(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 985 191**

(51) Int. Cl.:

G16B 40/10 (2009.01)
G16B 20/10 (2009.01)
G16B 20/20 (2009.01)
G16B 40/20 (2009.01)
G16B 20/30 (2009.01)
G16B 20/00 (2009.01)
G16B 30/00 (2009.01)
C12Q 1/6869 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **17.08.2020 PCT/CN2020/109602**

(87) Fecha y número de publicación internacional: **25.02.2021 WO21032060**

(96) Fecha de presentación y número de la solicitud europea: **17.08.2020 E 20853612 (8)**

(97) Fecha y número de publicación de la concesión europea: **31.01.2024 EP 3827092**

(54) Título: **Detección de metilación de nucleótidos en ácidos nucleicos**

(30) Prioridad:

16.08.2019 US 201962887987 P
05.02.2020 US 202062970586 P
19.03.2020 US 202062991891 P
04.05.2020 US 202063019790 P
13.07.2020 US 202063051210 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
04.11.2024

(73) Titular/es:

THE CHINESE UNIVERSITY OF HONG KONG (100.0%)
Office of Research and Knowledge Transfer Services, (ORKTS) Room 301 Pi Ch'iu Building, Shatin New Territories
Hong Kong, CN

(72) Inventor/es:

LO, YUK-MING DENNIS;
CHIU, ROSSA WAI KWUN;
CHAN, KWAN CHEE;
JIANG, PEIYONG;
CHENG, SUK HANG;
PENG, WENLEI y
TSE, ON YEE

(74) Agente/Representante:

MENDIGUTÍA GÓMEZ, María Manuela

DESCRIPCIÓN

Detección de metilación de nucleótidos en ácidos nucleicos

Antecedentes

La existencia de modificaciones de bases en ácidos nucleicos varía en diferentes organismos que incluyen virus, bacterias, plantas, hongos, nematodos, insectos y vertebrados (por ejemplo, humanos), etc. Las modificaciones de bases más comunes son la adición de un grupo metilo a diferentes bases de ADN en diferentes posiciones, lo que se denomina metilación. Se ha encontrado metilación en citosinas, adeninas, timinas y guaninas, tales como 5mC (5-metilcitosina), 4mC (N4-metilcitosina), 5hmC (5-hidroximetilcitosina), SfC (5-formilcitosina), ScaC (5-carboxilcitosina), 1mA (N1-metiladenina), 3mA (N3-metiladenina), 7mA (N7-metiladenina), 3mC (N3-metilcitosina), 2mG (N2-metilguanina), 6mG (06-metilguanina), 7mG (N7-metilguanina), 3mT (N3-metiltimina), y 4mT (O4-metiltimina). En los genomas de vertebrados, 5mC es el tipo más común de metilación de bases, seguido por el de guanina (es decir, en el contexto CpG).

La metilación del ADN es esencial para el desarrollo de los mamíferos y tiene funciones notables en la expresión y silenciamiento de genes, desarrollo embrionario, transcripción, estructura de la cromatina, inactivación del cromosoma X, protección contra la actividad de los elementos repetitivos, mantenimiento de la estabilidad genómica durante la mitosis y regulación de la impronta genómica del progenitor de origen.

La metilación del ADN desempeña muchas funciones importantes en el silenciamiento de promotores y potenciadores de manera coordinada (Robertson, 2005; Smith and Meissner, 2013). Se ha encontrado que muchas enfermedades humanas se asocian con aberraciones de la metilación del ADN, que incluyen, pero no se limitan a, el proceso de carcinogénesis, trastornos de impronta (por ejemplo síndrome de Beckwith-Wiedemann y síndrome de Prader-Willi), enfermedades de inestabilidad repetida (por ejemplo síndrome de X frágil), trastornos autoinmunitarios (por ejemplo lupus eritematoso sistémico), trastornos metabólicos (por ejemplo diabetes tipo I y tipo II), trastornos neurológicos, envejecimiento, etc.

La medición precisa de la modificación metilómica en moléculas de ADN tendría numerosas implicaciones clínicas. Un método ampliamente utilizado para medir la metilación del ADN es a través del uso de secuenciación con bisulfito (BS-seq) (Lister et al., 2009; Frommer et al., 1992). En este enfoque, las muestras de ADN se tratan primero con bisulfito, que convierte la citosina no metilada (es decir, C) a uracilo. Por el contrario, la citosina metilada permanece sin cambios. Luego, el ADN modificado con bisulfito se analiza mediante secuenciación del ADN. En otro enfoque, después de la conversión con bisulfito, el ADN modificado se somete a amplificación por reacción en cadena de la polimerasa (PCR) utilizando cebadores que pueden diferenciar el ADN convertido con bisulfito de diferentes perfiles de metilación (Herman et al., 1996). Este último enfoque se denomina PCR específica de metilación.

Una desventaja de dichos enfoques basados en bisulfito es que se ha informado que la etapa de conversión con bisulfito degrada significativamente la mayoría del ADN tratado (Grunau, 2001). Otra desventaja es que la etapa de conversión de bisulfito crearía fuertes sesgos de CG (Olova et al., 2018), lo que resultaría en la reducción de las relaciones señal-ruido normalmente para mezclas de ADN con estados de metilación heterogéneos. Además, la secuenciación con bisulfito no podría secuenciar moléculas de ADN largas debido a la degradación del ADN durante el tratamiento con bisulfito. Por tanto, subsiste la necesidad de determinar la modificación de las bases de los ácidos nucleicos, sin amplificación química previa (por ejemplo, conversión con bisulfito) ni de ácidos nucleicos (por ejemplo, utilizando la PCR).

Breve resumen

Hemos desarrollado un nuevo método que permite la detección de una metilación de un nucleótido en una molécula de ácido nucleico, como se establece en la reivindicación independiente 1 adjunta. Las características preferidas se establecen en las reivindicaciones dependientes adjuntas.

El método se puede utilizar para la detección de metilación, tal como 5mC en ácidos nucleicos sin pretratamiento de ADN de plantilla, tales como conversiones enzimáticas y/o químicas, o unión a proteínas y/o anticuerpos. Si bien dicho pretratamiento del ADN de plantilla no es necesario para la determinación de las metilaciones, en los ejemplos que se muestran, cierto pretratamiento (por ejemplo, digestión con enzimas de restricción) puede servir para potenciar aspectos de la invención (por ejemplo, permitir el enriquecimiento de sitios CpG para analizar). Las realizaciones presentes en esta divulgación se podrían utilizar para detectar diferentes tipos de modificación de bases, por ejemplo, que incluyen, pero no se limitan a, 4mC, 5hmC, SfC y ScaC, 1mA, 3mA, 7mA, 3mC, 2mG, 6mG, 7mG, 3mT. y 4mT, etc. Dichas realizaciones pueden hacer uso de características derivadas de la secuenciación, tales como características cinéticas, que se ven afectadas por las diversas metilaciones, así como una identidad de nucleótidos en una ventana alrededor de una posición diana cuyo estado de metilación se determina.

Las realizaciones de la presente invención se pueden utilizar para, pero no se limitan a, la secuenciación de única molécula. Un tipo de secuenciación de única molécula es la secuenciación en tiempo real de única molécula en la que el progreso de la secuenciación de única molécula de ADN se monitoriza en tiempo real. Un tipo de secuenciación en tiempo real de única molécula es la comercializada por Pacific Biosciences utilizando su sistema de Molécula Única

en Tiempo Real (SMRT). Los métodos pueden utilizar la anchura de pulso de una señal de las bases de secuenciación, la duración interpulso (IPD) de las bases y la identidad de las bases para detectar una metilación en una base o en una base vecina. Otro sistema de una única molécula es el basado en la secuenciación de nanoporos. Un ejemplo de sistema de secuenciación de nanoporos es el comercializado por Oxford Nanopore Technologies.

