manera contigua que se reconoce y se metila por una metilasa específica de secuencia. Una metilasa es una enzima que metila (es decir, une de manera covalente a un grupo metilo) uno o más nucleótidos en un sitio de metilación.

Tal como se usa en el presente documento, el término “transcrito de ARNm” se refiere a la molécula de ácido nucleico resultante de la transcripción de ADN.

Tal como se usa en el presente documento, el término “combinar” significa reunir una pluralidad de moléculas de ácido nucleico con al menos otra pluralidad de moléculas de ácido nucleico. En una realización, se combinan moléculas de ácido nucleico con etiqueta. La combinación puede realizarse tras cualquier etapa en la que se generan moléculas de ácido nucleico con etiqueta. En una realización, la combinación se realiza antes de la PCR y/o antes de la secuenciación. En una realización, moléculas de ácido nucleico con etiqueta generadas a partir de una célula individual se combinan con moléculas de ácido nucleico con etiqueta generadas a partir de una célula individual diferente.

Tal como se usa en el presente documento, el término “matriz de sonda” se refiere a una matriz de N moléculas de ácido nucleico diferentes depositadas sobre un sustrato de reacción que sirven para interrogar mezclas de moléculas diana o múltiples sitios en una molécula diana individual suministrada a la superficie de la matriz.

Tal como se usa en el presente documento, el término “lectura” o “lectura de secuencia” se refiere a la información de secuencia de nucleótidos o bases de un ácido nucleico que se ha generado mediante cualquier método de secuenciación. Por tanto, una lectura corresponde a la información de secuencia obtenida a partir de una cadena de un fragmento de ácido nucleico. Por ejemplo, un fragmento de ADN en el que se ha generado una secuencia a partir de una cadena en una única reacción dará como resultado una única lectura. Sin embargo, pueden generarse múltiples lecturas para la misma cadena de ADN en la que existen múltiples copias de ese fragmento de ADN en un proyecto de secuenciación o en la que la cadena se ha secuenciado múltiples veces. Por tanto, una lectura corresponde a las identificaciones de bases de purina o pirimidina o determinaciones de secuencia de una reacción de secuenciación particular.

Tal como se usa en el presente documento, el término “genoma de referencia” se refiere a un genoma de la misma especie que la que está analizándose, genoma para el que se conoce la información de secuencia.

Tal como se usa en el presente documento, el término “región del genoma” se refiere a una secuencia genómica continua que comprende múltiples ubicaciones diferenciadas.

Tal como se usa en el presente documento, el término “representación de ADN” o “representación de un genoma” se refiere a una toma de muestras del ADN o genoma producido mediante una digestión con endonucleasa de

restricción de ADN genómico u otro, seguido por unión de adaptadores y después amplificación con cebadores complementarios a los adaptadores.

Tal como se usa en el presente documento, el término “etiqueta de muestra” se refiere a un ácido nucleico que tiene una secuencia de no más de 1000 nucleótidos y no menos de dos, que puede unirse de manera covalente a cada miembro de una pluralidad de moléculas de ácido nucleico con etiqueta o moléculas de reactivo con etiqueta. Una “etiqueta de muestra” puede comprender parte de una “etiqueta”.

Tal como se usa en el presente documento, el término “conjunto de etiquetas de muestra” se refiere a una pluralidad de etiquetas de muestra únicas de la misma longitud, en el que la secuencia de nucleótidos de cada etiqueta de muestra única difiere de la secuencia de nucleótidos de cualquier otra etiqueta de muestra única en el conjunto de etiquetas de muestra en dos o más posiciones correspondientes en la secuencia. Una etiqueta de muestra única seleccionada del conjunto de etiquetas de muestra puede unirse de manera covalente a cada miembro de una muestra única que consiste en una pluralidad de moléculas de ácido nucleico con etiqueta con el fin de identificar una molécula de ácido nucleico con etiqueta como miembro de esa muestra única.

Tal como se usa en el presente documento, el término “segmentos de ADNc”, se refiere a las moléculas de ácido nucleico resultantes de la fragmentación de ADNc.

Tal como se usa en el presente documento, el término “segmentos de material genómico” se refiere a las moléculas de ácido nucleico resultantes de la fragmentación de ADN genómico.

Una molécula de ácido nucleico que contiene una secuencia de nucleótidos idéntica a un segmento de un ADNc o un segmento de material genómico comprende un “segmento de ADNc” o un “segmento de material genómico”, respectivamente.

Tal como se usa en el presente documento, la “complejidad de secuencia” o “complejidad” con respecto a una población de polinucleótidos se refiere al número de especies diferentes de polinucleótidos presentes en la población.

Tal como se usa en el presente documento, las secuencias “sustancialmente idénticas” o “complementarias de manera sustancialmente total” tienen una identidad de secuencia o complementariedad de al menos aproximadamente el 80%, respectivamente, con respecto a una secuencia de nucleótidos. Las secuencias sustancialmente idénticas o sustancialmente complementarias de manera sustancialmente total pueden tener una identidad de secuencia o complementariedad de al menos aproximadamente el 85%, 90%, 95% o 100%, respectivamente.

Tal como se usa en el presente documento, el término “cebadores sustancialmente únicos” se refiere a una pluralidad de cebadores, en la que cada cebador comprende una etiqueta, y en la que al menos el 50% de las etiquetas de la pluralidad de cebadores son únicas. Preferiblemente, las etiquetas son etiquetas únicas en al menos el 60%, 70%, 80%, 90% o 100%.

Tal como se usa en el presente documento, el término “etiquetas sustancialmente únicas” se refiere a etiquetas en una pluralidad de etiquetas, en la que al menos el 50% de las etiquetas de la pluralidad son únicas con respecto a la pluralidad de etiquetas. Preferiblemente, las etiquetas son etiquetas únicas en al menos el 60%, 70%, 80%, 90% o 100%.

Tal como se usa en el presente documento, el término “etiqueta” se refiere a un ácido nucleico que tiene una secuencia de no más de 1000 nucleótidos y no menos de dos que puede unirse de manera covalente a una molécula de ácido nucleico o molécula de reactivo. Una etiqueta puede comprender una parte de un adaptador o un cebador.

Tal como se usa en el presente documento, una “molécula de ácido nucleico con etiqueta” se refiere a una molécula de ácido nucleico que está unida de manera covalente a una “etiqueta”.

Cuando se proporciona un intervalo de valores, se entiende que cada valor intermedio, hasta una 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 valor mencionado o intermedio en ese intervalo mencionado, queda abarcado dentro de la invención. Los límites superior e inferior de estos intervalos menores pueden incluirse independientemente en los intervalos menores, y también quedan abarcados dentro de la invención, sujetos a cualquier límite específicamente excluido en el intervalo mencionado. Cuando el intervalo mencionado incluye uno o ambos de los límites, los intervalos que excluyen cualquiera o ambos de esos límites incluidos también quedan incluidos en la invención.

Todas las combinaciones de las realizaciones de la invención descritas en el presente documento están dentro del alcance de la invención a menos que se indique claramente lo contrario.

Determinación del perfil del número de copias de células individuales mediante “recuento de secuencias” y

etiquetado de muestras.

Los secuenciadores de alto rendimiento son en realidad secuenciadores de moléculas individuales. Dado que el genoma humano se ha secuenciado y ensamblado, un secuenciador de moléculas individuales también es un identificador de moléculas individuales. Por tanto, la secuenciación también es el recuento, tal como se ilustra en la figura 4 que ilustra cómo la secuenciación de baja cobertura puede proporcionar una estimación del número de copias que es comparable a una hibridación en matriz. La distinción para esto es que el "recuento de secuencias" puede realizarse con células individuales y, usando el sistema de multiplex dado a conocer en el presente documento, de manera relativamente económica, convirtiendo de hecho el secuenciador de alto rendimiento en un detector de células individuales con genomas anómalos. Se añadieron dos características a la tecnología de secuenciación: dilución limitante y etiquetado de muestras. Para la secuenciación de células individuales, en primer lugar se depositan las células individuales o los núcleos de las células individuales en pocillos independientes o se mantienen físicamente separadas mediante otros medios, es decir un gel. Las células individuales o los núcleos de células pueden depositarse en un pocillo mediante un clasificador celular. Alternativamente puede diluirse una colección de células o núcleos en un fluido que después se vuelve semisólido. Se prepara una etiqueta de muestra única para cada pocillo para la secuenciación, con una etiqueta de muestra independiente por pocillo. Después de esta etapa, se combinan los pocillos y se secuencian. La etiqueta de muestra se usa para realizar la deconvolución de la información, permitiendo realizar un perfil del número de copias para cada célula individual.

Protocolo: Aislar células tumorales individuales o núcleos individuales de la célula tumoral, etiquetar y opcionalmente etiquetar con etiqueta de muestra el ADN de cada célula o núcleo, y amplificar el ADN con etiqueta hasta cantidades de microgramos para generar bibliotecas con etiqueta para la secuenciación de tipo Solexa. El aislamiento de células tumorales individuales se producirá mediante clasificación celular activada por fluorescencia mediante contenido de ADN genómico total, para depositar núcleos individuales en pocillos individuales.

Alternativamente, pueden aislarse células tumorales mediante clasificación FACS con anticuerpos fluorescentes contra marcadores de epitelio de mama, incluyendo citoqueratina 8, 18 ó 19. Se someten las células a lisis *in situ* y se fragmenta el genoma mediante calentamiento. Se construyen bibliotecas a partir del ADN genómico fragmentado. Se realizaron experimentos preliminares usando micromatriz ROMA de alta resolución para cuantificar el número de copias en todo el genoma con ADN amplificado usando WGA a partir de células SK-BR-3 individuales (figura 5). Estos experimentos demuestran la viabilidad y eficacia de amplificar 6 pg de ADN hasta cantidades de microgramos para determinar el número de copias mediante métodos de micromatriz, lo que sugiere que el ADN también es adecuado para determinar el número de copias mediante secuenciación.

Dado que la secuenciación de genomas tumorales completos a partir de células individuales es costosa y requiere mucho tiempo, se proponen dos métodos para estratificar el genoma y secuenciar para determinar el número de copias: secuenciación de representación y secuenciación por grupos.

Secuenciación de representación

La secuenciación de representación implica digerir por restricción de ADN genómico para dar fragmentos y secuenciar sólo la población estratificada de fragmentos de restricción para determinar el número de copias. Este método tiene la ventaja de generar fragmentos de restricción precisos que pueden prepararse de manera fiable como bibliotecas a partir de diferentes células tumorales individuales. En este protocolo se amplifica ADN genómico a partir de una célula tumoral individual, se digiere con DpnII, se realiza una segunda digestión con una enzima AluI para retirar "elementos SINE Alu" altamente repetitivos, se ligan adaptadores de secuenciación (Illumina GA, Solexa) con una etiqueta de muestra de 4 nucleótidos única y se amplifican por PCR los fragmentos ligados para generar una biblioteca de secuenciación. El modelado *in silico* predice que las digestiones dobles aumentarán la fracción de lecturas de secuenciación únicas hasta el 75% desde el 55% observado normalmente en la secuenciación de genoma humano sin seleccionar. Después se combinan las bibliotecas de ADN con etiqueta de muestra y se aplican a las celdas de flujo de Solexa para la amplificación de agrupaciones y secuenciación de lecturas individuales. Se usan lecturas de extremos individuales porque son más rápidas y económicas que las lecturas de extremos emparejados y proporcionan toda la información necesaria para el "recuento de secuencias". Después, algoritmos de mapeo alinean cada lectura de secuencia con una puntuación de alta calidad (con una longitud de lectura superior a 35 pb) y una etiqueta de muestra inequívoca a una ubicación específica en el mapa de restricción de DpnII del genoma humano. El resultado es un recuento de lecturas en lugares en el genoma, de manera similar al procedimiento demostrado en la figura 4, excepto porque los datos se segregan por etiqueta de muestra, de modo que cada pocillo recibe un perfil de número de copias.

Secuenciación por grupos

Queda claro que la secuenciación de fragmentos seleccionados por tamaño al azar a partir del genoma es una alternativa viable. En este método, se amplifica el ADN de células tumorales individuales para generar fragmentos de 200-2000 pb. Después se usa la clonación de extremos romos para ligar adaptadores a estos fragmentos, seguida por amplificación mediante PCR para generar bibliotecas (protocolo de Solexa convencional para ADN genómico de lectura individual). Sin embargo, además de este protocolo, se ligan secuencias de adaptador con etiquetas de muestra de 4 nucleótidos para generar bibliotecas únicas a partir de cada genoma de célula tumoral.

Después se combinan las bibliotecas de ADN con etiqueta de muestra, y se aplican a las celdas de flujo de Solexa para la amplificación de agrupaciones y secuenciación de lecturas individuales. Después se mapean las secuencias de 35 pb a regiones específicas del genoma humano y se someten a deconvolución usando la etiqueta de muestra única para determinar qué lectura de secuencia se originó a partir de qué célula tumoral. La etapa crucial en este procedimiento que es diferente de los métodos de representación es dividir artificialmente el genoma humano en "grupos" de varias kilobases. La descomposición del análisis en datos de recuentos de secuencias tiene otro grado de libertad en comparación con lo que se pone en práctica para datos de matriz, concretamente que pueden ajustarse los límites de los grupos para minimizar la varianza. Cada grupo contiene un número de lecturas de secuencias genómicas al azar a partir del cual se calcula un valor medio para determinar el número de copias genómicas con una baja cobertura de secuencia.

El uso particular del método de WGA no depende de la amplificación absolutamente uniforme a lo largo del genoma. Pueden obtenerse datos útiles aunque sólo se amplifique la mitad de los posibles fragmentos a partir de una célula individual en una reacción dada. Adicionalmente, la redundancia de datos obtenida a partir de la segmentación es beneficiosa en la identificación de puntos de rotura.

Secuenciación de representación frente a secuenciación por grupos

Una comparación de los dos métodos revela el número de muestras cuyo perfil puede determinarse por serie de secuenciación. De los 3.149.324 fragmentos de Dpn II con un intervalo de tamaño de 200-800 (tamaño óptimo para la secuenciación) hay 618.629 que no se cortan con AluI y por tanto están disponibles para amplificarse. Aproximadamente el 75% de los extremos de fragmento, o aproximadamente 450.000, proporcionan mapeos únicos y por tanto proporcionan ubicaciones para el recuento. El procedimiento global se presenta como diagrama de flujo en la figura 7.

Suponiendo 15 millones de lecturas de calidad por carril de secuenciación de Solexa, la secuenciación al azar proporciona aproximadamente el 55% u 8,25 millones de lecturas únicas por carril. La representación de DpnII/AluI debe aumentar el número de lecturas únicas hasta 11,25 millones. Modelando un carril de celda de flujo individual con grupos de 100 kb distribuidos a lo largo del genoma, los resultados previstos son los siguientes:

Tabla 1

N.º de muestras	lecturas prom./grupo/muestra (al azar)	lecturas prom./grupo/muestra (Dpn/Alu)
1 muestra por carril	267	363
2 muestras por carril	133	181
12 muestras por carril	22	30

El número de muestras (células individuales) que van a secuenciarse por carril es de entre 4 y diez. Con 8 carriles en una celda de flujo, puede determinarse el perfil de 30-80 muestras al nivel de la figura 4 de manera relativamente económica.