Los métodos que hemos desarrollado pueden servir como herramientas para detectar metilaciones en muestras biológicas para evaluar los perfiles de metilación en las muestras para diversos fines que incluyen, pero no se limitan a, fines de investigación y diagnóstico. Los perfiles de metilación detectados se pueden utilizar para diferentes análisis. Los perfiles de metilación se pueden utilizar para detectar el origen del ADN (por ejemplo, materno o fetal, de tejido, bacteriano o ADN obtenido de células tumorales enriquecidas a partir de la sangre de un paciente con cáncer). La detección de perfiles de metilación aberrantes en tejidos ayuda a identificar trastornos del desarrollo en individuos, identificar y pronosticar tumores o neoplasias malignas.

También describimos métodos que pueden incluir el análisis de los niveles relativos de metilación de haplotipos de un organismo. Se puede utilizar un desequilibrio en los niveles de metilación entre los dos haplotipos para determinar la clasificación de un trastorno. Un desequilibrio mayor puede indicar la presencia de un trastorno o un trastorno más grave. El trastorno puede incluir cáncer.

Los patrones de metilación en una única molécula pueden identificar ADN quimérico y híbrido. Las moléculas quiméricas e híbridas pueden incluir secuencias de dos genes, cromosomas, orgánulos (por ejemplo, mitocondrias, núcleos, cloroplastos), organismos (mamíferos, bacterias, virus, etc.) y/o especies diferentes. La detección de uniones de moléculas de ADN quiméricas o híbridas puede permitir detectar fusiones genéticas para diversos trastornos o enfermedades, que incluyen cáncer, trastornos prenatales o congénitos.

Se puede obtener una mejor comprensión de la naturaleza y las ventajas de las realizaciones de la presente invención con referencia a la siguiente descripción detallada y a los dibujos acompañantes.

Breve descripción de los dibujos

La FIG. 1 ilustra la secuenciación SMRT de moléculas que llevan modificaciones de bases.

La FIG. 2 ilustra la secuenciación SMRT de moléculas que llevan sitios CpG metilados y no metilados.

La FIG. 3 ilustra las duraciones entre interpulsos y la anchura del pulso.

La FIG. 4 muestra un ejemplo de una ventana de medición de la hebra de Watson de ADN para detectar una modificación de base de acuerdo con realizaciones de la presente invención.

La FIG. 5 muestra un ejemplo de una ventana de medición de la hebra de ADN de Crick para detectar una modificación de base.

La FIG. 6 muestra un ejemplo de una ventana de medición al combinar datos de la hebra de ADN de Watson y su hebra de Crick complementaria para detectar cualquier modificación de base.

La FIG. 7 muestra un ejemplo de una ventana de medición al combinar datos de la hebra de ADN de Watson y la hebra de Crick de su región cercana para detectar cualquier modificación de base.

La FIG. 8 muestra ejemplos de ventanas de medición de la hebra de Watson, la hebra de Crick y ambas hebras para determinar los estados de metilación en los sitios CpG.

La FIG. 9 muestra un procedimiento general de construcción de modelos analíticos, computacionales, matemáticos o estadísticos para clasificar modificaciones de bases.

La FIG. 10 muestra un procedimiento general para clasificar modificaciones de bases.

La FIG. 11 muestra un procedimiento general para construir modelos analíticos, computacionales, matemáticos o estadísticos para clasificar estados de metilación en sitios CpG utilizando muestras con estados de metilación conocidos de la hebra de Watson.

La FIG. 12 muestra un procedimiento general para clasificar los estados de metilación de la hebra de Watson para una muestra desconocida.

La FIG. 13 muestra un procedimiento general para construir los modelos analíticos, computacionales, matemáticos o estadísticos para clasificar estados de metilación en sitios CpG utilizando muestras con estados de metilación conocidos de la hebra de Crick.

La FIG. 14 muestra un procedimiento general para clasificar los estados de metilación de la hebra de Crick para una muestra desconocida.

La FIG. 15 muestra un procedimiento general de construcción de modelos estadísticos para clasificar estados de metilación en sitios CpG utilizando muestras con estados de metilación conocidos de las hebras tanto de Watson como de Crick.

La FIG. 16 muestra un procedimiento general para clasificar los estados de metilación de una muestra desconocida de las hebras de Watson y Crick.

Las FIG. 17A y 17B muestran el rendimiento de un conjunto de datos de entrenamiento y un conjunto de datos de prueba para determinar la metilación.

La FIG. 18 muestra el rendimiento de un conjunto de datos de entrenamiento y un conjunto de datos de prueba para determinar la metilación.

La FIG. 19 muestra el rendimiento de un conjunto de datos de entrenamiento y un conjunto de datos de prueba a diferentes profundidades de secuenciación para determinar la metilación.

La FIG. 20 muestra el rendimiento de un conjunto de datos de entrenamiento y un conjunto de datos de prueba para diferentes hebras para determinar la metilación.

La FIG. 21 muestra el rendimiento de un conjunto de datos de entrenamiento y un conjunto de datos de prueba para diferentes ventanas de medición para determinar la metilación.

La FIG. 22 muestra el rendimiento de un conjunto de datos de entrenamiento y un conjunto de datos de prueba para diferentes ventanas de medición utilizando bases en dirección descendente solo para determinar la metilación.

La FIG. 23 muestra el rendimiento de un conjunto de datos de entrenamiento y un conjunto de datos de prueba para diferentes ventanas de medición utilizando bases en dirección ascendente solo para determinar la metilación.

La FIG. 24 muestra el rendimiento del análisis de metilación utilizando patrones cinéticos asociados con bases en dirección descendente y en dirección ascendente utilizando tamaños de flanqueo asimétricos en el conjunto de datos de entrenamiento.

La FIG. 25 muestra el rendimiento del análisis de metilación utilizando patrones cinéticos asociados con bases en dirección descendente y en dirección ascendente utilizando tamaños de flanqueo asimétricos en el conjunto de datos de prueba.

La FIG. 26 muestra la importancia relativa de las características con respecto a la clasificación de los estados de metilación en los sitios CpG.

La FIG. 27 muestra el rendimiento del análisis IPD basado en motivos para la detección de metilación sin utilizar la señal de anchura de pulso.

La FIG. 28 es un gráfico de una técnica de análisis de componentes principales que utiliza 2 nt en dirección ascendente y 6 nt en dirección descendente de una citosina que se somete a análisis de metilación.

La FIG. 29 es un gráfico de una comparación de rendimiento entre un método que utiliza análisis de componentes principales y un método que utiliza una red neuronal convolucional.

La FIG. 30 muestra el rendimiento de un conjunto de datos de entrenamiento y un conjunto de datos de prueba para diferentes modelos analíticos, computacionales, matemáticos o estadísticos que utilizan bases en dirección ascendente solo para determinar la metilación.

La FIG. 31A muestra un ejemplo de un enfoque para generar moléculas con adeninas no metiladas mediante amplificación del genoma completo.

La FIG. 31B muestra un ejemplo de un enfoque para generar moléculas con adeninas metiladas mediante amplificación del genoma completo.

Las FIG. 32A y 32B muestran valores de duración de interpulso (IPD) a través de bases A secuenciadas en el ADN de plantilla de la hebra de Watson entre conjuntos de datos metilados y no metilados.

La FIG. 32C muestra una curva característica operativa del receptor para determinar la metilación en la hebra de Watson.

Las FIG. 33A y 33B muestran valores de duración interpulso (IPD) a través de bases A secuenciadas en el ADN de plantilla de la hebra de Crick entre conjuntos de datos metilados y no metilados.

La FIG. 33C muestra una curva característica operativa del receptor para determinar la metilación en la hebra de Crick.

La FIG. 34 ilustra la determinación de 6mA de la hebra de Watson.

La FIG. 35 ilustra la determinación de 6mA de la hebra de Crick.

La FIG. 36A y FIG. 36B muestra la probabilidad determinada de ser metilado para bases A secuenciadas de la hebra de Watson entre conjuntos de datos de uA y mA utilizando un modelo de red neuronal convolucional basado en ventana de medición.