Finalmente, se determinará el número de copias genómicas completo (generado mediante secuenciación) para múltiples células tumorales individuales en un tumor sólido individual. Se segmentarán los datos para dar secciones que minimizan la varianza del número de copias, se someterán a pruebas de significación de Kolmogorov-Smirnov, para generar perfiles del número de copias de tumores individuales. Tal como se mencionó anteriormente, un grado de libertad adicional, que aún no incorporan los presentes algoritmos, es seleccionar el "límite" de los grupos. Los perfiles de genoma tumoral de células individuales se compararán para estudiar la heterogeneidad tumoral en tumores sólidos individuales, lo cual puede esclarecer la progresión de acontecimientos genéticos durante la tumorigénesis.

Recuento varietal de ácidos nucleicos

La medición de la concentración absoluta o relativa de secuencias de ácido nucleico en una muestra con frecuencia presenta distorsiones estocásticas graves debido a amplificaciones o pérdidas diferenciales durante el procesamiento y la manipulación de los ácidos nucleicos. Nunca hay más información disponible que el número de moléculas del material de partida en la propia muestra. El método descrito en el presente documento está diseñado para captar la información en la fase más temprana, de una forma que entonces es resistente a la distorsión por amplificación adicional.

Las moléculas de ácido nucleico de entrada (derivadas de orden 0), o copias de primera ronda de las mismas (derivadas de 1er orden), presentes en la muestra pueden etiquetarse al azar usando una gran elección de etiquetas distintas. Cada molécula con etiqueta se vuelve esencialmente única, la combinación de la información en la etiqueta y la información en el ácido nucleico, que posteriormente puede leerse mediante secuenciación. La combinación de una etiqueta con una molécula de ácido nucleico de orden 0, 1er orden o de orden N posterior se denomina "molécula de ácido nucleico con etiqueta". Tras la amplificación o selección u otro procesamiento, se cuenta el número de etiquetas diferentes.

Dentro de una combinación de moléculas de ácido nucleico con etiqueta, cada molécula de ácido nucleico con

etiqueta es probablemente única en la combinación cuando se usa un número suficientemente grande de etiquetas distintas. Después, en amplificaciones posteriores, se amplifican estas moléculas de ácido nucleico con etiqueta únicas para facilitar la detección, pero no se crea ninguna nueva especie de moléculas de ácido nucleico con etiqueta. Por tanto, no se crean recuentos distorsionados siempre que se cuente el número de moléculas de ácido nucleico con etiqueta diferentes.

En el método dado a conocer, cada lectura de secuencia generada a partir de una molécula de ácido nucleico con etiqueta comprende dos partes o subsecuencias. La primera parte corresponde a la etiqueta, que identifica la etiqueta (y la etiqueta de muestra si se combinan muestras). La segunda parte corresponde a un ácido nucleico en la muestra que se mapea a una ubicación en el genoma (o a un transcrito si están contándose moléculas de ARN). Tras la deconvolución mediante etiqueta de muestra, se cuenta el número de moléculas de ácido nucleico con etiqueta diferentes mapeadas a cada ubicación. Hay dos métodos de recuento independientes, y cada uno puede combinarse para obtener una mayor precisión y determinación de confianza.

El primer método de recuento se denomina “número de moléculas de ácido nucleico con etiqueta máximo”. El número de moléculas de ácido nucleico con etiqueta diferentes en una ubicación dada no puede superar el número de copias verdadero (absoluto). Por tanto, para una región dada que comprende múltiples ubicaciones, el número de copias máximo de la región no es menor que el número de moléculas de ácido nucleico con etiqueta diferentes máximo mapeado a cualquier ubicación en esa región. El número de moléculas de ácido nucleico con etiqueta diferentes máximo centrado alrededor de una ventana móvil de un número de ubicaciones fijado puede tomarse como la medida del número de copias verdadero de la ventana. Esto proporcionará en el peor de los casos una subestimación del número de copias verdadero máximo para esa ventana, y nunca una sobreestimación. El conjunto de número de moléculas de ácido nucleico con etiqueta diferentes máximo es un tipo de perfil del número de copias.

El método de recuento descrito anteriormente es lo más preciso cuando la eficacia de procesamiento es excelente y el número de moléculas mapeadas a cada ubicación es menor que el número de etiquetas. En tales condiciones prácticamente todas las moléculas en la muestra se etiquetan y recuentan, y puede derivarse el recuento verdadero. Pero hay un segundo método.

El segundo método de recuento se denomina “moléculas de ácido nucleico con etiqueta diferentes totales”, y es útil, por ejemplo, cuando la eficacia de etiquetado es baja. El número de moléculas de ácido nucleico con etiqueta diferentes total dentro de una región de un número de ubicaciones fijado será, dentro de un error de recuento estadístico, una función monotónica del número de copias verdadero, independientemente de la eficacia del etiquetado. La eficacia baja del etiquetado simplemente aumenta el error estadístico. El número de copias relativo de una región puede estimarse comparando el número de moléculas de ácido nucleico con etiqueta diferentes total mapeadas a la región, por ejemplo, con el valor medio de moléculas de ácido nucleico con etiqueta mapeadas a otras regiones del genoma que tienen un número de ubicaciones comparable.

Si el número de moléculas en una ubicación en un genoma o biblioteca de ADNc supera el número de etiquetas, y el etiquetado es demasiado eficaz, no puede recuperarse información del número de copias. En vez de eso, simplemente se medirá el número de etiquetas total en cada ubicación. Hay al menos tres maneras de solucionar esta condición: 1) disminuir la eficacia de la primera etapa del procedimiento, por ejemplo reduciendo el tiempo de reacción; 2) reducir la cantidad de muestra; y 3) aumentar el número de etiquetas total. Por ejemplo, si la longitud de la etiqueta es de N nucleótidos, hay 4^N etiquetas posibles. Para N = 15, hay aproximadamente 10^9 etiquetas, indudablemente superior al número de moléculas por ubicación.

La colección de lecturas se presta a modelar la potencia de deducción. Para cada ubicación, existe el número de moléculas de ácido nucleico con etiqueta observado, y para cada especie de molécula de ácido nucleico con etiqueta, el número de veces que se observó. Con una buena estimación de la entrada, el número de genomas, pueden deducirse con precisión parámetros críticos porque la gran mayoría de los números de copias son de dos por genoma. El caso limitante de análisis de células individuales se presta al modelado más preciso, con la proporción de lecturas nulas que permiten la deducción de la probabilidad, theta, de que se detecte una molécula individual en una ubicación. Theta es una función de la eficacia de etiquetado, la eficacia de amplificación y procesamiento, y la profundidad de lectura.

Usando el lenguaje inventado de la reacción de WGA, en el protocolo específico descrito en este caso, se etiquetan todas las moléculas de ADN derivadas de primer orden (es decir, primeras copias), sin nuevas etiquetas introducidas cuando se crean derivadas de orden posterior. Esto permite el “recuento” de moléculas originales, alelos en el caso del análisis de genoma. Este método se denomina en el presente documento “recuento varietal”. Siempre que la profundidad de secuencia sea suficiente, la amplificación posterior o métodos de atrapamiento que enriquecen regiones de interés, o la combinación, no pueden distorsionar el recuento. La combinación a partir de múltiples células pasa a ser una extensión natural de este método, que mejora la eficacia, aumenta el rendimiento y reduce los costes.

Etiquetas

Las etiquetas son secuencias de nucleótidos con porciones constantes y porciones variables. La longitud de la

porción constante y la longitud de la porción variable pueden ser iguales o diferentes. La longitud de las etiquetas dentro de una pluralidad de etiquetas es normalmente, pero no necesariamente, igual para todas las etiquetas. En una realización, la longitud total de una etiqueta es de menos de 100 nucleótidos.

Las porciones constantes se usan para ayudar en el etiquetado de moléculas de ácido nucleico. Por ejemplo, la porción constante puede contener una secuencia de nucleótidos que permiten la hibridación con una cola de poli-N añadida mediante una transferasa terminal. La porción constante también puede usarse para ayudar en la manipulación de las moléculas de ácido nucleico con etiqueta. Por ejemplo, la porción constante puede contener una secuencia que es un sitio de unión a cebador útil para la amplificación mediante PCR. La porción constante también puede usarse para identificar una muestra si hay múltiples muestras que están procesándose en paralelo, es decir, la porción constante puede contener una etiqueta de muestra.

No se necesita conocer la secuencia de etiquetas individuales antes de generar moléculas de ácido nucleico con etiqueta. Una pluralidad de etiquetas pueden tener porciones variables cuyas secuencias se esclarecen por primera vez durante la secuenciación. La porción variable de una etiqueta puede ser una secuencia de nucleótidos en la que cada nucleótido es individualmente uno de A, T, C o G, o uno de dos cualesquiera de A, T, C y G, o uno de tres cualesquiera de A, T, C y G, es decir que la porción variable de una etiqueta puede consistir únicamente en dos o tres especies de nucleótidos. Pueden diseñarse etiquetas de manera que la porción variable tiene una secuencia al azar. Pueden diseñarse etiquetas de manera que las porciones variables se construyen a partir de un conjunto de dinucleótidos o trinucleótidos.

Pueden asociarse etiquetas con moléculas de ácido nucleico usando, de manera individual o en combinación, ligación, ligación por mellas, transferasa terminal, hibridación, cebado y similares. Por ejemplo, pueden cortarse moléculas de ácido nucleico con una(s) endonucleasa(s) de restricción, tratarse con una transferasa terminal para añadir una cola de poli-N, hibridarse con una molécula de etiqueta, y después someterse a ligación por mellas para formar una molécula de ácido nucleico con etiqueta. A continuación se comentan en detalle métodos específicos de generación de moléculas de ácido nucleico con etiqueta.

Etiquetado con ayuda de una transferasa terminal

La figura 8 ilustra una realización de este método, en la que N_{10} designa un oligonucleótido que tiene un componente de etiqueta y un componente de etiqueta de muestra. Los seis primeros nucleótidos (4096 secuencias posibles), generados al azar, proporcionan el componente de etiqueta, y los cuatro últimos nucleótidos (256 elecciones) se eligen para codificar para el micropocillo y, por tanto, una etiqueta de muestra. Las longitudes de la etiqueta y la etiqueta de muestra pueden cambiarse para adaptarse a las necesidades, y no se limitan a un total de diez nucleótidos. Las longitudes se eligen para ilustración. La etiqueta de muestra permite la combinación, produciéndose posteriormente la deconvolución. La combinación de una etiqueta con una molécula de ácido nucleico a partir de la muestra de entrada proporciona una molécula de ácido nucleico con etiqueta esencialmente única para la cadena de primera derivada, que sólo puede replicarse con muy baja probabilidad por casualidad a partir de otra molécula similar (una probabilidad de aproximadamente 1/N donde N es el número de etiquetas disponibles). Al final, se cuentan individualmente cada una de múltiples copias de moléculas de secuencia idéntica en la muestra de entrada. En aplicaciones para el recuento de ARNm, la longitud del componente de etiqueta puede aumentarse para adaptarse posiblemente a miles de transcritos a partir del mismo gen.

Aunque hay múltiples realizaciones del método básico además de la ilustrada en la figura 8, la realización ilustrada en la figura 8 tiene varias características de diseño útiles. En primer lugar, evita el uso de ligasas (un método alternativo para asociar una etiqueta con una molécula de ácido nucleico), lo que se ha encontrado que es una reacción ineficaz. En vez de eso, esta realización usa transferasa terminal (TTasa), que es una enzima robusta a partir de la cual se espera una alta eficacia. En segundo lugar, una vez incorporadas la etiqueta y etiqueta de muestra en la cadena de ADN de primera derivada, pueden combinarse micropocillos y llevarse a cabo las etapas posteriores de manera más eficaz en volúmenes mayores con números de moléculas mayores, posiblemente con la adición de ácidos nucleicos transportadores. Debido a una etapa de limpieza, que elimina adaptador libre, no pueden crearse más cadenas derivadas de primer orden únicas con etiquetas, y pueden expandirse de manera aritmética cadenas de segundo orden si se encuentra que resulta útil. En tercer lugar, el método se extiende inmediatamente de ADN a ARNm. En vez de escindir y añadir una cola de polinucleótido con transferasa terminal, puede usarse la cola de poliA que se produce de manera natural de ARNm. En cuarto lugar, resulta fácil ver cómo combinar el análisis tanto de ARN como de ADN a partir de la misma célula con este esquema. En vez de aplicar colas de A a ADN, pueden aplicarse colas, por ejemplo, de C, distinguiendo por tanto el ADN del ARN. Los derivados de ADN y ARN pueden o bien leerse juntos en las mismas series de secuenciación, o bien por separado mediante amplificación usando sus adaptadores de cebadores de PCR diferentes. En quinto lugar, usando una etiqueta de muestra más larga, pueden amplificarse de manera preferible secuencias que pertenecen a una célula particular para un estudio más en profundidad cuando esto quede justificado.

Este método puede aplicarse a la evaluación de la metilación del ADN. Tras preparar las cadenas derivadas de primer orden, puede combinarse el ADN de tipo dúplex a partir de múltiples muestras. Estas moléculas se someterán a semimetilación y todavía conservarán marcas epigenéticas. Por tanto, después pueden separarse en muestras metiladas y sin metilar o bien mediante escisión con enzimas de restricción sensibles a la metilación o bien

mediante reparto con anticuerpos o proteínas de unión a metil-C dirigidos a citosina metilada (o hidroximetilada). Las tasas de error se medirán basándose en expectativas a partir de regiones con un estado de metilación conocido.

Etiquetado mediante ligación

En otra realización de la invención, pueden ligarse adaptadores con etiqueta(s) a ADN de muestra o ARN que se ha convertido en ADN. En la segunda etapa pueden combinarse estas moléculas (o no) y amplificarse para su secuenciación. En la tercera etapa se compilan los números de cada molécula de ácido nucleico con etiqueta diferente de cada locus (o lectura de secuencia que puede mapearse), proporcionando datos estadísticos a partir de los cuales puede deducirse el número de copias.

En otra realización, para medir ARN, en primer lugar se convierte ARN en ADN usando métodos que producen una molécula bicatenaria individual para cada ARN en la muestra. (Por ejemplo, cebado con oligómero de dT y transcriptasa inversa, seguido por cebado al azar y polimerasa de Klenow). Entonces pueden seguirse los procedimientos para analizar ADN.

En otra realización, el ADN de muestra puede someterse a cizallamiento y prepararse los extremos para la ligación. Sin embargo, hay tres ventajas principales de escindir con una o más endonucleasas de restricción (RE). La primera es que el número esperado de moléculas de ácido nucleico con etiqueta en cada ubicación puede predecirse con precisión a partir del número de células de partida en la muestra, o a la inversa el número de células en la muestra puede deducirse a partir del número esperado de moléculas de ácido nucleico con etiqueta por ubicación. La segunda es que se reduce la complejidad esperada del producto, permitiendo la determinación del número de copias a partir de menos lecturas de secuencias. La tercera son determinadas ventajas en cuanto a la flexibilidad y funcionalidad que pueden incorporarse en las etiquetas.