La FIG. 37 muestra una curva ROC para la detección de 6mA utilizando un modelo CNN basado en ventana de medición para bases A secuenciadas de la hebra de Watson.

La FIG. 38 muestra una comparación de rendimiento entre la detección de 6mA basada en métrica IPD y una detección de 6mA basada en una ventana de medición.

Las FIG. 39A y 39B muestran la probabilidad determinada de ser metilados para aquellas bases A secuenciadas de la hebra de Crick entre conjuntos de datos de uA y mA utilizando el modelo CNN basado en ventana de medición.

La FIG. 40 muestra el rendimiento de la detección de 6mA utilizando el modelo CNN basado en ventana de medición en bases A secuenciadas de la hebra de Crick.

La FIG. 41 muestra ejemplos de estados de metilación en bases A en una molécula que incluye las hebras de Watson y Crick.

La FIG. 42 muestra un ejemplo de entrenamiento mejorado al utilizar selectivamente bases A en un conjunto de datos de mA con valores de IPD superiores a su percentil 10.

La FIG. 43 es un gráfico de los porcentajes de adeninas no metiladas en el conjunto de datos de mA frente al número de sublecturas en cada pocillo.

La FIG. 44 muestra patrones de metiladenina entre las hebras de Watson y Crick de una molécula de ADN de hebra doble en un conjunto de datos de prueba.

La FIG. 45 es una tabla que muestra el porcentaje de moléculas completamente no metiladas, moléculas hemimetiladas, moléculas completamente metiladas y moléculas con patrones de metiladenina entrelazados en conjuntos de datos de entrenamiento y prueba.

La FIG. 46 ilustra ejemplos representativos de moléculas con moléculas completamente no metiladas con respecto a sitios de adenina, moléculas hemimetiladas, moléculas completamente metiladas y moléculas con patrones de metiladenina entrelazados.

La FIG. 47 muestra un ejemplo de una lectura larga (6,265 bp) que alberga una isla CpG (que se sombrea en amarillo).

La FIG. 48 es una tabla que muestra que las 9 moléculas de ADN se secuenciaron mediante secuenciación SMRT de Pacific Biosciences y que se superpusieron con regiones impresas.

La FIG. 49 muestra un ejemplo de una impronta genómica.

La FIG. 50 muestra un ejemplo para la determinación de patrones de metilación en una región impresa.

La FIG. 51 muestra una comparación de los niveles de metilación deducidos entre el nuevo enfoque y la secuenciación con bisulfito convencional.

La FIG. 52 muestra el rendimiento de la detección de la metilación del ADN plasmático. (A) La relación entre la probabilidad prevista de metilación versus los rangos de niveles de metilación cuantificados mediante secuenciación con bisulfito. (B) La correlación entre los niveles de metilación determinados mediante secuenciación de Pacific Biosciences (PacBio) de acuerdo con las realizaciones presentes en esta divulgación (eje y) y los niveles de metilación cuantificados mediante secuenciación con bisulfito (eje x) en una resolución de 10 Mb.

La FIG. 53 muestra una correlación de la presentación genómica (GR) del cromosoma Y entre la secuenciación SMRT de Pacific Biosciences y BS-seq.

La FIG. 54 muestra un ejemplo de detección de metilación basada en bloques CpG utilizando bloques CpG, cada uno de los cuales alberga una serie de sitios CpG. SmC: metilación; C: sin metilación.

La FIG. 55 muestra el entrenamiento y las pruebas de metilación que requieren moléculas de ADN humano utilizando el enfoque basado en bloques CpG. (A) Rendimiento en el conjunto de datos de entrenamiento. (B) Rendimiento en un conjunto de datos de prueba independiente.

Las FIG. 56A y 56B muestran cambios en el número de copias en tejido tumoral.

Las FIG. 57A y 57B muestran cambios en el número de copias en tejido tumoral.

La FIG. 58 muestra una ilustración esquemática del mapeo de tejido del ADN plasmático del plasma de una mujer embarazada utilizando los niveles de metilación deducidos.

La FIG. 59 muestra una correlación entre la contribución placentaria al ADN del plasma materno deducida y la fracción de ADN fetal deducida por las lecturas del cromosoma Y.

La FIG. 60 muestra una tabla que resume los datos de secuenciación de diferentes muestras de ADN de tejido humano.

La FIG. 61 muestra una ilustración de varias formas de analizar patrones de metilación.

Las FIG. 62A y 62B muestran una comparación de las densidades de metilación a nivel de genoma completo cuantificadas mediante secuenciación con bisulfito y secuenciación en tiempo real de única molécula.

Las FIG. 63A, 63B y 63C muestran diferentes correlaciones de los niveles generales de metilación cuantificados mediante secuenciación con bisulfito y secuenciación en tiempo real de única molécula.

Las FIG. 64A y 64B muestran patrones de metilación con una resolución de 1 Mnt para una estirpe celular de carcinoma hepatocelular (HCC) y una muestra de capa leucocitaria de un sujeto de control sano con niveles de metilación determinados mediante secuenciación con bisulfito y mediante secuenciación en tiempo real de única molécula.

Las FIG. 65A y 65B muestran diagramas de dispersión de niveles de metilación con una resolución de 1 Mnt determinada mediante secuenciación con bisulfito y secuenciación en tiempo real de única molécula para una estirpe celular HCC (HepG2) y una muestra de capa leucocitaria de un sujeto de control sano.

Las FIG. 66A y 66B muestran diagramas de dispersión de niveles de metilación con una resolución de 100 Mnt determinada mediante secuenciación con bisulfito y secuenciación en tiempo real de única molécula para una estirpe celular HCC (HepG2) y una muestra de capa leucocitaria de un sujeto de control sano.

Las FIG. 67A y 67B muestran patrones de metilación con una resolución de 1 Mnt para un tejido tumoral de HCC y tejido normal adyacente con niveles de metilación determinados mediante secuenciación con bisulfito y mediante secuenciación en tiempo real de única molécula.

Las FIG. 68A y 68B muestran diagramas de dispersión de niveles de metilación con una resolución de 1 Mnt determinada mediante secuenciación con bisulfito y secuenciación en tiempo real de única molécula para tejido tumoral de HCC y tejido normal adyacente.

Las FIG. 69A y 69B muestran diagramas de dispersión de niveles de metilación con una resolución de 100 knt determinados mediante secuenciación con bisulfito y secuenciación en tiempo real de única molécula para tejido tumoral de HCC y tejido normal adyacente.

Las FIG. 70A y 70B muestran patrones de metilación con una resolución de 1 Mnt para un tejido tumoral de HCC y tejido normal adyacente con niveles de metilación determinados mediante secuenciación con bisulfito y mediante secuenciación en tiempo real de única molécula.

Las FIG. 71A y 71B muestran diagramas de dispersión de niveles de metilación con una resolución de 1 Mnt determinada mediante secuenciación con bisulfito y secuenciación en tiempo real de única molécula para tejido tumoral de HCC y tejido normal adyacente.

Las FIG. 72A y 72B muestran diagramas de dispersión de niveles de metilación con una resolución de 100 knt determinados mediante secuenciación con bisulfito y secuenciación en tiempo real de única molécula para tejido tumoral de HCC y tejido normal adyacente.

La FIG. 73 muestra un ejemplo del patrón aberrante de metilación cerca del gen supresor de tumores CDKN2A.

Las FIG. 74A y 74B muestran regiones de metilación diferencial detectadas mediante secuenciación en tiempo real de única molécula.

La FIG. 75 muestra patrones de metilación del ADN del virus de la hepatitis B entre tejidos de HCC y tejidos no tumorales adyacentes utilizando secuenciación en tiempo real de única molécula.

La FIG. 76A muestra los niveles de metilación del ADN del virus de la hepatitis B en tejidos hepáticos de pacientes con cirrosis pero sin HCC utilizando secuenciación con bisulfito.

La FIG. 76B muestra niveles de metilación del ADN del virus de la hepatitis B en tejidos de HCC utilizando secuenciación con bisulfito.