En una realización, se escinde muestra ácido nucleico con una enzima de restricción específica y se ligan moléculas de muestra a adaptadores de oligonucleótidos con etiqueta (y con etiqueta de muestra si es apropiado) que contienen secuencias que permiten la amplificación mediante PCR. Hay diversos diseños para estos adaptadores, y también alternativas a la PCR (tales como WGA), basándose en las funcionalidades incorporadas en las etiquetas. A continuación se comentan algunas de estas funcionalidades adicionales.

Ahora las moléculas ligadas son esencialmente únicas, siendo cada una de las moléculas ligadas la combinación de una secuencia de ADN y etiqueta sustancialmente única a la que se ligaron. Una amplificación adicional puede distorsionar los rendimientos de moléculas, pero no creará nuevas etiquetas (aunque algunas pueden perderse mediante procedimientos estadísticos que se modelan fácilmente durante el análisis).

Mediante error de secuencia durante el procesamiento, muy ocasionalmente la etiqueta puede mutar, creando la aparición de una nueva molécula de ácido nucleico con etiqueta. Pero si el conjunto de etiquetas es suficientemente grande, las moléculas de ácido nucleico con etiqueta diferirán generalmente en más de un nucleótido, y por tanto las moléculas de ácido nucleico con etiqueta que difieren en un único nucleótido pueden (opcionalmente) ignorarse cuando se realiza el recuento.

Para células individuales, se cargan células individuales a como mucho una por micropocillo. Después se prepara el ácido nucleico en pocillos y se etiquetan las moléculas. Una vez etiquetadas, pueden combinarse para su procesamiento adicional, lo cual mejora en gran medida la uniformidad y eficacia de métodos de amplificación de ADN.

En aún otra realización, se corta el ADN y se ligan adaptadores a ambos extremos de cada fragmento, permitiendo que puedan amplificarse mediante PCR. Para evitar la autoligación adaptador-adaptador, pueden usarse adaptadores bicatenarios sin fosforilación en 5', de modo que un adaptador sólo se liga en una cadena en cada extremo del ADN de muestra escindido. Después se combinan las moléculas y se tratan como para las representaciones: el adaptador monocatenario más corto se funde a una temperatura elevada, pero muy por debajo de la Tf para el dúplex de muestra, se rellenan los extremos del ADN bicatenario con polimerasa de Klenow, se retiran los adaptadores sin ligar y después se amplifica mediante PCR la muestra.

En la realización anterior, se usan adaptadores no sólo para el etiquetado sino también para la amplificación. Los adaptadores son para identificar cada molécula de manera individual. También pueden usarse para la amplificación, tal como se describió anteriormente, o para otros usos (véase a continuación). Dado que en la realización anterior se requiere para la amplificación la ligación de adaptador en ambos extremos de fragmentos de muestra, el rendimiento puede reducirse. Para mejorar el rendimiento, el método puede funcionar con moléculas que tienen un adaptador sólo en un extremo.

En una realización, se realizan las reacciones como anteriormente, pero las etiquetas no contienen secuencias que permiten PCR. Antes de la PCR, pero después de una combinación opcional, se ligan adaptadores nuevos (composición diferente y sin etiquetas) de modo que pueden amplificarse y eventualmente contarse moléculas con tan sólo un único adaptador de etiqueta. Si la etiqueta original tiene una secuencia de selección, opcionalmente pueden enriquecerse mediante selección de híbridos moléculas con el adaptador de etiqueta antes de la secuenciación.

En otra realización, se diseñan adaptadores iniciales con un extremo 5' fosforilado y un extremo 3'-didesoxilo, de modo que la ligación inicial se produce con la cadena opuesta tal como se describió en la realización anterior. (En vez de usar una RE con una proyección en 5', puede preferirse una que deja una proyección en 3'). Tras la retirada de adaptadores sin ligar, siguen varias rondas de amplificación aritmética, seguido por la aplicación a moléculas de colas de dT usando transferasa terminal, y amplificación del ADN de muestra procesado mediante PCR.

En otra realización, se marca ADN a partir de muestra como anteriormente, se combina si se requiere y se retiran adaptadores sin ligar de la reacción. Se somete la reacción a amplificación de genoma completo (WGA) y se usa el producto de WGA para preparar bibliotecas de secuenciación. Para esta realización, los adaptadores iniciales también pueden contener: (A) sitios de escisión para facilidad de preparación de bibliotecas de secuenciación; y (B) una secuencia lo suficientemente grande como para enriquecimiento opcional mediante selección de híbridos antes de preparar bibliotecas para secuenciación.

En las realizaciones que acaban de describirse, se supone que la etiqueta es bicatenaria. La producción de una gran variedad de moléculas bicatenarias no es completamente directa. En una realización, se sintetiza ADN monocatenario como de tipo I y de tipo II. El tipo II tiene tres conjuntos de funcionalidades A, B, C y D. Las funcionalidades B contienen la secuencia de etiqueta, C los extremos adhesivos con un sitio de RE, y D el tampón. B contendrá alguna longitud de series con secuencia al azar, por ejemplo 10 nucleótidos con una complejidad total de 4^{10}, o aproximadamente 10^6 secuencias posibles. El oligómero de tipo I será complementario al tipo II en la funcionalidad A, se unen y aparean los dos, y después se extienden con polimerasa para crear una molécula bicatenaria. Tras la escisión con el sitio de RE, la molécula bicatenaria resultante es la etiqueta usada para el posterior etiquetado de la muestra.

En la técnica se conocen métodos adicionales para producir oligonucleótidos mono y bicatenarios adecuados para su uso como etiquetas, por ejemplo, en las patentes estadounidenses n.^{os} 5.639.603 y 7.622.281, publicación de solicitud de patente estadounidense n.º 2006/0073506 y Parameswaran *et al*. (2007) (7, 5, 8, 4).

Ligación por mellas

En esta realización de la invención, en primer lugar se extienden extremos 3'OH de moléculas de ácido nucleico, por ejemplo, segmentos de material genómico o transcritos de ARNm, con transferasa terminal, añadiendo una cola de poli-N. Después se aparean adaptadores de cebadores a las colas de poli-N y se sella con ADN ligasa el dúplex resultante con una mella. Esta serie de reacciones es mucho más eficaz que una ligación bicatenaria en una única etapa. El producto resultante se amplifica mediante PCR. Este método se ha usado de manera satisfactoria con tan sólo 1 ng de ADN. Pueden generarse moléculas de ácido nucleico con etiqueta generadas usando el método de ligación por mellas a partir de ADN de una célula individual, y después combinarse antes de la PCR. La combinación a partir de 100 células proporciona la concentración de ADN necesaria para una amplificación satisfactoria. Pueden añadirse etiquetas de muestra en la etapa de ligación por mellas y esto se logra fácilmente (véase la figura 9).

Tecnología para el ensayo económico de marcadores regionales

La descripción anterior es de una metodología diseñada para medir el número de copias y las cargas de mutación a lo largo de gran parte del genoma en células individuales. Pero la aplicación de descubrimientos, una vez realizados, de marcadores predictivos para el desenlace o la respuesta terapéutica no requerirá información del genoma completo. De hecho, los presentes marcadores de desenlaces son marcadores de número de copias en loci muy específicos. Puede emplearse esencialmente la misma metodología que la dada a conocer anteriormente en el presente documento para hacer que estas evaluaciones sean muy asequibles. Considérese, por ejemplo, el caso en el que es deseable evaluar un gran estudio para determinar el poder predictivo de un conjunto de N marcadores, y estos marcadores son marcadores del número de copias que pueden evaluarse mediante recuento varietal. En principio, puede usarse codificación de muestras, combinación y captura y secuenciación en paralelo para someter a ensayo un gran número de muestras en un carril de secuenciación individual. La principal diferencia es que no hay limitación a cantidades diminutas de ADN, de modo que no se necesita WGA, y los ensayos serán más robustos. Se ha mostrado en el trabajo sobre el autismo que pueden obtenerse excelentes datos del número de copias a partir de la profundidad de lectura tras la captura. Por tanto, puede usarse una preparación de bibliotecas directa simplemente con etiquetas de muestra. Es probable que la zona objetivo para la captura sea diferente y más pequeña que el exoma. Por tanto, pueden combinarse más muestras por carril de secuenciación, y pueden reducirse los costes de secuenciación para someter a ensayo cientos de marcadores en cientos de muestras. Evidentemente, hay costes asociados con el desarrollo y la realización de pruebas de los reactivos de captura, y el coste de la propia captura, y no son despreciables, pero si una prueba llega a ser de uso clínico extendido, será económica, y éste es uno de los principales objetivos. Existe una alternativa clara, concretamente crear un banco de cebadores de PCR específicos de región con etiquetas y etiquetas de muestra. Las etiquetas serán necesarias porque no habrá extremos únicos al azar creados mediante cizalladura disponibles para métodos basados en recuento. Este plan alternativo puede ser más económico y más robusto que la captura (véase la figura 10).

Mediciones de ARN de células individuales

Ahora se han publicado varios métodos mediante los cuales los investigadores han determinado el perfil del

contenido en ARN de células individuales mediante secuenciación (Tang *et al*., 2009). Esto es potencialmente un método muy potente para el análisis de células tumorales. Un método de recuento varietal puede ser muy útil para la determinación del perfil de ARN a partir de células individuales, y por los mismos motivos por los que es útil con ADN. Cuando no se introduce variedad mediante procesamiento de ácido nucleico el recuento varietal proporciona una manera de evitar las distorsiones creadas mediante PCR. Puede analizarse ARN y ADN a partir de las mismas células. El análisis de ADN permite agrupar células en subpoblaciones estromales y tumorales, y esto facilita a su vez la interpretación de la determinación de perfil de ARN. Como ejemplo de utilidad, se supone que interesa investigar miles de células en circulación, o células de una biopsia, o de los bordes de un tumor, para detectar células tumorales poco frecuentes. El número de copias de ADN puede facilitar su identificación, y entonces puede usarse de manera fiable el ARN de esa célula para deducir propiedades adicionales, tales como el tejido de origen. Esta combinación es especialmente potente para la detección temprana de la nueva incidencia o recidiva de cáncer.

Ejemplos

Secuenciación de metilación

Actualmente, la secuenciación de alto rendimiento requiere la purificación física de regiones subgenómicas de interés con el fin de obtener una alta cobertura de esas regiones. Se ha aplicado el método de captura de secuencias para examinar la metilación en grandes conjuntos de islas de CpG a nivel de secuencia. La secuenciación de metilación se basa en la conversión de residuos de C en T usando reactivos de bisulfito, detectándose las C metiladas como las protegidas frente a la conversión de C en T. Esto significa o bien que la captura de secuencias debe tener lugar antes del tratamiento con bisulfito, lo cual requiere más ADN de partida de lo que está disponible para muestras de pacientes, o bien que la captura de secuencias debe diseñarse para adaptarse a la conversión de C en T. El método de secuenciación basado en bisulfito usado en el presente documento adopta esto último y por tanto es adecuado para la secuenciación de moléculas individuales de alto rendimiento mediante los métodos o bien de Illumina o bien de Roche 454, usando tan sólo 100 ng de ADN de entrada. El método de secuenciación depende de la captura de híbridos de ADN tratado con bisulfito seguido por secuenciación mediante el secuenciador Illumina Solexa y mapeo de vuelta al genoma usando un algoritmo desarrollado por el Dr. Andrew Smith (19). Se observó que las islas de CpG pueden mostrar múltiples niveles diferentes de metilación, desde improntadas (alélicas) hasta parcialmente metiladas al azar, hasta completamente metiladas o sin metilar. En particular, una observación que es importante para esta invención es que muchas islas cambian de completamente sin metilar en tejidos normales hasta metiladas de manera prácticamente completa en tejido tumoral coincidente ("dominios de cambio de metilación del ADN"). La figura 3 muestra tres ejemplos de los casos más sencillos en células normales frente a la línea celular de cáncer de mama MDA-MB-231: sin metilar de manera estable (A y B); el gen ALX, cambiado de sin metilar a completamente metilado (C y D); y un caso, SSTR4, en el que sólo se cambia un segmento de la isla (E y F). Este cambio completo hace posible detectar ADN de células tumorales poco frecuentes en líquidos corporales u otras poblaciones celulares usando PCR anidada de ADN tratado con bisulfito y abre la posibilidad de la creación de un método para la detección temprana de recidiva en forma de células metastásicas circulantes.

Hibridación genómica comparativa (CGH) mediante secuenciación de alto rendimiento

Hasta ahora, el método más económico para el análisis genómico de alta resolución ha empleado alguna forma de micromatriz. El rápido surgir de la secuenciación de ADN de "nueva generación" está cambiando ese panorama. Los instrumentos tanto de Illumina (Solexa) como de Roche 454 pueden proporcionar millones de lecturas de secuencias individuales distribuidas al azar a lo largo de cualquier ADN de entrada (Craig *et al*., 2008). La profundidad de cobertura de estas tecnologías depende de la longitud total del ADN de entrada, desde el genoma completo (amplitud grande, profundidad inferior) hasta regiones estrechamente enfocadas logradas mediante captura de híbridos de secuencias específicas (amplitud estrecha, profundidad superior) (Albert *et al*., 2007; Hodges *et al*., 2007; Okou *et al*., 2007). Convertir lecturas de secuenciación en análisis de número de copias es una cuestión de mapear las lecturas y después ponerlas en "grupos" de un determinado tamaño, dependiendo del propósito, contar los números en los grupos y realizar un análisis estadístico apropiado para definir puntos de rotura de cambio de número de copias.

Se aplicó recuento de secuencias al ADN incrustado en parafina, fijado con formalina, degradado (FFPE) que estaba disponible de ensayos clínicos retrospectivos con resultados muy alentadores. La figura 4 presenta una comparación de resultados de micromatriz y recuento de secuencias en una región alrededor de Her2 en el cromosoma 17 para una muestra de ensayo clínico incrustada en parafina de 10 años de antigüedad, JZ33. Los puntos grises y naranjas son (respectivamente) los datos de razón sin procesar para la micromatriz 244K Agilent y un carril de lecturas de secuenciación Solexa divididas en grupos de 30 K. Obsérvese que la pista de secuenciación (naranja) muestra significativamente menos ruido. Las líneas verde y azul son el resultado de segmentar cada tipo de datos. Los segmentos concuerdan generalmente, pero el método de recuento de secuencias revela un patrón algo más detallado. Otra característica de interés de la figura 4 es la robustez del presente programa de segmentación, independientemente de la variación de ruido en los dos conjuntos de datos sin procesar.

La eficacia del recuento de secuencias con ADN FFPE y el coste en rápido decrecimiento de la secuenciación de ADN en general proporcionan una oportunidad para convertir los métodos de micromatriz anteriores en una

plataforma altamente flexible y ajustable a escala para el diagnóstico de genoma de cáncer de mama. Tanto la amplitud de ADN que va a secuenciarse como el tamaño y la ubicación (y por tanto resolución) de grupos individuales pueden ajustarse fácilmente para adaptarse a los loci particulares que van a someterse a ensayo.

Micromatriz y secuenciación de células individuales

Se usó una representación de ADN amplificado a partir de células individuales para la determinación de perfil genómico mediante "recuento de secuencias". Como etapa preliminar en el desarrollo de esa tecnología, se amplificó ADN a partir de lotes de 1, 10 y 100 células de la línea celular SKBR3 usando el kit de amplificación de genoma completo (WGA) de Sigma-Aldrich y se compararon esas representaciones amplificadas con micromatrices equivalentes en el mismo formato realizadas con una preparación de ADN de SKBR3 convencional. Se escogieron células individuales a partir de placas de cultivo adherentes usando una micropipeta de transferencia de células. Tras la amplificación mediante WGA1 al azar se digirieron las preparaciones de ADN con DpnII y AluI, se añadieron adaptadores y sólo se amplificaron mediante PCR fragmentos de DpnII sin sitios de AluI. Se marcaron estas representaciones y se aplicaron a micromatrices personalizadas ROMA con 390.000 sondas diseñadas para detectar fragmentos de DpnII de tamaño apropiado a lo largo del genoma. La idea era someter a prueba cómo de reproducible sería este procedimiento en dos etapas cuando sólo están presentes 2-6 cromosomas en cada preparación de células individuales (SKBR3 es aproximadamente triploide).

Los resultados se muestran en la figura 5, que presenta una comparación de los perfiles segmentados de una preparación en lotes de ADN de SKBR3 (figura 5A) con la presente línea celular de referencia masculina normal (SKN-1) y tres ADN de células individuales independientes (figuras 5B, C, D). Una inspección visual de estos perfiles deja claro que, tras la segmentación, los resultados para células individuales son prácticamente idénticos a un ADN convencional, llevando a confiar en que la secuenciación de ADN a partir de una representación proporcionará resultados útiles. Aunque en este caso no se representan los datos de razón sin procesar, se observa que la desviación estándar de los datos de razón es mucho mayor para los tres experimentos de células individuales (0,706, 0,725, 0,676) que para la preparación de ADN de lotes (0,313). El ruido adicional no afecta de manera adversa a la segmentación, aunque en determinadas regiones pueden perderse detalles muy finos (flecha roja en la figura 5A). Esto es comparable a otros experimentos de WGA1 tales como el representado en gris en la figura 4. A la vista de esto, también merece la pena observar que el presente uso particular del método de WGA1 no depende de una amplificación absolutamente uniforme a lo largo del genoma. Se aprovecha la redundancia de datos que se obtiene de la segmentación para identificar puntos de rotura. Aunque sólo se amplifique la mitad de los posibles fragmentos a partir de una célula individual en una reacción dada, todavía se obtendrán datos útiles.

Secuenciación a lo largo de puntos de rotura del cromosoma

Uno de los métodos para la detección de células tumorales poro frecuentes, en ganglios linfáticos, sangre u otros líquidos corporales, es la amplificación o secuenciación de puntos de rotura únicos resultantes de deleciones y/o translocaciones que se producen durante el desarrollo de tumor de mama. Aunque acontecimientos tales como deleciones y duplicaciones resultan evidentes a partir de CGH en matriz, la estructura real subyacente al cambio de número de copias no lo es.

Se enriquecieron las secuencias de ADN que rodean a 22 puntos de rotura a partir de dos líneas celulares de cáncer de mama mediante captura de híbridos en micromatrices personalizadas tal como se describió por Hodges *et al* (20). Tras la captura de híbridos, se amplificó el ADN enriquecido y se preparó para la secuenciación de "extremos emparejados" en el instrumento Illumina GA2. La secuenciación de extremos emparejados proporciona aproximadamente 36 pb de secuencia desde cada extremo de una molécula de ADN individual (tras la preparación para secuenciación, cada molécula tiene aproximadamente 200 pb de longitud), y después se mapean esas secuencias de vuelta al genoma consenso. Para una molécula de una región normal, no reordenada, las dos lecturas coincidentes deben mapearse con una separación de aproximadamente 130 pb. En cambio, en los bordes de una deleción, las dos lecturas de una molécula individual deben estar separadas en el mapa por la anchura de la deleción, o en el caso de una translocación, se mapearán en otro cromosoma. Estos se denominan "fragmentos de puente" porque conectan regiones que normalmente no son adyacentes entre sí. En la parte superior de la figura 6 se muestran ejemplos de dos fragmentos de puente de este tipo, junto con el patrón de CGH original, y gráficos de la distancia entre extremos emparejados (azul) y el número total de lecturas (rojo) a lo largo de la región de captura (eje X). Mapeando la ubicación exacta de un punto de rotura, pueden prepararse sondas altamente sensibles y específicas que distinguirán células tumorales de grandes números de células normales. De 22 regiones que presentaban deleciones capturadas a partir de las líneas celulares T47D y MDA-MB-436, 15 se comportaron como deleciones individuales a partir de las cuales pudieron identificarse "fragmentos de puente" únicos como en la figura 6. Dos sitios adicionales "saltaron" a cromosomas diferentes lo que indica translocaciones. Los otros 5 sitios implicaron secuencias repetidas en uno de la pareja para el que no era posible la identificación del punto de rotura exacto y que no sería adecuado para crear sondas únicas.

Determinación del perfil de número de copias usando etiquetas varietales con fragmentos de enzimas de restricción.

Como ilustración de recuento varietal, se obtuvo información del número de copias usando un "método de moléculas de ácido nucleico con etiqueta diferentes totales". En primer lugar se asociaron fragmentos de ADN con etiquetas

usando transferasa terminal y ligación por mellas (figura 9). Se obtuvo ADN genómico a partir de la línea celular SKBR3. Se digirió ADN (2 μg) mediante NlaIII, dando como resultado fragmentos de ADN con proyecciones en 3', "CATG", de 4 pb. Este extremo 3'OH se extendió con transferasa terminal añadiendo una cola de poli-A. Uno de los adaptadores de cebador se diseñó para tener etiquetas, 12 T y CATG en el extremo 3', de modo que podía aparearse con los extremos 3' de los fragmentos de ADN con un sitio de mellas (figura 9, producto 3). Este sitio de mellas se ligó mediante ADN ligasa, formando fragmentos de ADN en dúplex con etiquetas. Después se añadió otro adaptador de cebador a estos fragmentos de ADN en dúplex mediante un ciclo de PCR (figura 9, etapa 5), permitiendo amplificar adicionalmente mediante PCR los productos resultantes de esta reacción, seguido por secuenciación (Illumina/Solexa). Después, mediante mapeo de la subsecuencia de cada lectura de secuencia asociada a etiqueta a una ubicación en el genoma, se contó el número de fragmentos de ADN (asignados a la misma ubicación genómica) con diferentes etiquetas.

Anteriormente se obtuvieron perfiles del número de copias de ADN genómico de SKBR3 a partir de una biblioteca de un millón de células, mediante secuenciación y recuento de secuencias. La figura 11 compara el uso de recuento varietal (un "método de moléculas de ácido nucleico con etiqueta diferentes totales") con el "recuento de secuencias" para analizar datos del número de copias, en todo el genoma (figura 11A) y en parte del genoma (figura 11B). Los resultados esencialmente equivalentes obtenidos usando el recuento varietal ilustran que, añadiendo etiquetas varietales a fragmentos de ADN antes de la PCR, puede usarse el recuento varietal tal como se describe en el presente documento para obtener de manera fiable información del número de copias no afectada por distorsión por amplificación.

El análisis de datos para el procedimiento descrito anteriormente empleó una modificación del método de "recuento de grupos" porque los sitios de restricción de NlaIII no están distribuidos de manera uniforme en el genoma. Por tanto, con el fin de garantizar un peso igual para cada uno de los grupos de 50 K, los límites de grupo para este experimento se diseñaron para abarcar un número igual de fragmentos de NlaIII con un tamaño de entre 100 pb y 250 pb, el intervalo de tamaño a partir del cual se prepararon las bibliotecas de Illumina.

Discusión

Los datos de células individuales tienen tanta información como los datos obtenidos a partir de múltiples células, por ejemplo a partir de hibridación en matriz de múltiples secciones de un tumor primario. En primer lugar, es posible permitirse analizar muchos más genomas mediante secuenciación de células individuales en multiplex que mediante hibridación en matriz (del orden de cien células individuales frente a una docena de secciones). En segundo lugar, es posible evaluar directamente la proporción exacta de células en una región de un tumor que son malignas. En tercer lugar, es posible realizar un análisis mucho más detallado del orden temporal de lesiones genéticas, ya que no se estará considerando el estado promedio de una población. En cuarto lugar, es posible evaluar si la heterogeneidad tumoral es de tipo anatómicamente segregado o cohabitación, o ambos.

La invención descrita en el presente documento puede usarse para determinar el número de copias de genoma original (o el perfil de expresión génica) de una muestra sin distorsión por amplificación. Todas las mediciones del número de copias o expresión génica, especialmente que implican múltiples muestras, son aplicaciones posibles. Las muestras pueden ser ácidos nucleicos derivados de células individuales o poblaciones de células. La invención puede usarse para estudiar la heterogeneidad celular en tumores o poblaciones de células normales.

La invención puede usarse para generar un perfil genómico de una célula, por ejemplo, una célula cancerosa, o población de células. A partir del perfil genómico puede determinarse si genes específicos están amplificados o delecionados. Cuando la muestra se deriva de ARN expresado, y las regiones son genes expresados, los recuentos relativos comprenden un perfil de expresión. A partir del perfil de expresión puede identificarse el tejido de origen de una célula cancerosa. Cuando la muestra es una mezcla de genomas a partir de especies diferentes, y las regiones genómicas distinguen las especies diferentes, los recuentos relativos comprenden un censo de esas especies diferentes. Por ejemplo, a partir de un censo puede determinarse la población de flora microbiana y relacionarse la misma con la enfermedad.

Una de las ventajas del uso de etiquetas es que permite el uso de fragmentos de endonucleasa de restricción en el análisis del genoma. Los métodos de recuento para la determinación del número de copias del genoma usan recuentos de lecturas independientes. Dos lecturas no se consideran independientes si son el resultado de amplificación de ADN que se produce durante el procesamiento de la muestra. Normalmente, esto conlleva someter a cizalladura fragmentos de ADN de muestra, y se usa la posición de extremos de los fragmentos para distinguir duplicados de PCR. Si dos lecturas de ADN tienen los mismos extremos, pueden ser duplicados de PCR, y no independientes, por tanto se cuentan como una única lectura verdadera. No pueden distinguirse fragmentos de restricción de esta manera, porque dos lecturas del mismo fragmento tendrán los mismos extremos. Sin embargo, mediante la adición de etiquetas, pueden distinguirse lecturas independientes, si se añaden antes de cualquier etapa de amplificación, ya que tendrán etiquetas diferentes. En cambio, si dos lecturas de la misma región del genoma tienen la misma etiqueta, es altamente probable que sean duplicados de PCR, y no independientes.

La capacidad de usar la secuencia de fragmentos de restricción en el análisis del genoma puede aprovecharse para reducir el coste de determinación del perfil del número de copias, de manera muy similar a cómo realizar ROMA

facilita el análisis del genoma. Mediante selección por tamaño de fragmentos, pueden tomarse muestras del genoma a mayor profundidad en menos loci, y todavía obtener un perfil del número de copias de alta resolución pero con menos lecturas y por tanto a menor coste. Esta es la esencia de la representación, y da como resultado una reducción de la complejidad, el elemento central de ROMA. En ROMA, el análisis de la representación se realiza mediante hibridación con matrices, y por tanto se analiza la representación mediante secuenciación.

Puede lograrse una reducción adicional de la complejidad mediante captura de secuencia. La captura de secuencias puede diseñarse específicamente para los fragmentos que puede predecirse mediante métodos *in silico* que se mapearán a las regiones en las que se desea la mutación del número de copias. De esta manera, pueden seleccionarse como diana regiones especiales del genoma, tales como el exoma, o genes cancerosos conocidos. Entonces se necesita muy poca secuencia de la muestra para la determinación del número de copias, y pueden multiplexarse muestras con un gran ahorro de coste.

La invención dada a conocer en el presente documento también puede usarse para determinar la proporción de componentes en una mezcla. Por ejemplo, se supone que se dispone de un conjunto de reactivos, tales como moléculas de anticuerpo, que se unen a un conjunto de sustratos, tales como diferentes proteínas, y se desea detectar la proporción relativa de esos sustratos en una muestra, tal como sangre. En primer lugar, se combina de manera covalente una etiqueta que tiene una etiqueta de muestra con cada reactivo, en el que la porción de etiqueta de muestra identifica el reactivo, y la etiqueta varietal se usará en el recuento. Después se combinan los reactivos con etiqueta con la muestra, y se separan físicamente los reactivos que se combinan con sustratos. Por ejemplo, los sustratos pueden biotinilarse antes de mezclarse con reactivos, después se separan los reactivos que se combinan con sustratos mediante cromatografía de afinidad con avidina. Las etiquetas se amplifican mediante PCR y se secuencian, proporcionando eficazmente recuentos relativos para cada especie de reactivo y por tanto cada sustrato. En este caso, las porciones constantes de las etiquetas incluyen el identificador de muestra y nucleótidos usados para la amplificación.

La invención proporciona las siguientes ventajas: (1) el método es sencillo; (2) con análisis estadístico puede determinarse la confianza de una medición con gran precisión; (3) este es un método basado en secuencia, que a largo plazo será más económico que las matrices; (4) no se necesitan longitudes de lectura largas, proporcionando ahorros; (5) dado que se cuentan moléculas de ácido nucleico con etiqueta, muchos métodos de procesamiento que crean distorsión en los rendimientos pueden aplicarse sin una pérdida principal en el recuento de moléculas de ácido nucleico con etiqueta; (6) entre estos métodos de procesamiento se encuentra la selección de híbridos (también conocida como micromatrices de captura), que limita la información de secuencia a regiones de importancia, y por tanto reduce el coste reduciendo el número total de lecturas por muestra necesarias para la determinación de la información deseada; (7) las moléculas de una muestra pueden combinarse con etiquetas y una etiqueta de muestra, creando moléculas de ácido nucleico con etiqueta de muestra que entonces pueden combinarse, y tras el procesamiento y la secuenciación, se descodifican las etiquetas de muestra, proporcionando un perfil para cada muestra; (8) la combinación añade eficacia y uniformidad; (9) las muestras pueden ser células individuales, proporcionando grandes números de mediciones de células individuales de manera económica.

La invención descrita en el presente documento tiene aplicaciones clínicas tal como se describió anteriormente, incluyendo: (1) determinación de la heterogeneidad genómica de tumores; (2) detección de células tumorales en biopsias de aguja, en particular una evaluación de la heterogeneidad de las células en una biopsia; (3) evaluación de la propagación de un cáncer hasta ganglios linfáticos regionales; y (4) detección de células malignas en sitios tales como sangre y médula ósea. La presencia de células con la firma molecular del primario en la sangre o médula ósea abre la posibilidad de usar ensayos de células/moléculas individuales más generales como primera línea de defensa en la detección temprana de cáncer, es decir cáncer de mama en mujeres en riesgo.