La FIG. 77 ilustra el análisis de haplotipos de metilación.

La FIG. 78 muestra la distribución de tamaño de las moléculas secuenciadas determinadas a partir de secuencias consenso.

Las FIG. 79A, 79B, 79C y 79D muestran ejemplos de patrones de metilación alélica en las regiones impresas.

Las FIG. 80A, 80B, 80C y 80D muestran ejemplos de patrones de metilación alélica en regiones no impresas.

La FIG. 81 muestra una tabla de niveles de metilación de fragmentos específicos de alelo.

La FIG. 82 muestra un ejemplo para determinar el origen placentario del ADN plasmático durante el embarazo utilizando perfiles de metilación.

La FIG. 83 ilustra el análisis de metilación del ADN específico de fetos.

Las FIG. 84A, 84B y 84C muestran el rendimiento de diferentes tamaños de ventanas de medición en diferentes kits de reactivos para SMRT-seq.

Las FIG. 85A, 85B y 85C muestran el rendimiento de diferentes tamaños de ventanas de medición en diferentes kits de reactivos para SMRT-seq.

Las FIG. 86A, 86B y 86C muestran la correlación de los niveles de metilación generales cuantificados mediante secuenciación con bisulfito y SMRT-seq (Kit de Secuenciación Sequel II 2.0).

Las FIG. 87A y 87B muestran una comparación del nivel de metilación global entre diversos tejidos tumorales y tejidos no tumorales adyacentes emparejados.

La FIG. 88 muestra la determinación del estado de metilación utilizando un contexto de secuencia determinado a partir de una secuencia consenso circular (CCS).

La FIG. 89 muestra una curva ROC para la detección de sitios CpG metilados utilizando un contexto de secuencia determinado a partir de CCS.

La FIG. 90 muestra una curva ROC para la detección de sitios CpG metilados sin información DE CCS y sin alineación previa con un genoma de referencia.

La FIG. 91 muestra un ejemplo de preparación de moléculas para secuenciación en tiempo real de única molécula.

La FIG. 92 muestra una ilustración del sistema CRISPR/Cas9.

La FIG. 93 muestra un ejemplo de un complejo Cas9 para introducir dos cortes que abarcan una molécula de interés bloqueada en el extremo.

La FIG. 94 muestra la distribución de metilación de regiones Alu determinada mediante secuenciación con bisulfito y secuenciación en tiempo real de única molécula.

La FIG. 95 muestra la distribución de los niveles de metilación de las regiones Alu determinadas por el modelo utilizando resultados de secuenciación en tiempo real de única molécula.

La FIG. 96 muestra una tabla de tejidos y los niveles de metilación de regiones Alu en los tejidos.

La FIG. 97 muestra un análisis de agrupamiento para diferentes tipos de cáncer utilizando señales de metilación relacionadas con repeticiones de Alu.

La FIG. 98A y 98B muestran el efecto de la profundidad de lectura en la cuantificación general del nivel de metilación en los conjuntos de datos de prueba que estuvieron involucrados con la amplificación del genoma completo y el tratamiento con M.Ssssl.

La FIG. 99 muestra una comparación entre los niveles de metilación generales determinados por SMRT-seq (Kit de Secuenciación Sequel II 2.0) y BS-seq con el uso de diferentes valores de corte de profundidad de sublectura.

La FIG. 100 es una tabla que muestra el efecto de la profundidad de la sublectura en la correlación de los niveles de metilación entre dos mediciones mediante SMRT-seq (Kit de Secuenciación Sequel II 2.0) y BS-seq.

La FIG. 101 muestra la distribución de profundidad de sublectura con respecto a los tamaños de fragmentos en los datos generados por el Kit de Secuenciación Sequel II 2.0.

La FIG. 102 muestra un método para detectar una modificación de un nucleótido en una molécula de ácido nucleico.

La FIG. 103 muestra un método para detectar una modificación de un nucleótido en una molécula de ácido nucleico.

La FIG. 104 ilustra el análisis de desequilibrio de metilación relativo basado en haplotipos.

Las FIG. 105A y 105B son una tabla de los bloques de haplotipos que muestran niveles de metilación diferenciales entre Hap I y Hap II en el ADN tumoral en comparación con el ADN del tejido no tumoral adyacente para el caso TBR3033.

La FIG. 106 es una tabla de los bloques de haplotipos que muestran niveles de metilación diferenciales entre Hap I y Hap II en el ADN del tumor en comparación con el ADN del tejido normal adyacente para el caso TBR3032.

La FIG. 107A es una tabla que resume el número de bloques de haplotipos que muestran un desequilibrio de metilación entre dos haplotipos entre tejidos tumorales y no tumorales adyacentes en base a los datos generados por el Kit de Secuenciación Sequel II 2.0.

La FIG. 107B es una tabla que resume el número de bloques de haplotipos que muestran un desequilibrio de metilación entre dos haplotipos en tejidos tumorales para diferentes estadios tumorales en base a los datos generados por el Kit de Secuenciación Sequel II 2.0.

La FIG. 108 ilustra el análisis de desequilibrio de metilación relativo basado en haplotipos.

La FIG. 109 muestra un método para clasificar un trastorno en un organismo que tiene un primer haplotipo y un segundo haplotipo.

La FIG. 110 ilustra la creación de fragmentos híbridos humano-ratón para los cuales la parte humana está metilada mientras que la parte del ratón no está metilada.

La FIG. 111 ilustra la creación de fragmentos híbridos humano-ratón para los cuales la parte humana no está metilada mientras que la parte del ratón está metilada.

La FIG. 112 muestra la distribución de longitud de las moléculas de ADN en una mezcla de ADN (muestra MIX01) después de la ligación.

La FIG. 113 ilustra una región de unión mediante la cual se unen un primer ADN (A) y un segundo ADN (B).

La FIG. 114 ilustra el análisis de metilación para la mezcla de ADN.

La FIG. 115 muestra un diagrama de caja de las probabilidades de ser metilado para sitios CpG en la muestra MIX01.

La FIG. 116 muestra la distribución de longitud de las moléculas de ADN en la mezcla de ADN después de ligación cruzada de la muestra MIX02.

La FIG. 117 muestra un diagrama de caja de las probabilidades de ser metilado para sitios CpG en la muestra MIX02.

La FIG. 118 es una tabla que compara la metilación determinada mediante secuenciación con bisulfito y secuenciación de Pacific Biosciences para MIX01.

La FIG. 119 es una tabla que compara la metilación determinada mediante secuenciación con bisulfito y secuenciación de Pacific Biosciences para MIX02.

Las FIG. 120A y 120B muestran niveles de metilación en intervalos de 5 Mb para ADN solo de humanos y de ratón para MIX01 y MIX02.

Las FIG. 121A y 121B muestran niveles de metilación en intervalos de 5 Mb para la parte humana y la parte de ratón de fragmentos de ADN híbridos humano-ratón para MIX01 y MIX02.

Las FIG. 122A y 122B son gráficos representativos que muestran estados de metilación en una única molécula híbrida humano-ratón.

La FIG. 123 muestra un método para detectar moléculas quiméricas en una muestra biológica.

La FIG. 124 ilustra un sistema de medición.

La FIG. 125 muestra un diagrama de bloques de un sistema informático de ejemplo utilizable con los sistemas y métodos descritos en el presente documento.

La FIG. 126 muestra una secuenciación en tiempo real de única molécula dirigida basada en MspI con el uso de reparación de extremos de ADN y cola A.

Las FIG. 127A y 127B muestran la distribución de tamaño de los fragmentos digeridos con MspI.

La FIG. 128 muestra una tabla con el número de moléculas de ADN para ciertos rangos de tamaño seleccionados.

La FIG. 129 es un gráfico del porcentaje de cobertura de sitios CpG dentro de islas CpG versus el tamaño de los fragmentos de ADN después de la digestión con enzimas de restricción.

La FIG. 130 muestra una secuenciación en tiempo real de única molécula dirigida basada en MspI sin el uso de reparación de extremos de ADN y cola A.

La FIG. 131 muestra una secuenciación en tiempo real de única molécula dirigida basada en MspI con una probabilidad reducida de autoligación del adaptador.

La FIG. 132 es un gráfico de los niveles generales de metilación entre muestras de ADN de placenta y leucocitos determinados mediante secuenciación en tiempo real de única molécula dirigida basada en MspI.

La FIG. 133 muestra un análisis de agrupamiento de muestras de placenta y capa leucocitaria utilizando sus perfiles de metilación del ADN determinados mediante secuenciación en tiempo real de única molécula dirigida basada en MspI.

Términos

Un “tejido” corresponde a un grupo de células que se agrupan formando una unidad funcional. Se puede encontrar más de un tipo de células en un solo tejido. Diferentes tipos de tejido pueden consistir en diferentes tipos de células (por ejemplo, hepatocitos, células alveolares o células sanguíneas), pero también pueden corresponder a tejido de diferentes organismos (madre versus feto; tejidos de un sujeto que ha recibido un trasplante; tejidos de un organismo que está infectado por un microorganismo o un virus) o a células sanas versus células tumorales. Los “tejidos de referencia” pueden corresponder a tejidos utilizados para determinar niveles de metilación específicos de tejido. Se pueden utilizar múltiples muestras de un mismo tipo de tejido de diferentes individuos para determinar un nivel de metilación específico de tejido para ese tipo de tejido.

Una “muestra biológica” se refiere a cualquier muestra que se toma de un sujeto humano. La muestra biológica puede ser una biopsia de tejido, un aspirado con aguja fina o células sanguíneas. La muestra también puede ser, por ejemplo, plasma o suero u orina de una mujer embarazada. También se pueden utilizar muestras de heces. En diversas realizaciones, la mayoría del ADN en una muestra biológica de una mujer embarazada que se ha enriquecido con ADN libre de células (por ejemplo, una muestra de plasma obtenida mediante un protocolo de centrifugación) puede estar libre de células, por ejemplo, más del 50 %, 60 %, 70 %, 80 %, 90 %, 95 % o 99 % del ADN pueden estar libres de células. El protocolo de centrifugación puede incluir, por ejemplo, 3.000 g x 10 minutos, obtener la parte fluida y volver a centrifugar a, por ejemplo, 30,000 g durante otros 10 minutos para eliminar las células residuales. En determinadas realizaciones, después de la etapa de centrifugación de 3.000 g, se puede seguir con la filtración de la parte fluida (por ejemplo, utilizando un filtro con un tamaño de poro de 5 µm o menor de diámetro).

Una “lectura de secuencia” se refiere a una cadena de nucleótidos secuenciados de cualquier parte o la totalidad de una molécula de ácido nucleico. Por ejemplo, una secuencia leída puede ser una cadena corta de nucleótidos (por ejemplo, 20-150) secuenciada a partir de un fragmento de ácido nucleico, una cadena corta de nucleótidos en uno o ambos extremos de un fragmento de ácido nucleico, o la secuenciación del fragmento de ácido nucleico entero que existe en la muestra biológica. Una lectura de secuencia se puede obtener de diversas maneras, por ejemplo, utilizando técnicas de secuenciación o utilizando sondas, por ejemplo, en matrices de hibridación o sondas de captura, o técnicas de amplificación, tales como la reacción en cadena de la polimerasa (PCR) o amplificación lineal utilizando un único cebador o amplificación isotérmica.

Una “sublectura” es una secuencia generada a partir de todas las bases en una hebra de una plantilla de ADN circularizado que se ha copiado en una hebra contigua por una ADN polimerasa. Por ejemplo, una sublectura puede corresponder a una hebra de ADN plantilla de ADN circularizado. En dicho ejemplo, después de la circularización, una molécula de ADN de hebra doble tendría dos sublecturas: una para cada paso de secuenciación. En algunas realizaciones, la secuencia generada puede incluir un subconjunto de todas las bases en una hebra, por ejemplo, debido a la existencia de errores de secuenciación.

Un “sitio” (también llamado “sitio genómico”) corresponde a un único sitio, que puede ser una única posición de base o un grupo de posiciones de bases correlacionadas, por ejemplo, un sitio CpG o un grupo más grande de posiciones de bases correlacionadas. Un “locus” puede corresponder a una región que incluye múltiples sitios. Un locus puede incluir sólo un sitio, lo que haría que el locus sea equivalente a un sitio en ese contexto.

Un “estado de metilación” se refiere al estado de metilación en un sitio determinado. Por ejemplo, un sitio puede estar metilado, no metilado o, en algunos casos, indeterminado.

El “índice de metilación” para cada sitio genómico (por ejemplo, un sitio CpG) se puede referir a la proporción de fragmentos de ADN (por ejemplo, determinados a partir de lecturas de secuencia o sondas) que muestran metilación en el sitio sobre el número total de lecturas que cubren ese sitio. Una “lectura” puede corresponder a información (por ejemplo, estado de metilación en un sitio) obtenida de un fragmento de ADN. Se puede obtener una lectura utilizando reactivos (por ejemplo, cebadores o sondas) que se hibridan preferentemente con fragmentos de ADN de un estado de metilación particular en uno o más sitios. Normalmente, dichos reactivos se aplican después del tratamiento con

un proceso que modifica o reconoce diferencialmente las moléculas de ADN dependiendo de su estado de metilación, por ejemplo, conversión de bisulfito, o enzima de restricción sensible a la metilación, o proteínas de unión a metilación, o anticuerpos anti-metilcitosina, o técnicas de secuenciación de única molécula (por ejemplo, secuenciación en tiempo real de única molécula y secuenciación de nanoporos (por ejemplo, de Oxford Nanopore Technologies)) que reconocen metilcitosinas y hidroximetilcitosinas.

La "densidad de metilación" de una región se puede referir al número de lecturas en sitios dentro de la región que muestran metilación dividido por el número total de lecturas que cubren los sitios en la región. Los sitios pueden tener características específicas, por ejemplo, ser sitios CpG. Por lo tanto, la "densidad de metilación de CpG" de una región se puede referir al número de lecturas que muestran metilación de CpG dividido por el número total de lecturas que cubren sitios CpG en la región (por ejemplo, un sitio CpG particular, sitios CpG dentro de una isla CpG, o una región más grande). Por ejemplo, la densidad de metilación para cada intervalo de 100 kb en el genoma humano se puede determinar a partir del número total de citosinas no convertidas después del tratamiento con bisulfito (que corresponde a la citosina metilada) en los sitios CpG como una proporción de todos los sitios CpG cubiertos por las lecturas de secuencia mapeadas a la región de 100 kb. Este análisis también se puede realizar para otros tamaños de intervalos, por ejemplo, 500 bp, 5 kb, 10 kb, 50 kb o 1 Mb, etc. Una región podría ser el genoma completo o un cromosoma o parte de un cromosoma (por ejemplo, un brazo cromosómico). El índice de metilación de un sitio CpG es el mismo que la densidad de metilación de una región cuando la región solo incluye ese sitio CpG. La "proporción de citosinas metiladas" se puede referir al número de sitios de citosina, "C's", que se muestra que están metilados (por ejemplo, sin convertir después de la conversión con bisulfito) sobre el número total de residuos de citosina analizados, es decir, que incluyen citosinas fuera del contexto CpG, en la región. El índice de metilación, la densidad de metilación, el recuento de moléculas metiladas en uno o más sitios y la proporción de moléculas metiladas (por ejemplo, citosinas) en uno o más sitios son ejemplos de "niveles de metilación". Además de la conversión de bisulfito, se pueden utilizar otros procesos conocidos por los expertos en la técnica para interrogar el estado de metilación de moléculas de ADN, que incluyen, pero no se limitan a, enzimas sensibles al estado de metilación (por ejemplo, enzimas de restricción sensibles a la metilación), proteínas de unión a metilación, secuenciación de única molécula utilizando una plataforma sensible al estado de metilación (por ejemplo, secuenciación de nanoporos (Schreiber et al. Proc Natl Acad Sci 2013; 110: 18910-18915) y mediante secuenciación en tiempo real de única molécula (por ejemplo, la de Pacific Biosciences) (Flusberg et al, Nat Methods 2010; 7: 461-465)).