Cuando hay pruebas radiológicas de bulto en la mama, la paciente se somete lo más habitualmente a biopsia con aguja fina, que examina el patólogo quien realiza una evaluación de la malignidad del bulto. Dado que la aspiración con aguja destruye la histología de la biopsia, el patólogo no dispone de una de sus herramientas canónicas para evaluar la malignidad. La secuenciación de células individuales no sólo predecirá con mayor precisión la presencia de células malignas, sino que también proporcionará lo que puede llamarse “estadificación genómica”. ¿Es el tumor un tipo sencillo de tipo complejo? ¿Es monogenómico o poligenómico? ¿Hay loci críticos que están amplificados o delecionados que puedan sugerir inmediatamente una terapia preoperatoria?

El enfoque de células/moléculas individuales permite la evaluación de la propagación de cáncer a ganglios linfáticos regionales y establecer los estados de metilación del ADN en regiones de cambio.

La aplicación satisfactoria de la detección de células/moléculas individuales en los contextos anteriores conducirá a la aplicación a la detección de recidiva de cáncer, y la medición de la carga de cáncer. La detección de recidiva lo antes posible puede prolongar la supervivencia para una gran clase de pacientes mediante intervención más temprana de lo que recibirían de otro modo. Además, si se monitoriza la carga de cáncer en el paciente, incluso antes de que pueda detectarse el sitio del crecimiento de cáncer mediante obtención de imágenes, podrá medirse la respuesta del cáncer a una variedad de medidas terapéuticas. Es decir, puede determinarse empíricamente en el paciente la opción más eficaz de terapia.

Existen aplicaciones adicionales en el diagnóstico y la selección de tratamiento para pacientes con enfermedades autoinmunitarias, y examen de fecundación *in vitro* (IVF) de embriones previos al implante para detectar defectos genéticos tales como trisomía 21.

Bibliografía

1. Pinard R, de Winter A, Sarkis GJ, Gerstein MB, Tartaro KR, Plant RN, *et al.* Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. BMC Genomics. 2006; 7:216. PMCID: PMC1560136.

2. Klein, C. *et al.* Comparative genomic hybridization, loss of heterozygosity, and DNA sequence analysis of single cells. PNAS. 1999; 96, págs. 4494-4499.

3. Stoecklein, N. *et al.* SCOMP Is Superior to Degenerated Oligonucleotide Primed-Polymerase Chain Reaction for Global Amplification of Minute Amounts of DNA from Microdissected Archival Tissue Samples. American Journal of Pathology. 2002; 161(1):43-51.

4. Parameswaran *et al.* A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. Nucleic Acids Research. 2007; 35(19): e130.

5. Patente estadounidense n.º 7.622.281, Ronaghi *et al.*

6. Eid *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. Science. 2009; 323:133-138.

7. Patente estadounidense n.º 5.639.603, Dower *et al.*

8. Publicación de solicitud de patente estadounidense n.º 2006/0073506, Christians *et al.*

9. Navin N, Krasnitz A, Rodgers L, Cook K, Meth J, Kendall J, *et al.* Inferring tumor progression from genomic heterogeneity. Genome Res. 2010; 20(1):68-80. PMCID: 2798832.

10. Fisher B, Redmond CK, Fisher ER. Evolution of knowledge related to breast cancer heterogeneity: a 25-year retrospective. J Clin Oncol. 2008; 26(13):2068-71.

11. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10(3):R25. PMCID: PMC2690996.

12. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987; 4(4):406-25.

13. De Baetselier P, Roos E, Brys L, Remels L, Gobert M, Dekegel D, *et al.* Nonmetastatic tumor cells acquire metastatic properties following somatic hybridization with normal cells. Cancer Metastasis Rev. 1984; 3(1):5-24.

14. Duelli DM, Padilla-Nash HM, Berman D, Murphy KM, Ried T, Lazebnik Y. A virus causes cancer by inducing massive chromosomal instability through cell fusion. Curr Biol. 2007; 17(5):431-7.

15. Jorgensen HF, Adie K, Chaubert P, Bird AP. Engineering a high-affinity methyl-CpG-binding protein. Nucleic Acids Res. 2006; 34(13):e96. PMCID: PMC1540740.

16. Meehan RR, Lewis JD, Bird AP. Characterization of MeCP2, a vertebrate DNA binding protein with affinity for methylated DNA. Nucleic Acids Res. 1992; 20(19):5085-92. PMCID: PMC334288.

17. Huang J., Pang J., Watanabe T., Ng HK, Ohgaki H. Whole genome amplification for array comparative genomic hybridization using DNA extracted from formalin-fixed, paraffin-embedded histological sections. J Mol Diagn. marzo de 2009; 11(2):109-16. Epub 5 de febrero de 2009.

18. Talseth-Palmer BA, Bowden NA, Hill A, Meldrum C, Scott RJ. Whole genome amplification and its impact on CGH array profiles. BMC Res Notes. 29 de julio de 2008; 1:56.

19. Hodges E, Smith AD, Kendall J, Xuan Z, Ravi K, Rooks M, Zhang MQ, Ye K, Bhattacharjee A, Brizuela L, McCombie WR, Wigler M, Hannon GJ, Hicks JB. High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. Genome Res. septiembre de 2009; 19(9):1593-605. Epub 6 de julio de 2009.

20. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR. Genome-wide in situ exon capture for selective resequencing. Nat Genet. diciembre de 2007; 39 (12):1522-7. Epub 4 de noviembre de 2007.

21. genome.ucsc.edu/index.html?org=Human&db=hg19&hgsid=171216665

22. Miner B.E., Stoger, R.J., Burden, A.F., Laird, C.D., Hansen R.S. Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. Nucleic Acids Research, 2004:32(17):e135.

23. Patente estadounidense n.º 7.537.897, concedida el 26 de mayo de 2009 (Brenner *et al.*).

24. McCloskey M.L., Stoger, R., Hansen, R.S., Laird, C.D. Encoding PCR Products with Batch-stamps and Barcodes. Biochem. Genet. 2007:45:761-767.

REIVINDICACIONES

1. Método para obtener, a partir de material genómico, información del número de copias genómicas no afectada por distorsión por amplificación, que comprende:

a) obtener segmentos del material genómico;

b) etiquetar los segmentos con etiquetas de ácido nucleico para generar moléculas de ácido nucleico con etiqueta únicas, de manera que cada una de las moléculas de ácido nucleico con etiqueta únicas comprende un segmento del material genómico de la etapa (a) y una etiqueta;

c) someter las moléculas de ácido nucleico con etiqueta a amplificación mediante reacción en cadena de la polimerasa (PCR);

d) generar lecturas de secuencias asociadas a etiquetas mediante secuenciación del producto de la etapa (c);

e) asignar cada molécula de ácido nucleico con etiqueta a una ubicación en un genoma asociado con el material genómico mediante mapeo de la subsecuencia de cada lectura de secuencia asociada a etiqueta correspondiente a un segmento del material genómico a una ubicación en el genoma; y

f) contar el número de moléculas de ácido nucleico con etiqueta que tienen una etiqueta diferente que se han asignado a la misma ubicación en el genoma,

obteniendo así información del número de copias genómicas no afectada por distorsión por amplificación.

2. Método según la reivindicación 1, que comprende además estimar un número de copias genómicas de una región del genoma que comprende más de una ubicación en el genoma asignando como número de copias de la región el mayor recuento obtenido en la etapa (f) para las ubicaciones dentro de la región.

3. Método según la reivindicación 1, que comprende además comparar un recuento obtenido en la etapa (f) para una ubicación en el genoma con un recuento para la misma ubicación obtenido a partir de una muestra de referencia, estimando así un número de copias genómicas relativo de la ubicación.

4. Método según la reivindicación 1, que comprende además

g) sumar los recuentos obtenidos en la etapa (f) para ubicaciones en el genoma que comprenden una primera región del genoma, en el que la primera región comprende más de una ubicación;

h) sumar los recuentos obtenidos en la etapa (f) para ubicaciones en el genoma que comprenden una segunda región del genoma, en el que la segunda región está compuesta por un número de ubicaciones que es comparable al número de ubicaciones de la primera región;

i) comparar el valor obtenido en la etapa (g) con el valor obtenido en la etapa (h),

estimando así el número de copias genómicas relativo de la primera región del genoma con respecto al número de copias genómicas de la segunda región del genoma.

5. Método según la reivindicación 4, en el que la etapa (h) comprende además

j) sumar los recuentos obtenidos en la etapa (f) para ubicaciones en el genoma que comprenden una tercera región del genoma, en el que la tercera región está compuesta por un número de ubicaciones que es comparable al número de ubicaciones de la primera región; y

k) obtener un promedio de la suma de los recuentos obtenidos en la etapa (f) para ubicaciones que comprenden la segunda región y la suma de los recuentos obtenidos en la etapa (f) para ubicaciones que comprenden la tercera región.

6. Método según la reivindicación 4, en el que la segunda región del genoma comprende un centrómero.

7. Método según la reivindicación 1, que comprende además sumar los recuentos obtenidos en la etapa (f) para ubicaciones que comprenden una región del genoma, y comparar la suma con una suma obtenida a partir de una muestra de referencia para la misma región del genoma, estimando así un número de copias genómicas relativo de la región del genoma.

8. Método para obtener, a partir de transcritos de ARNm, información del número de copias de ARNm no afectada por distorsión por amplificación, que comprende:

a) generar moléculas de ácido nucleico con etiqueta únicas, que comprende:

i) someter los transcritos de ARNm a una reacción de polimerasa en condiciones que fomentan la formación de tan sólo un complemento, generando así cadenas derivadas de primer orden;

ii) añadir una cola de polinucleótido a las cadenas derivadas de primer orden; y

iii) someter las cadenas derivadas de primer orden a una reacción de polimerasa en presencia de cebadores que pueden hibridarse con la cola de polinucleótido añadida en la etapa (ii) en condiciones que fomentan la formación de tan sólo un complemento, generando así cadenas derivadas de segundo orden,

en el que los cebadores de al menos una de las etapas (i) y (iii) comprenden etiquetas de ácido nucleico, de manera que cada molécula de ácido nucleico con etiqueta es única, generando así moléculas de ácido nucleico con etiqueta únicas;

b) someter las moléculas de ácido nucleico con etiqueta a amplificación mediante reacción en cadena de la polimerasa (PCR);

c) generar lecturas de secuencias asociadas a etiquetas mediante secuenciación del producto de la etapa (b);

d) asignar cada molécula de ácido nucleico con etiqueta a una ubicación en una biblioteca de ADNc asociada con los transcritos de ARNm mediante mapeo de la subsecuencia de cada lectura de secuencia asociada a etiqueta correspondiente a un transcrito de ARNm a una ubicación en la biblioteca de ADNc; y

e) contar el número de moléculas de ácido nucleico con etiqueta que tienen una etiqueta diferente que se han asignado a la misma ubicación en la biblioteca de ADNc,

obteniendo así información del número de copias de ARNm no afectada por distorsión por amplificación.

9. Método según una cualquiera de las reivindicaciones 1-7, en el que etiquetar los segmentos para generar moléculas de ácido nucleico con etiqueta comprende:

(i) añadir una cola de polinucleótido a los extremos de los segmentos del material genómico para generar cadenas derivadas de orden cero;

(ii) someter las cadenas derivadas de orden cero de la etapa (i) a una reacción de polimerasa en presencia de cebadores que pueden hibridarse con la cola de polinucleótido de las cadenas de orden cero en condiciones que fomentan la formación de tan sólo un complemento, generando así cadenas derivadas de primer orden;

(iii) añadir una cola de polinucleótido a las cadenas derivadas de primer orden;

(iv) someter las cadenas derivadas de primer orden a una reacción de polimerasa en presencia de cebadores que pueden hibridarse con la cola de polinucleótido de las cadenas derivadas de primer orden en condiciones que fomentan la formación de tan sólo un complemento, generando así cadenas derivadas de segundo orden,

en el que los cebadores de al menos una de las etapas (ii) y (iv) comprenden etiquetas de ácido nucleico, de manera que cada molécula de ácido nucleico con etiqueta es única, generando así moléculas de ácido nucleico con etiqueta únicas.

10. Método según una cualquiera de las reivindicaciones 1-7, en el que etiquetar los segmentos para generar moléculas de ácido nucleico con etiqueta comprende:

i) añadir una cola de polinucleótido a los extremos de los segmentos de material genómico para generar cadenas derivadas de orden cero,

ii) someter las cadenas derivadas de orden cero de la etapa (i) a una reacción de ligación en presencia de cebadores que pueden hibridarse con la cola de polinucleótido de las cadenas de orden cero añadidas en la etapa (i) en condiciones que fomentan la ligación de un cebador a los extremos 5’ de las cadenas derivadas de orden cero,

iii) someter el producto de la etapa (ii) a una reacción de polimerasa en presencia de cebadores que pueden hibridarse con la cola de polinucleótido añadida en la etapa (i) en condiciones que fomentan la formación de tan sólo un complemento, en el que los cebadores de la etapa (iii) tienen secuencias de nucleótidos diferentes de los cebadores de la etapa (ii), y en el que la polimerasa tiene actividad de corrección de lectura 3’-5’,

en el que los cebadores de al menos una de las etapas (ii) y (iii) comprenden etiquetas de ácido nucleico,

de manera que cada molécula de ácido nucleico con etiqueta es única, generando así moléculas de ácido nucleico con etiqueta únicas.

11. Método según la reivindicación 10, en el que añadir una cola de polinucleótido comprende el uso de una transferasa terminal.

12. Método según una cualquiera de las reivindicaciones 1-11,

en el que las moléculas de ácido nucleico con etiqueta se someten a captura de híbridos antes de PCR o antes de la secuenciación,

en el que las moléculas de ácido nucleico con etiqueta se generan a partir de una especie individual,

en el que se generan moléculas de ácido nucleico con etiqueta a partir de una célula individual,

en el que las moléculas de ácido nucleico con etiqueta se generan a partir de dos o más organismos, o

en el que las moléculas de ácido nucleico con etiqueta se generan a partir de dos o más especies, en el que las moléculas de ácido nucleico con etiqueta se generan a partir de una a población de microbios, y en el que se compara la información del número de copias genómicas obtenida para especies diferentes de la población para determinar el recuento relativo de esas especies diferentes en la población.

13. Método según una cualquiera de las reivindicaciones 1-12, en el que las secuencias de etiqueta comprenden además una etiqueta de muestra.

14. Método según la reivindicación 13, en el que las moléculas de ácido nucleico con etiqueta se combinan con una pluralidad de moléculas de ácido nucleico con etiqueta que tienen una etiqueta de muestra diferente antes de la amplificación mediante PCR o antes de la secuenciación.

15. Método según la reivindicación 14, que comprende además realizar la deconvolución de las lecturas de secuencias asociadas a etiquetas agrupando las lecturas de secuencias asociadas a etiquetas según la etiqueta de muestra.

16. Método según la reivindicación 1, en el que la ubicación en un genoma asociado con el material genómico es idéntica a la subsecuencia de una lectura de secuencia asociada a etiqueta correspondiente a una especie de molécula de ácido nucleico.

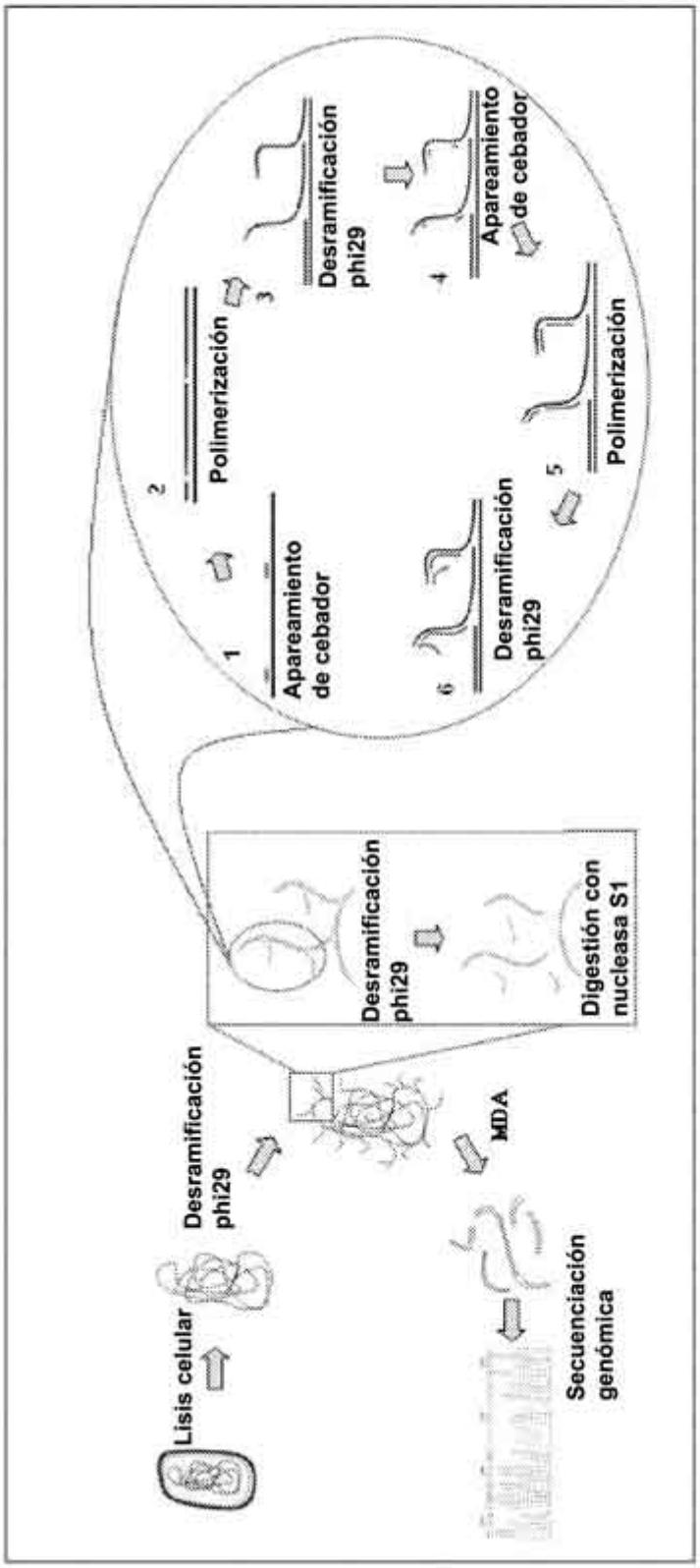
Lisis celular
Desramificación phi29
MDA
Secuenciación genómica
Desramificación phi29
Digestión con nucleasa S1
1
Apareamiento de cebador
2
Polimerización
3
Desramificación phi29
4
Apareamiento de cebador
5
Polimerización
6
Desramificación phi29

Fig. 1

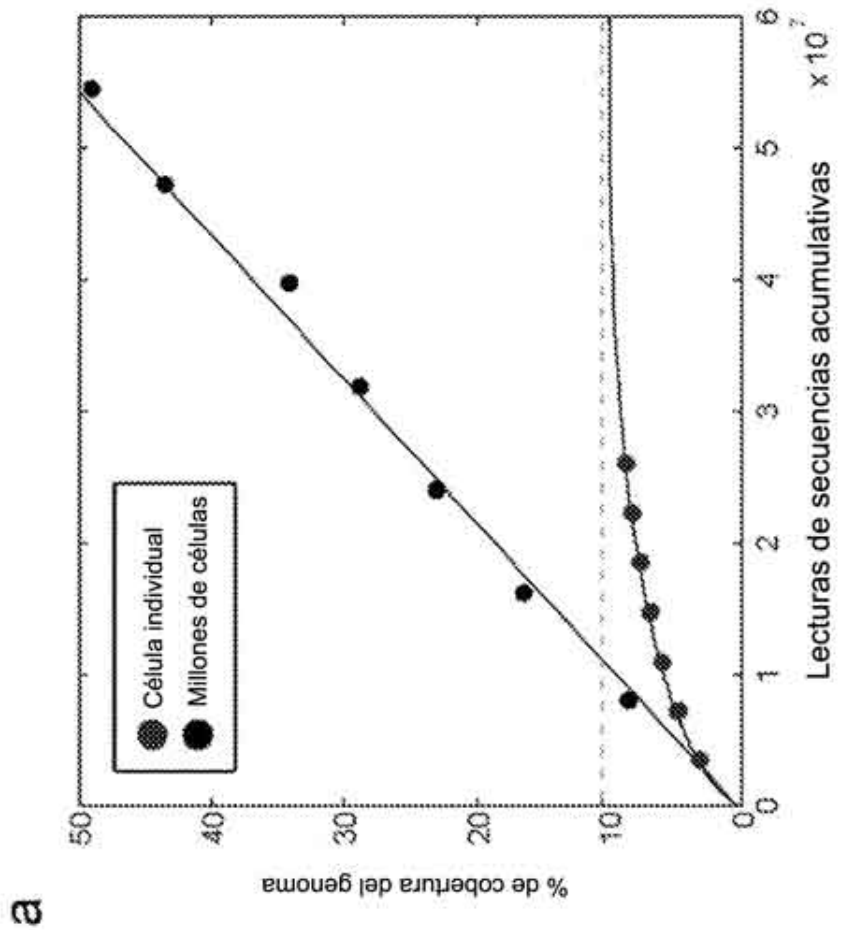
a
% de cobertura del genoma
Lecturas de secuencias acumulativas
x 10^7
Célula individual
Millones de células
0
10
20
30
40
50
1
2
3
4
5
6

Fig. 2

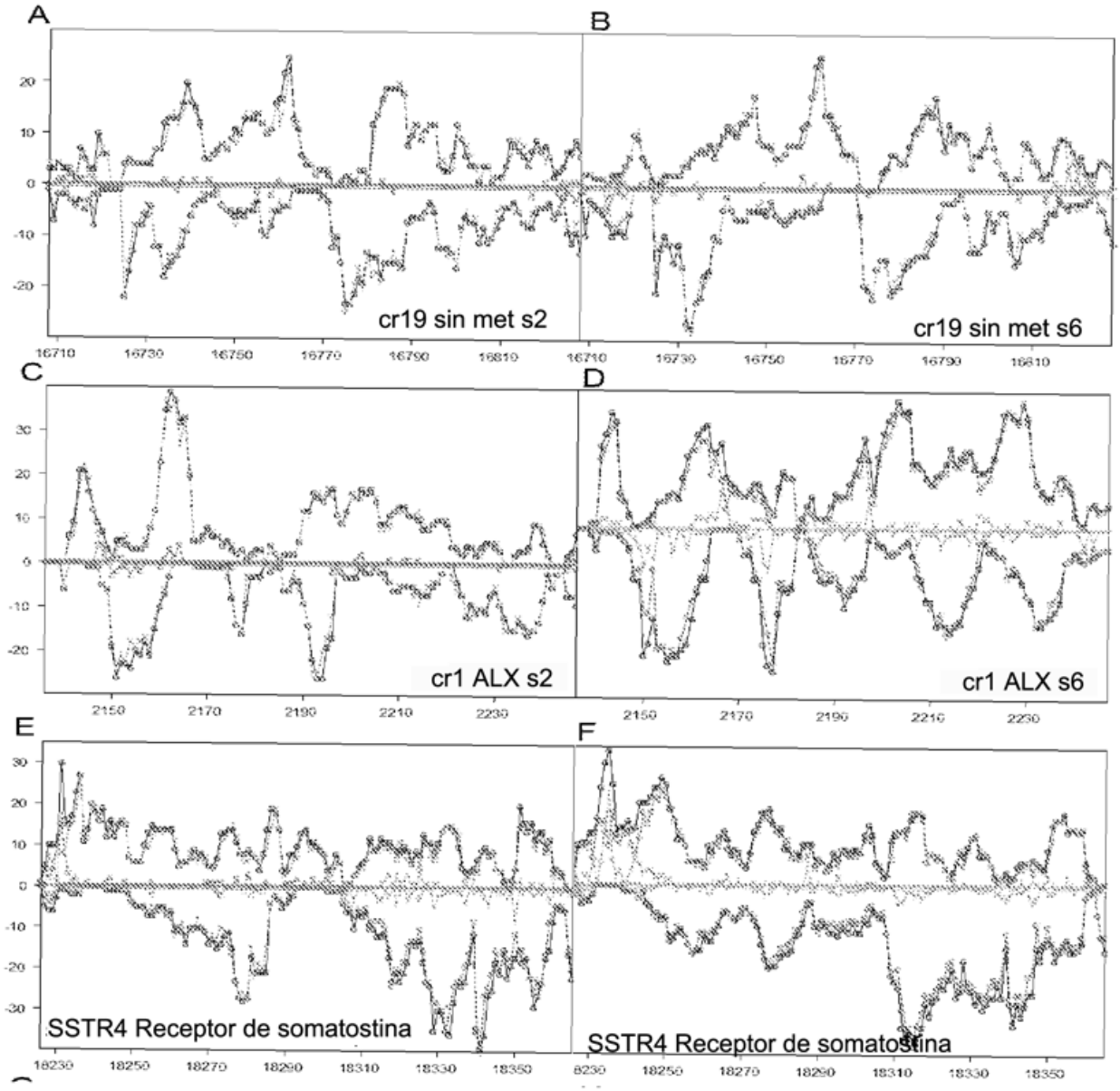
A
B
cr19 sin met s2
cr19 sin met s6
C
D
cr1 ALX s2
cr1 ALX s6
E
F
SSTR4 Receptor de somatostina
SSTR4 Receptor de somatostina

Fig. 3

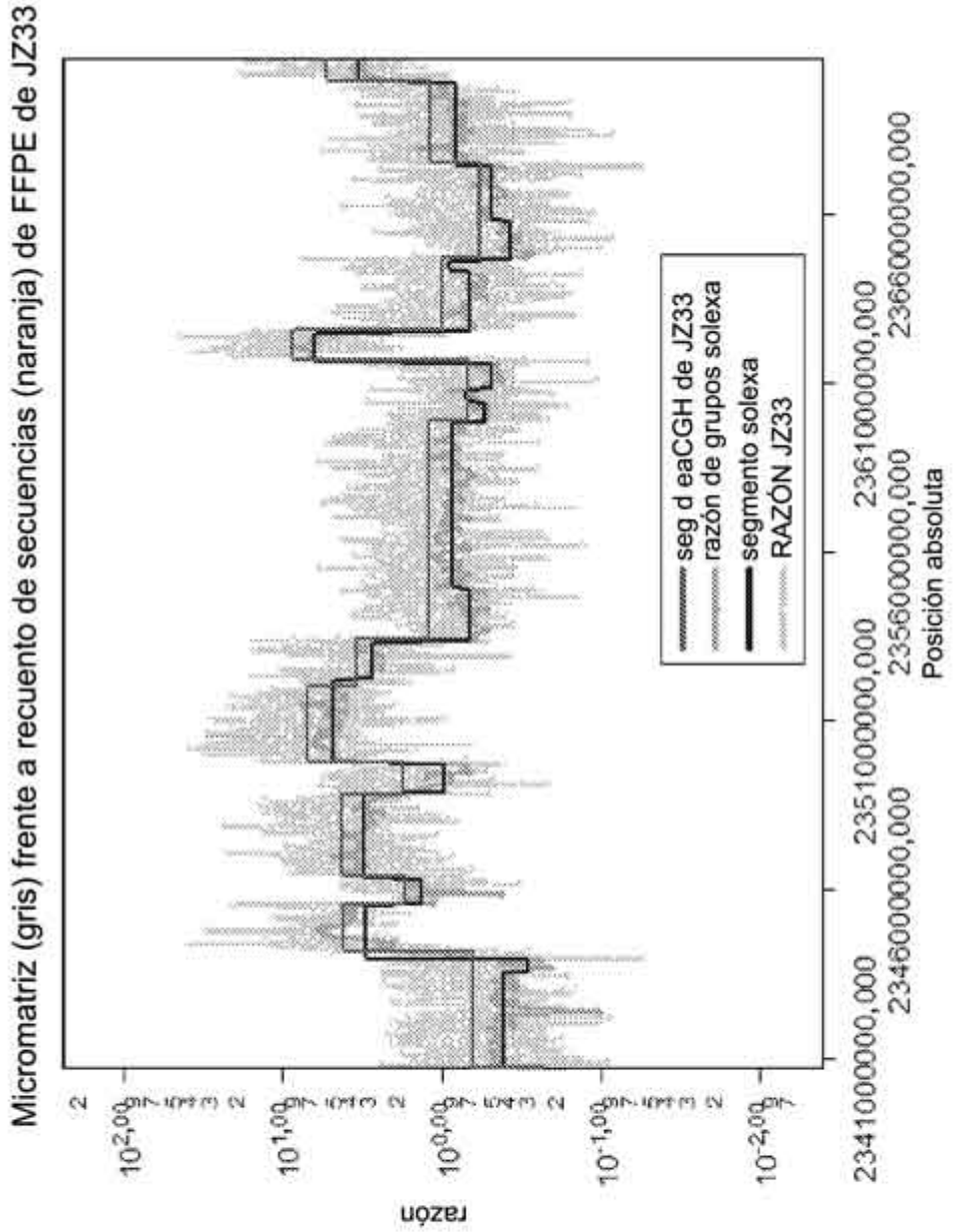
Micromatriz (gris) frente a recuento de secuencias (naranja) de FFPE de JZ33
razón
Posición absoluta
seg d eaCGH de JZ33
razón de grupos solexa
segmento solexa
RAZÓN JZ33
2341000000,000
2346000000,000
2351000000,000
2356000000,000
2361000000,000
2366000000,000