Un "metiloma" proporciona una medida de una cantidad de metilación del ADN en una pluralidad de sitios o loci en un genoma. El metiloma puede corresponder a todo el genoma, a una parte sustancial del genoma o a porción(es) relativamente pequeñas del genoma.

Un "metiloma de plasma en gestante" es el metiloma determinado a partir del plasma o suero de un animal gestante (por ejemplo, un humano). El metiloma de plasma en gestante es un ejemplo de metiloma libre de células, ya que el plasma y el suero incluyen ADN libre de células. El metiloma del plasma en gestante también es un ejemplo de metiloma mixto, ya que es una mezcla de ADN de diferentes órganos, tejidos o células dentro de un cuerpo. En una realización, dichas células son células hematopoyéticas, que incluyen, pero no se limitan a, células del linaje eritroide (es decir, glóbulos rojos), el linaje mieloide (por ejemplo, neutrófilos y sus precursores) y el linaje megacariocítico. Durante el embarazo, el metiloma de plasma puede contener información metilómica del feto y de la madre. El "metiloma celular" corresponde al metiloma determinado a partir de células (por ejemplo, células sanguíneas) del paciente. El metiloma de las células sanguíneas se llama metiloma de las células sanguíneas (o metiloma sanguíneo).

Un "perfil de metilación" incluye información relacionada con la metilación del ADN o ARN para múltiples sitios o regiones. La información relacionada con la metilación del ADN puede incluir, pero o se limita a, un índice de metilación de un sitio CpG, una densidad de metilación (MD para abreviar) de sitios CpG en una región, una distribución de sitios CpG sobre una región contigua, un patrón o nivel de metilación para cada sitio CpG individual dentro de una región que contiene más de un sitio CpG, y metilación no CpG. En una realización, el perfil de metilación puede incluir el patrón de metilación o no metilación de más de un tipo de base (por ejemplo, citosina o adenina). Un perfil de metilación de una parte sustancial del genoma se puede considerar equivalente al metiloma. La "metilación del ADN" en genomas de mamíferos normalmente se refiere a la adición de un grupo metilo al carbono 5' de los residuos de citosina (es decir, 5-metilcitosinas) entre los dinucleótidos CpG. La metilación del ADN puede ocurrir en citosinas en otros contextos, por ejemplo CHG y CHH, donde H es adenina, citosina o timina. La metilación de la citosina también se puede realizar en forma de 5-hidroximetilcitosina. También se ha informado de metilaciones distintas de la citosina, como la N^6-metiladenina.

Un "patrón de metilación" se refiere al orden de las bases metiladas y no metiladas. Por ejemplo, el patrón de metilación puede ser el orden de las bases metiladas en una única hebra de ADN, una única molécula de ADN de hebra doble u otro tipo de molécula de ácido nucleico. Como un ejemplo, tres sitios CpG consecutivos pueden tener cualquiera de los siguientes patrones de metilación: UUU, MMM, LTMM, UMU, UUM, MLTM, MUU o MMU, donde "U" indica un sitio no metilado y "M" indica un sitio metilado. Cuando se extiende este concepto a modificaciones de bases que incluyen, pero no se restringen a, metilación, se utilizaría el término "patrón de modificación", que se refiere al orden de las bases modificadas y no modificadas. Por ejemplo, el patrón de modificación puede ser el orden de las bases modificadas en una única hebra de ADN, una única molécula de ADN de hebra doble u otro tipo de molécula de ácido nucleico. Como un ejemplo, tres sitios consecutivos potencialmente modificables pueden tener cualquiera de los siguientes patrones de modificación: UUU, MMM, LTMM, UMU, UUM, MLTM, MUU o MMU, donde "U" indica un sitio

no modificado y “M” indica un sitio modificado. Un ejemplo de modificación de bases que no se basa en la metilación son los cambios de oxidación, tal como en la 8-oxoguanina.

Los términos “hipermetilado” e “hipometilado” se pueden referir a la densidad de metilación de una única molécula de ADN medida por su nivel de metilación de única molécula, por ejemplo, el número de bases o nucleótidos metilados dentro de la molécula dividido por el número total de bases metilables o nucleótidos dentro de esa molécula. Una molécula hipermetilada es aquella en la que el nivel de metilación de única molécula está en o por encima de un umbral, que se puede definir desde la aplicación hasta la aplicación. El umbral puede ser 5 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 % o 95 %. Una molécula hipometilada es aquella en la que el nivel de metilación de una única molécula está en o por debajo de un umbral, que se puede definir desde la aplicación hasta aplicación, y que puede cambiar desde aplicación a aplicación. El umbral puede ser del 5 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 % o 95 %.

Los términos “hipermetilado” e “hipometilado” también se pueden referir al nivel de metilación de una población de moléculas de ADN según se mide por los niveles de metilación de múltiples moléculas de estas moléculas. Una población de moléculas hipermetiladas es aquella en la que el nivel de metilación de múltiples moléculas está en o por encima de un umbral que se puede definir desde la aplicación hasta la aplicación y que puede cambiar desde la aplicación hasta la aplicación. El umbral puede ser del 5 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 % o 95 %. Una población de moléculas hipometiladas es aquella en la que el nivel de metilación de múltiples moléculas está en o por debajo de un umbral que se puede definir desde la aplicación hasta la aplicación. El umbral puede ser del 5 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 % y 95 %. En una realización, la población de moléculas puede alinearse con una o más regiones genómicas seleccionadas. En una realización, la(s) región(es) genómica(s) seleccionada(s) puede(n) estar relacionada(s) con una enfermedad tales como cáncer, un trastorno genético, un trastorno de impronta, un trastorno metabólico o un trastorno neurológico. La(s) región(es) genómica(s) seleccionada(s) puede(n) tener una longitud de 50 nucleótidos (nt), 100 nt, 200 nt, 300 nt, 500 nt, 1000 nt, 2 knt, 5 knt, 10 knt, 20 knt, 30 knt, 40 knt, 50 knt, 60 knt, 70 knt, 80 knt, 90 knt, 100 knt, 200 knt, 300 knt, 400 knt, 500 knt o 1 Mnt.

El término “profundidad de secuenciación” se refiere al número de veces que un locus está cubierto por una secuencia leída alineada con el locus. El locus podría ser tan pequeño como un nucleótido, tan grande como un brazo cromosómico o tan grande como el genoma completo. La profundidad de secuenciación se puede expresar como 50x, 100x, etc., donde “x” se refiere al número de veces que se cubre un locus con una lectura de secuencia. La profundidad de secuenciación también se puede aplicar a múltiples loci, o al genoma completo, en cuyo caso x se puede referir al número medio de veces que se secuencian los loci o el genoma haploide, o el genoma completo, respectivamente. La secuenciación ultraprofunda se puede referir a al menos 100 veces la profundidad de secuenciación.

El término “clasificación” como se utiliza en el presente documento se refiere a cualesquier número(s) u otro(s) carácter(es) que estén asociados con una propiedad particular de una muestra. Por ejemplo, un símbolo “+” (o la palabra “positivo”) podría significar que una muestra está clasificada como con supresiones o amplificaciones. La clasificación puede ser binaria (por ejemplo, positiva o negativa) o tener más niveles de clasificación (por ejemplo, una escala desde 1 hasta 10 o desde 0 hasta 1).