Fig. 4

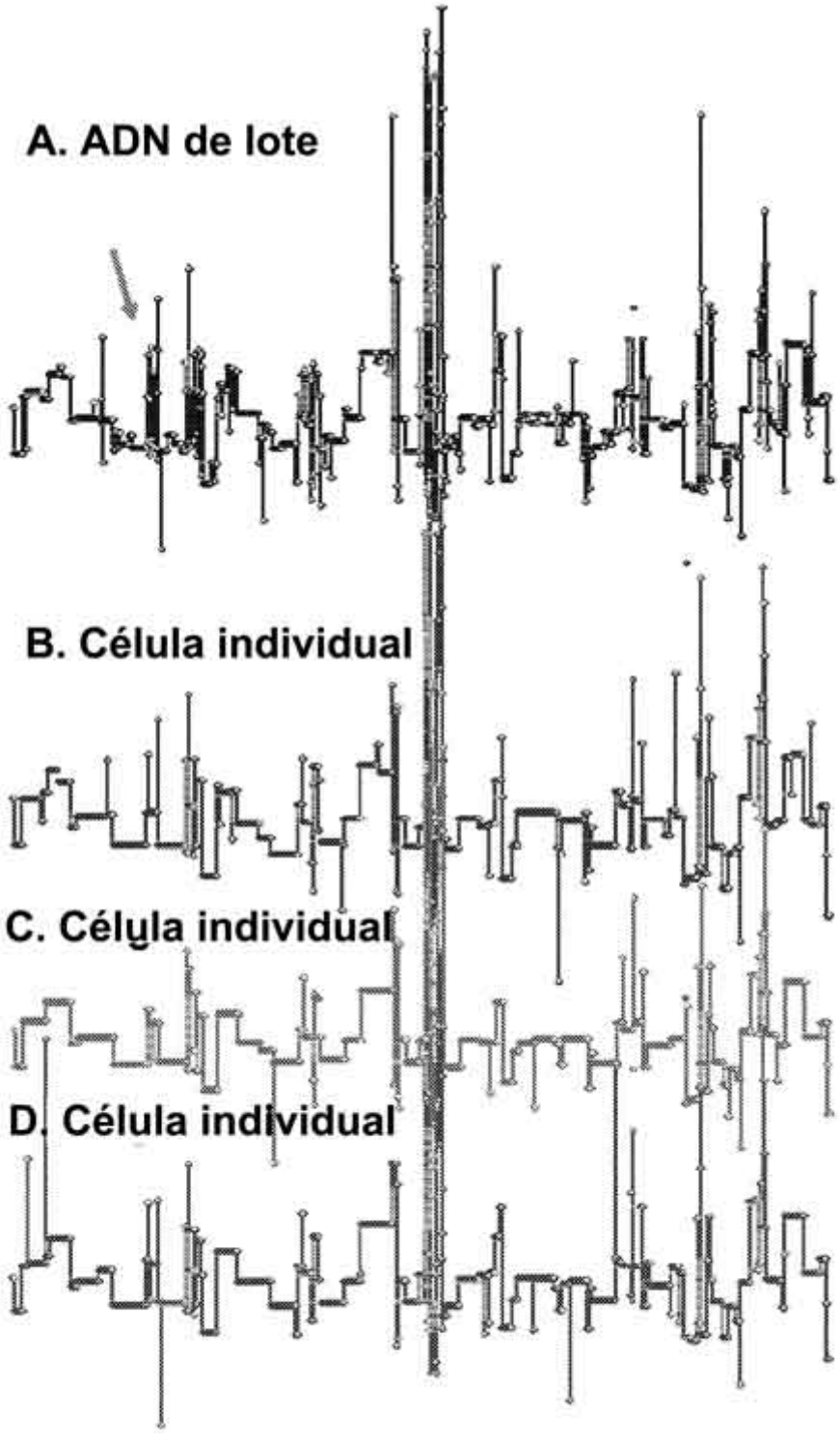
A. ADN de lote
B. Célula individual
C. Célula individual
D. Célula individual

Fig. 5

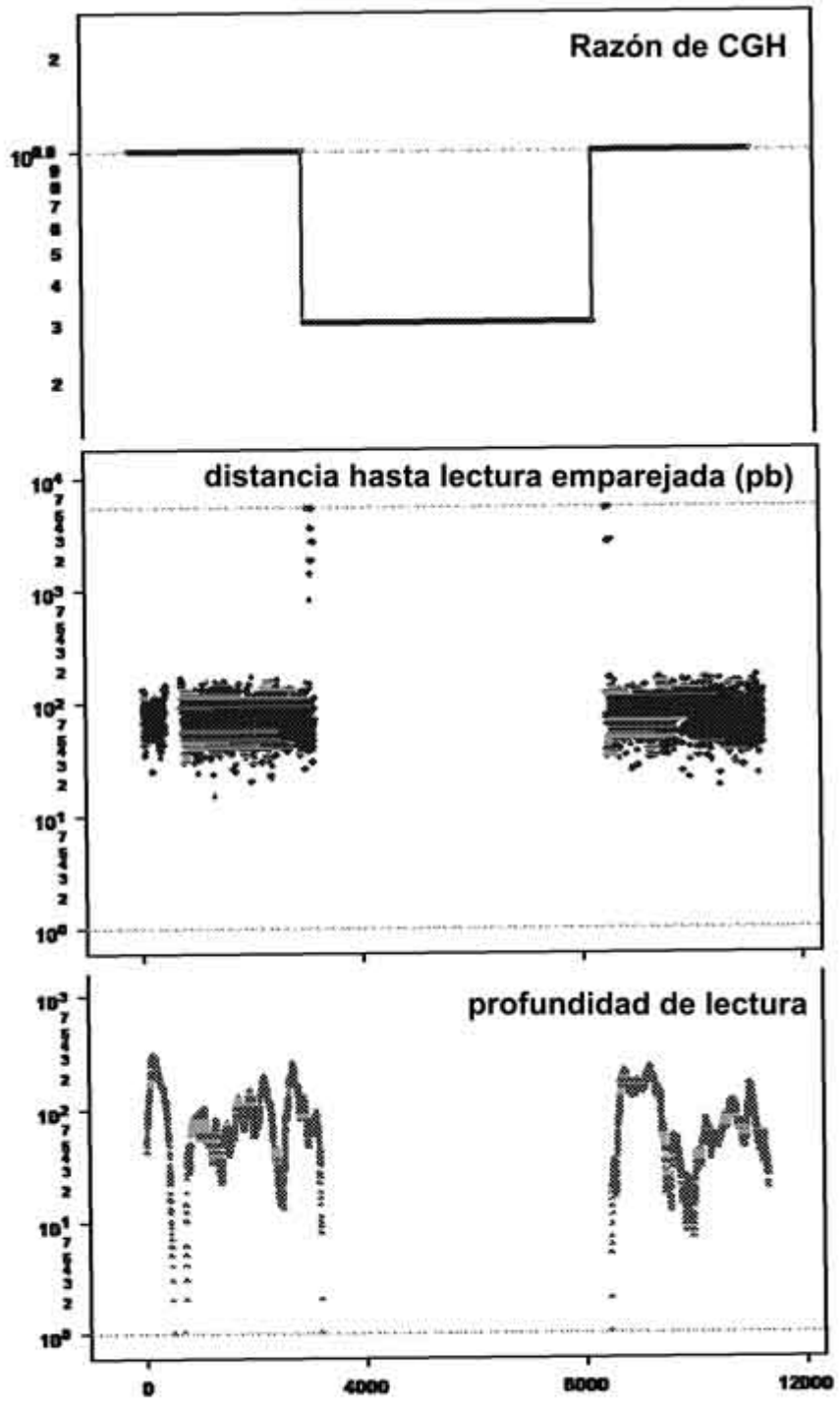
Razón de CGH
distancia hasta lectura emparejada (pb)
profundidad de lectura
0
4000
8000
12000

Fig. 6

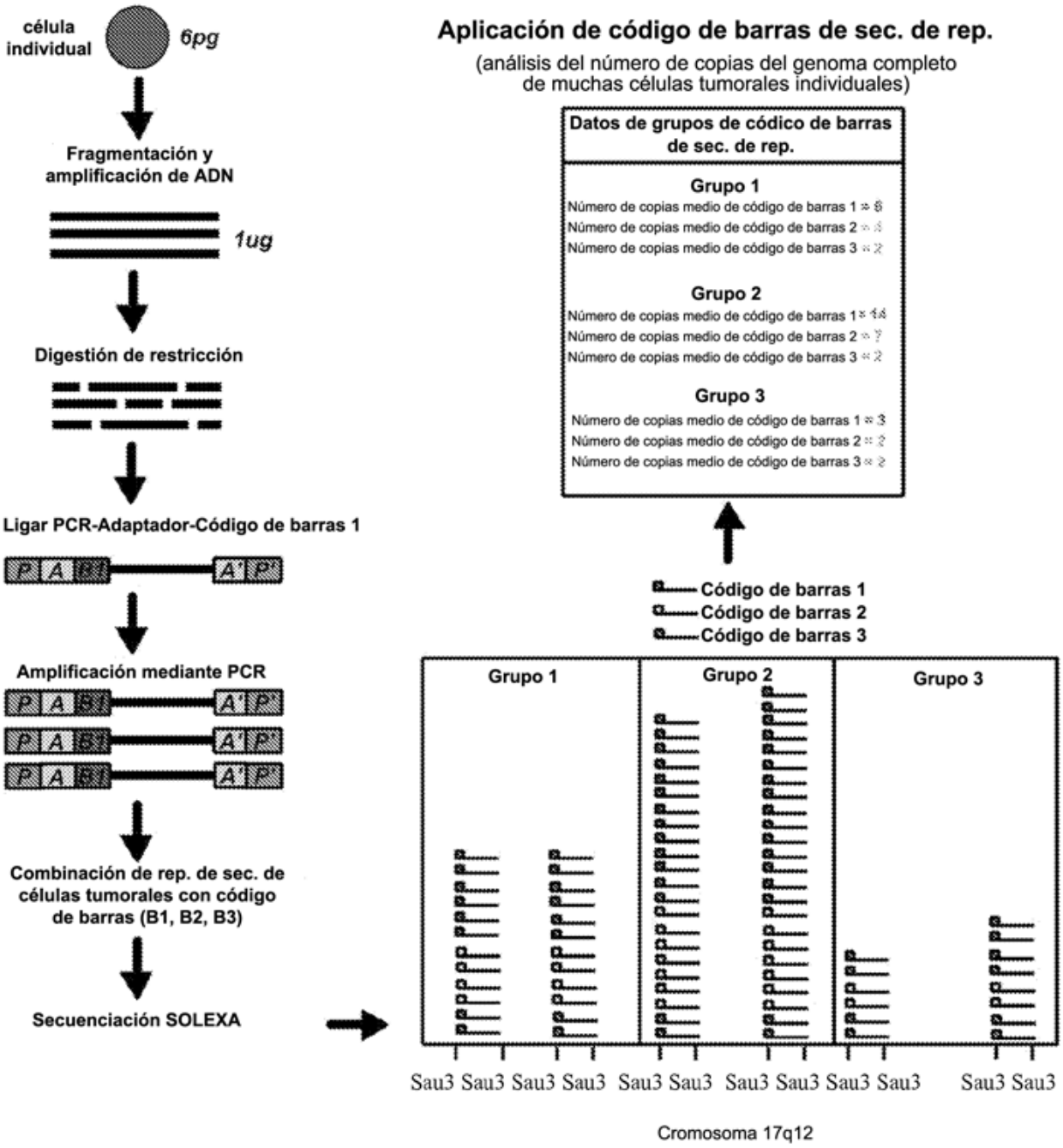
célula individual
6pg
Fragmentación y amplificación de ADN
1ug
Digestión de restricción
Ligar PCR-Adaptador-Código de barras 1
Amplificación mediante PCR
Combinación de rep. de sec. de células tumorales con código de barras (B1, B2, B3)
Secuenciación SOLEXA
Aplicación de código de barras de sec. de rep.
(análisis del número de copias del genoma completo de muchas células tumorales individuales)
Datos de grupos de códico de barras de sec. de rep.
Grupo 1
Número de copias medio de código de barras 1
Número de copias medio de código de barras 2
Número de copias medio de código de barras 3
Grupo 2
Número de copias medio de código de barras 1
Número de copias medio de código de barras 2
Número de copias medio de código de barras 3
Grupo 3
Número de copias medio de código de barras 1
Número de copias medio de código de barras 2
Número de copias medio de código de barras 3
Código de barras 1
Código de barras 2
Código de barras 3
Grupo 1
Grupo 2
Grupo 3
Sau3 Sau3 Sau3 Sau3 Sau3 Sau3 Sau3 Sau3 Sau3 Sau3 Sau3 Sau3
Cromosoma 17q12

Fig. 7

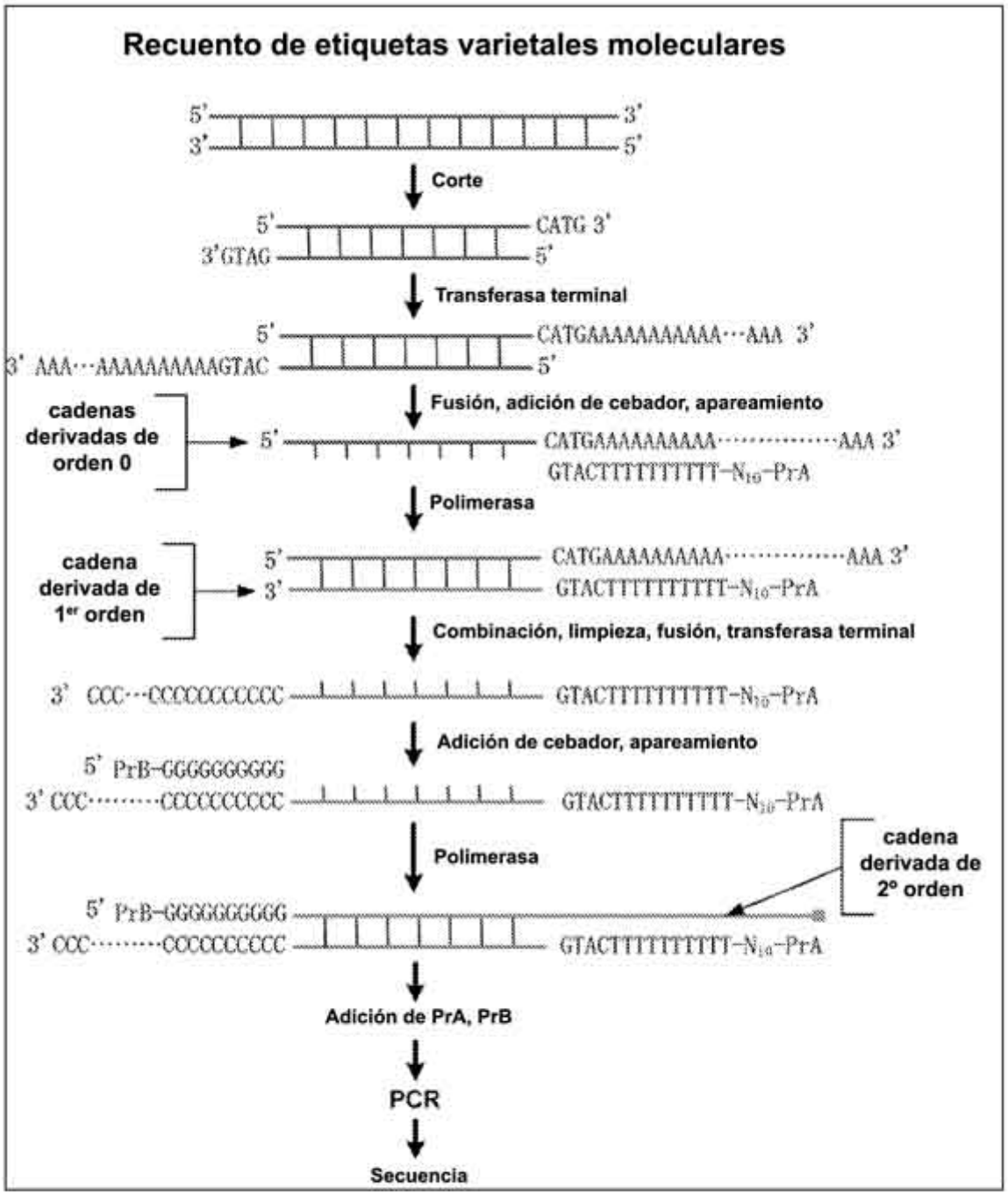
Recuento de etiquetas varietales moleculares
5'
3'
3'
5'
Corte
5'
CATG 3'
3'GTAG
5'
Transferasa terminal
5'
CATGAAAAAAAAAAA···AAA 3'
3' AAA···AAAAAAAAAAGTAC
5'
Fusión, adición de cebador, apareamiento
cadenas derivadas de orden 0
5'
CATGAAAAAAAAAA·····AAA 3'
GTACTTTTTTTTTT-N10-PrA
Polimerasa
cadena derivada de 1er orden
5'
3'
CATGAAAAAAAAAA·····AAA 3'
GTACTTTTTTTTTT-N10-PrA
Combinación, limpieza, fusión, transferasa terminal
3' CCC···CCCCCCCCCC
GTACTTTTTTTTTT-N10-PrA
Adición de cebador, apareamiento
5' PrB-GGGGGGGGG
3' CCC·····CCCCCCCCC
GTACTTTTTTTTTT-N10-PrA
Polimerasa
cadena derivada de 2º orden
5' PrB-GGGGGGGGG
3' CCC·····CCCCCCCCC
GTACTTTTTTTTTT-N10-PrA
Adición de PrA, PrB
PCR
Secuencia

Fig. 8

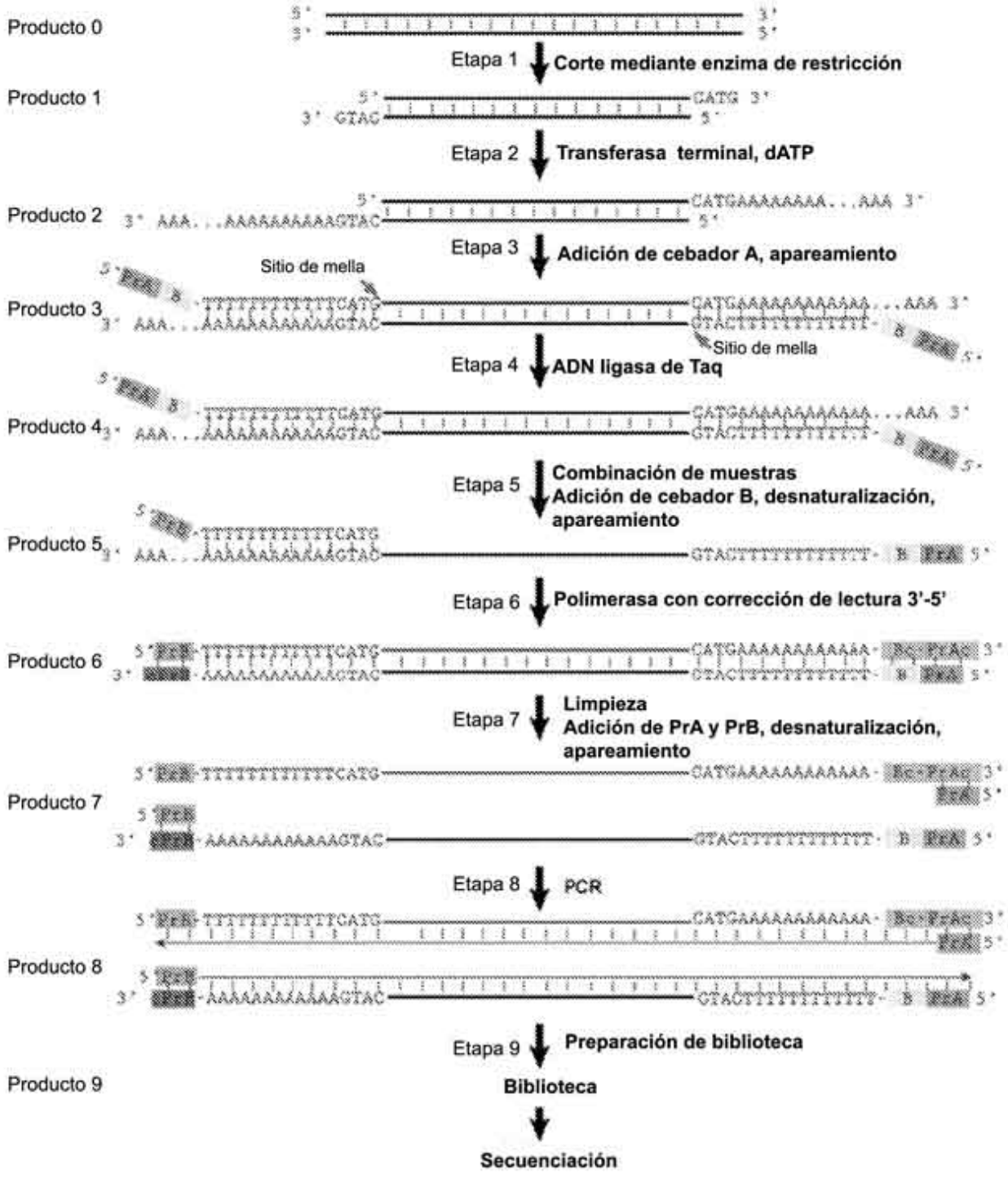
Producto 0
Etapa 1
Corte mediante enzima de restricción
Producto 1
Etapa 2
Transferasa terminal, dATP
Producto 2
Etapa 3
Adición de cebador A, apareamiento
Sitio de mella
Producto 3
Sitio de mella
Etapa 4
ADN ligasa de Taq
Producto 4
Etapa 5
Combinación de muestras
Adición de cebador B, desnaturalización, apareamiento
Producto 5
Etapa 6
Polimerasa con corrección de lectura 3'-5'
Producto 6
Etapa 7
Limpieza
Adición de PrA y PrB, desnaturalización, apareamiento
Producto 7
Etapa 8
PCR
Producto 8
Etapa 9
Preparación de biblioteca
Producto 9
Biblioteca
Secuenciación

Fig. 9

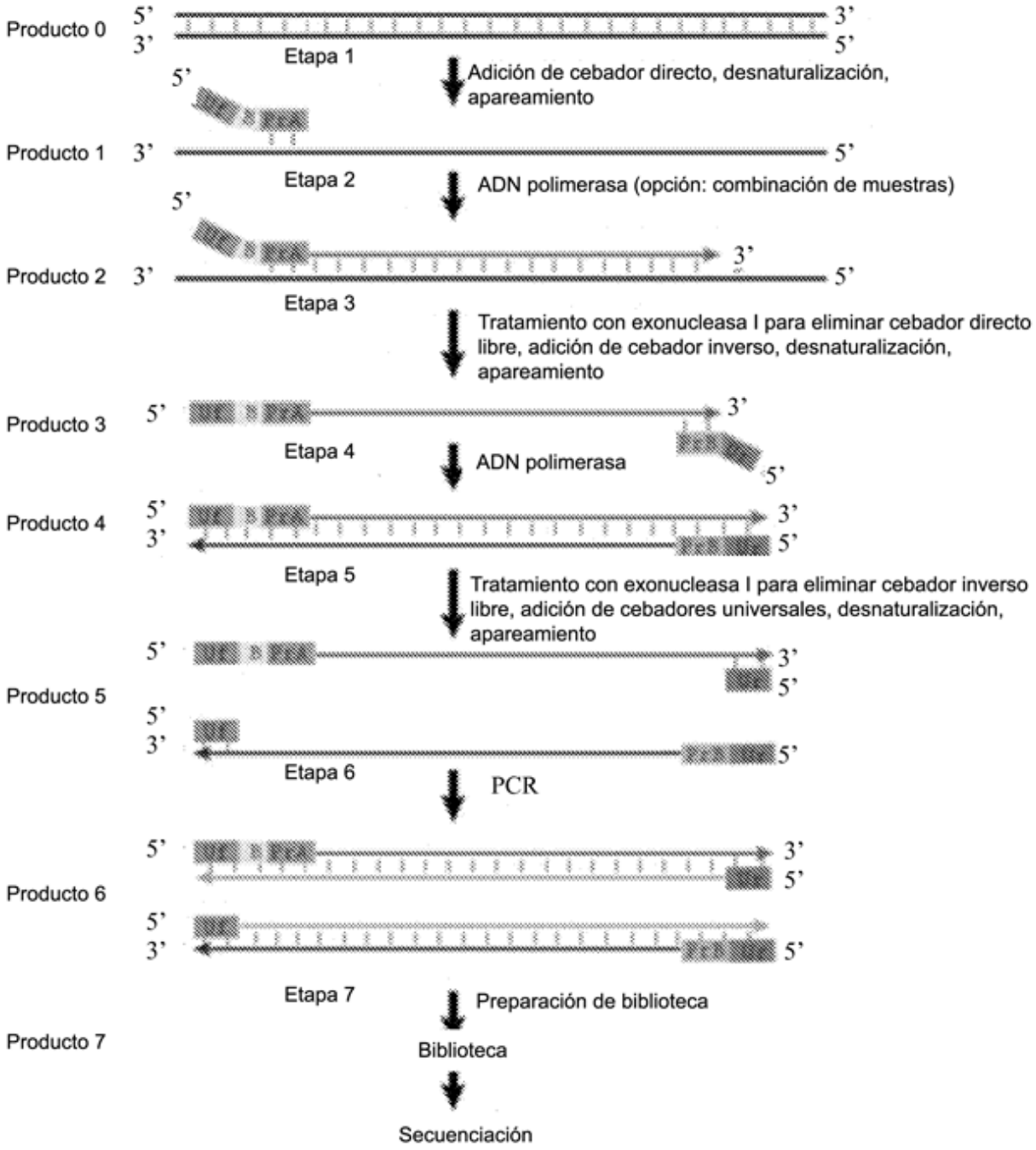
Producto 0
Etapa 1
Adición de cebador directo, desnaturalización, apareamiento
Producto 1
Etapa 2
ADN polimerasa (opción: combinación de muestras)
Producto 2
Etapa 3
Tratamiento con exonucleasa I para eliminar cebador directo libre, adición de cebador inverso, desnaturalización, apareamiento
Producto 3
Etapa 4
ADN polimerasa
Producto 4
Etapa 5
Tratamiento con exonucleasa I para eliminar cebador inverso libre, adición de cebadores universales, desnaturalización, apareamiento
Producto 5
Etapa 6
PCR
Producto 6
Etapa 7
Preparación de biblioteca
Producto 7
Biblioteca
Secuenciación

Fig. 10

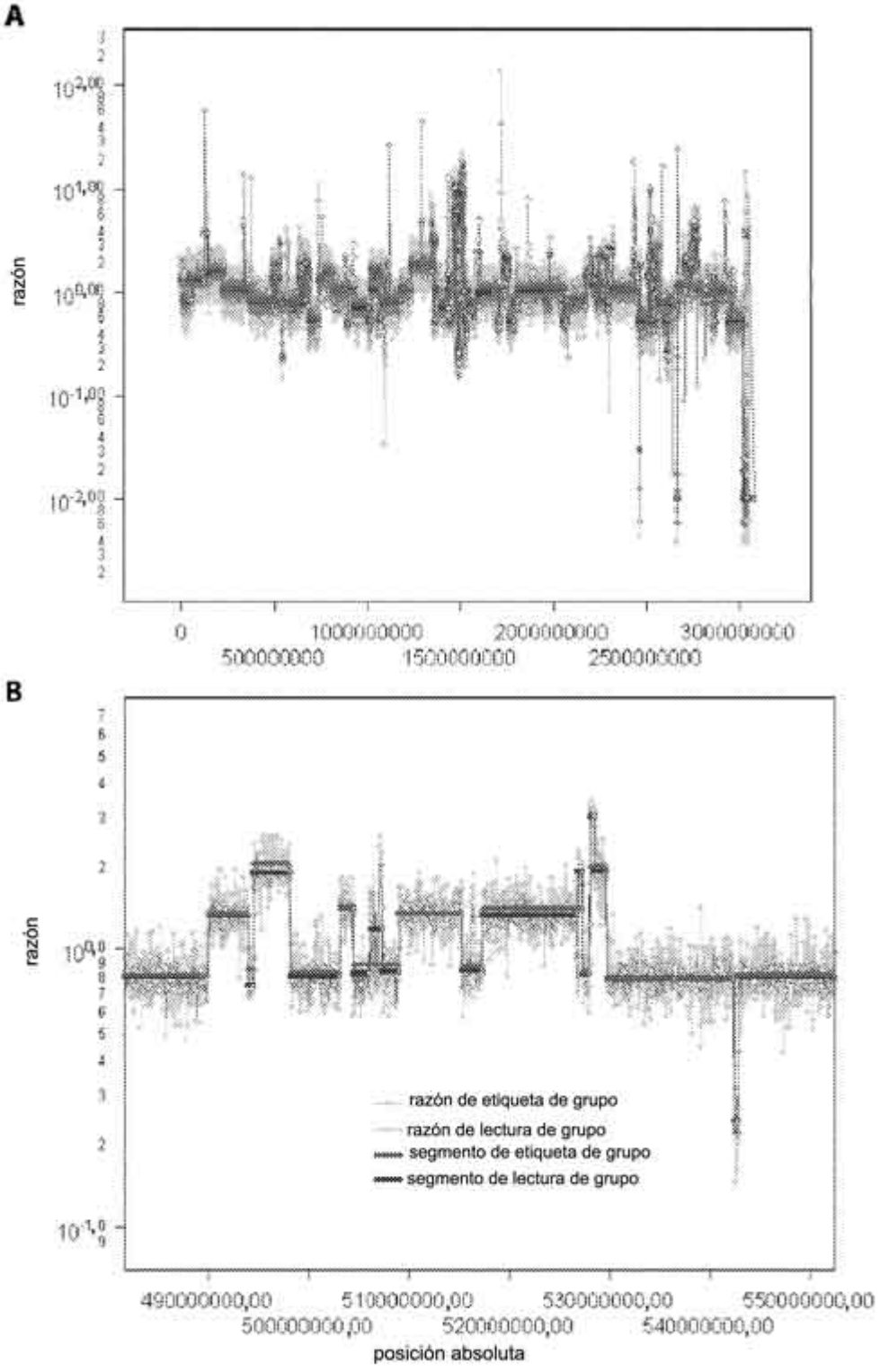
A
razón
0
500000000
1000000000
1500000000
2000000000
2500000000
3000000000
B
razón
razón de etiqueta de grupo
razón de lectura de grupo
segmento de etiqueta de grupo
segmento de lectura de grupo
posición absoluta

Fig. 11