Los términos “valor de corte” y “umbral” se refieren a números predeterminados utilizados en una operación. Por ejemplo, un tamaño de valor de corte se puede referir a un tamaño por encima del cual se excluyen los fragmentos. Un valor umbral puede ser un valor por encima o por debajo del cual se aplica una clasificación particular. Cualquiera de estos términos se puede utilizar en cualquiera de estos contextos. Un valor de corte o umbral puede ser “un valor de referencia” o derivarse de un valor de referencia que sea representativo de una clasificación particular o discrimine entre dos o más clasificaciones. Dicho valor de referencia se puede determinar de varias maneras, como apreciará el experto. Por ejemplo, se pueden determinar métricas para dos cohortes diferentes de sujetos con diferentes clasificaciones conocidas, y se puede seleccionar un valor de referencia como representativo de una clasificación (por ejemplo, una media) o un valor que está entre dos grupos de métricas (por ejemplo, elegidos para obtener la sensibilidad y especificidad deseadas). Como otro ejemplo, se puede determinar un valor de referencia en base al análisis estadísticos o simulaciones de muestras.

El término “nivel de cáncer” se puede referir a si existe cáncer (es decir, presencia o ausencia), un estadio de un cáncer, un tamaño de tumor, si hay metástasis, la carga tumoral total del cuerpo, la respuesta del cáncer al tratamiento y/u otra medida de la gravedad de un cáncer (por ejemplo, recurrencia del cáncer). El nivel de cáncer puede ser un número u otros indicios, tales como símbolos, letras del alfabeto y colores. El nivel puede ser cero. El nivel de cáncer también puede incluir condiciones (estados) premalignos o precancerosos. El nivel de cáncer se puede utilizar de varias maneras. Por ejemplo, el cribado puede comprobar si hay cáncer presente en alguien de quien no se sabía previamente que tuviera cáncer. La evaluación puede investigar a alguien a quien se le ha diagnosticado cáncer para monitorizar el progreso del cáncer a lo largo del tiempo, estudiar la efectividad de las terapias o determinar el pronóstico. En una realización, el pronóstico se puede expresar como la probabilidad de que un paciente muera de cáncer, o la probabilidad de que el cáncer progrese después de una duración o tiempo específico, o la probabilidad o extensión de que el cáncer haga metástasis. La detección puede significar ‘cribado’ o puede significar comprobar si alguien, con características sugestivas de cáncer (por ejemplo, síntomas u otras pruebas positivas), tiene cáncer.

Un “nivel de patología” (o nivel de trastorno) se puede referir a la cantidad, grado o gravedad de la patología asociada con un organismo, donde el nivel puede ser como se describió anteriormente para el cáncer. Otro ejemplo de patología es el rechazo de un órgano trasplantado. Otros ejemplos de patologías pueden incluir trastornos de impronta genética, ataques autoinmunitarios (por ejemplo, nefritis lúpica que daña el riñón o esclerosis múltiple), enfermedades inflamatorias (por ejemplo, hepatitis), procesos fibróticos (por ejemplo, cirrosis), infiltración grasa (por ejemplo, enfermedades del hígado graso), procesos degenerativos (por ejemplo, enfermedad de Alzheimer) y daño de tejido isquémico (por ejemplo, infarto de miocardio o accidente cerebrovascular). El estado de salud de un sujeto se puede considerar una clasificación de no patología.

Un “trastorno asociado al embarazo” incluye cualquier trastorno caracterizado por niveles de expresión relativa anormales de genes en tejido materno y/o fetal. Estos trastornos incluyen, pero no se limitan a, preeclampsia, restricción del crecimiento intrauterino, placentación invasiva, parto prematuro, enfermedad hemolítica del recién nacido, insuficiencia placentaria, hidropesía fetal, malformación fetal, síndrome HELLP, lupus eritematoso sistémico y otras enfermedades inmunológicas de la madre.

La abreviatura “bp” se refiere a pares de bases. En algunos casos, se puede utilizar “bp” para indicar una longitud de un fragmento de ADN, incluso aunque el fragmento de ADN pueda ser de hebra sencilla y no incluya un par de bases. En el contexto del ADN de hebra sencilla, se puede interpretar que “bp” proporciona la longitud en nucleótidos.

La abreviatura “nt” se refiere a nucleótidos. En algunos casos, se puede utilizar “nt” para indicar una longitud de un ADN de hebra sencilla en una unidad de base. También, se puede utilizar “nt” para indicar las posiciones relativas, tales como en dirección ascendente o en dirección descendente del locus que se está analizando. En algunos contextos relacionados con la conceptualización tecnológica, la presentación, el procesamiento y el análisis de datos, “nt” y “bp” se pueden utilizar indistintamente.

El término “contexto de secuencia” se puede referir a las composiciones de bases (A, C, G o T) y los órdenes de bases en un tramo de ADN. Dicho tramo de ADN podría estar rodeando una base que está sujeta o es la diana del análisis de modificación de bases. Por ejemplo, el contexto de secuencia se puede referir a bases en dirección ascendente y/o en dirección descendente de una base que se somete a análisis de modificación de bases.

El término “características cinéticas” se puede referir a características derivadas de la secuenciación, que incluyen la secuenciación en tiempo real de única molécula. Dichas características se pueden utilizar para el análisis de modificación de bases. Las características cinéticas de ejemplo incluyen contexto de secuencia en dirección ascendente y en dirección descendente, información de hebra, duración de interpulso, anchuras de pulso e intensidad de pulso. En la secuenciación en tiempo real de única molécula, se monitorizan continuamente los efectos de las actividades de una polimerasa en una plantilla de ADN. Por lo tanto, las mediciones generadas a partir de dicha secuenciación pueden considerarse características cinéticas, por ejemplo, secuencias de nucleótidos.

El término “modelos de aprendizaje automático” puede incluir modelos basados en el uso de datos de muestra (por ejemplo, datos de entrenamiento) para hacer predicciones sobre datos de prueba y, por lo tanto, puede incluir aprendizaje supervisado. Los modelos de aprendizaje automático a menudo se desarrollan utilizando un ordenador o procesador. Los modelos de aprendizaje automático pueden incluir modelos estadísticos.

El término “marco de análisis de datos” puede incluir algoritmos y/o modelos que pueden tomar datos como entrada y luego emitir un resultado previsto. Ejemplos de “marcos de análisis de datos” incluyen modelos estadísticos, modelos matemáticos, modelos de aprendizaje automático, otros modelos de inteligencia artificial y combinaciones de los mismos.

El término “secuenciación en tiempo real” se puede referir a una técnica que implica la recopilación de datos o la monitorización durante el progreso de una reacción involucrada en la secuenciación. Por ejemplo, la secuenciación en tiempo real puede implicar la monitorización óptica o la filmación de la ADN polimerasa que incorpora una nueva base.

El término “alrededor” o “aproximadamente” puede significar dentro de un rango de error aceptable para el valor particular determinado por un experto con conocimientos básicos en la técnica, que dependerá en parte de cómo se mide o determina el valor, es decir, las limitaciones del sistema de medición. Por ejemplo, “aproximadamente” puede significar dentro de 1 o más de 1 desviación estándar, de acuerdo con la práctica en la técnica. Alternativamente, “aproximadamente” puede significar un rango de hasta el 20 %, hasta el 10 %, hasta el 5 % o hasta el 1 % de un valor dado. Alternativamente, particularmente con respecto a sistemas o procesos biológicos, el término “alrededor de” o “aproximadamente” puede significar dentro de un orden de magnitud, dentro de 5 veces, y más preferiblemente dentro de 2 veces, de un valor. Cuando se describen valores particulares en la solicitud y las reivindicaciones, a menos que se indique lo contrario, se debe asumir que el término “aproximadamente” significa dentro de un rango de error aceptable para el valor particular. El término “aproximadamente” puede tener el significado que entiende comúnmente un experto con conocimientos básicos en la técnica. El término “aproximadamente” se puede referir a ±10 %. El término “aproximadamente” se puede referir a ±5 %.

Descripción detallada

Lograr la determinación sin bisulfito de una modificación de base, incluida una base metilada, es objeto de diferentes esfuerzos de investigación, pero ninguno ha demostrado ser comercialmente viable. Recientemente, se publicó un método sin bisulfito para detectar 5mC y 5hmC (Y. Liu et al., 2019) utilizando una condición suave para la conversión de bases de 5mC y 5hmC. Este método implica múltiples etapas de reacciones enzimáticas y químicas que incluyen oxidación por translocación diez-once (TET), reducción de piridina borano y PCR. La eficiencia de cada etapa de la reacción de conversión, así como el sesgo de la PCR, afectarían negativamente a la precisión final en el análisis de 5mC. Por ejemplo, se ha informado que la tasa de conversión de 5mC es de alrededor del 96 %, con una tasa de falsos negativos de alrededor del 3 %. Dicho desempeño limitaría potencialmente la capacidad de detectar ciertos cambios sutiles de metilación en un genoma. Por otro lado, la conversión enzimática no podría funcionar igual de bien en todo el genoma. Por ejemplo, la tasa de conversión de 5hmC fue un 8.2 % menor que la de 5mC, y la tasa de conversión para contextos no CpG fue un 11.4 % menor que la de contextos CpG (Y. Liu et al., 2019). Por tanto, la situación ideal es el desarrollo de enfoques para medir las modificaciones de bases de una molécula de ADN nativa sin ninguna etapa previa de conversión (química o enzimática, o combinaciones de las mismas) e incluso sin una etapa de amplificación.

Hubo una serie de estudios de prueba de concepto (Q. Liu et al., 2019; Ni et al., 2019) en los que las señales eléctricas producidas por un enfoque de secuenciación de nanoporos de lectura larga (por ejemplo, utilizando el sistema desarrollado por Oxford Nanopore Technologies) permitió detectar estados de metilación con el uso de un método de aprendizaje profundo. Además de Oxford Nanopore, existen otros enfoques de secuenciación de moléculas individuales que permiten lecturas largas. Un ejemplo es la secuenciación en tiempo real de única molécula. Un ejemplo de secuenciación en tiempo real de única molécula es el que comercializa el sistema SMRT de Pacific Biosciences. Como principio de una secuenciación en tiempo real de única molécula (por ejemplo, el sistema SMRT de Pacific Biosciences) es diferente de la de un sistema de nanoporos de base no óptica (por ejemplo, de Oxford Nanopore Technologies), enfoques para la detección de modificaciones de bases desarrollados para dichos sistemas de nanoporos de base no óptica no se puede utilizar para la secuenciación en tiempo real de única molécula. Por ejemplo, un sistema de nanoporos no óptico no está diseñado para capturar los patrones de señales fluorescentes producidas por la síntesis de ADN basada en ADN polimerasa inmovilizada (empleada mediante secuenciación en tiempo real de única molécula, tal como la del sistema SMRT de Pacific Biosciences). Como un ejemplo adicional, en la plataforma de secuenciación Oxford Nanopore, cada evento eléctrico medido está asociado con un k-mer (por ejemplo, 5-mer) (Q. Liu et al., 2019). Sin embargo, en la plataforma de secuenciación SMRT de Pacific Biosciences, cada evento fluorescente generalmente está asociado con una única base incorporada. Además, una única molécula de ADN se secuenciaría varias veces en la secuenciación SMRT de Pacific Biosciences, que incluyen las hebras de Watson y Crick. Por el contrario, para el enfoque de secuenciación de lectura larga de Oxford Nanopore, la lectura de la secuencia se realiza una vez para cada una de las hebras de Watson y Crick.

Se ha informado que la cinética de la polimerasa se vería afectada por los estados de metilación en las secuencias de E. coli (Flusberg et al., 2010). Estudios anteriores demostraron que, en comparación con la detección de 6mA, 4mC, 5hmC y 8-oxoguanina, es mucho más difícil utilizar la cinética de la polimerasa de una secuenciación en tiempo real de única molécula para deducir los estados de metilación (5mC versus C) de un CpG particular en una única molécula. La razón es que el grupo metilo es pequeño y está orientado hacia el surco principal y no participa en el emparejamiento de bases, lo que lleva a una interrupción muy sutil en la cinética causada por 5mC (Clark et al., 2013). Por lo tanto, existe una escasez de enfoques para determinar los estados de metilación de las citosinas a nivel de única molécula.

Suzuki et al desarrollaron un algoritmo (Suzuki et al., 2016) que intenta combinar las relaciones de duración interpulso (IPD) para sitios CpG vecinos para aumentar la confianza en la identificación de los estados de metilación de esos sitios. Sin embargo, este algoritmo solo permitía predecir si una región genómica estaba completamente metilada o completamente desmetilada, pero carecía de la capacidad de determinar patrones de metilación intermedios.

Con respecto a la secuenciación en tiempo real de única molécula, los enfoques actuales solo utilizaron uno o dos parámetros de forma independiente, logrando una precisión muy limitada en la detección de 5mC debido a la diferencia de medición entre 5-metilcitosina y citosina. Por ejemplo, Flusberg et al. demostró que la IPD estaba alterada en modificaciones de bases que incluían N6-metiladenosina, 5-metilcitosina y 5-hidroximetilcitosina. Sin embargo, no se encontró que la anchura de pulso (PW) de la cinética de secuenciación tuviera un efecto significativo. Por lo tanto, en el método que utilizaron para predecir la modificación de bases, utilizando la detección de N6-metiladenosina como ejemplo, solo se utilizó IPD pero no PW.

En publicaciones de seguimiento del mismo grupo (Clark et al., 2012; Clark et al. 2013), se incorporó IPD pero no PW en los algoritmos para la detección de 5-metilcitosina. En Clark et al. 2012, la tasa de detección de 5-metilcitosina sin convertirla en 5-metilcitosina solo varió entre el 1.9 % y el 4.3 %. Además, en Clark et al. 2013, los autores reafirmaron aún más la sutileza de la firma cinética de la 5-metilcitosina. Para superar la baja sensibilidad de la detección de 5-metilcitosina, Clark et al. desarrolló aún más un método que convertía 5-metilcitosina en 5-carboxilmetilcitosina utilizando proteínas de translocación Diez-once (Tet) para mejorar la sensibilidad de la 5-metilcitosina (Clark et al. 2013) porque la alteración de la IPD causada por la 5-carboxilcitosina era mucho más que por la 5-metilcitosina.

En un informe más reciente de Blow et al., el método basado en la relación de IPD descrito previamente por Flusberg et al. se utilizó para detectar las modificaciones de bases en 217 especies bacterianas y 13 arqueas con una cobertura

de lectura de 130 veces por organismo (Blow et al., 2016). Entre todas las modificaciones de bases que identificaron, sólo el 5 % involucraba 5-metilcitosina. Atribuyeron esta baja tasa de detección de 5-metilcitosina a la baja sensibilidad de la secuenciación en tiempo real de única molécula para detectar 5-metilcitosina. En la mayoría de las bacterias, un conjunto de motivos de secuencia se dirigió por las ADN metiltransferasas (MTasas) para la metilación (por ejemplo, 5'-GmATC-3' por Dam o 5'-CmCWGG-3' por Dcm en E. coli) en casi todos estos motivos en el genoma, y solo una pequeña fracción de estos sitios de motivos permanecen no metilados (Beaulaurier et al. 2019). Además, el uso del método basado en IPD para clasificar el estado de metilación de la segunda C en el motivo 5'-CCWGG-3' con o sin tratamiento con proteínas Tet produjo tasas de detección de 5-metilcitosina del 95.2 % y del 1.9 %, respectivamente (Clark et al. 2013). En conjunto, el método IPD sin conversión de bases previa (por ejemplo, utilizando proteínas Tet) omitió la mayor parte de la 5-metilcitosina.

En los estudios mencionados anteriormente (Clark et al., 2012; Clark et al., 2013; Blow et al., 2016), se utilizaron algoritmos basados en IPD sin tener en cuenta el contexto de secuencia en el que se encontraba la modificación de la base candidata. Otros grupos han intentado tener en cuenta el contexto de secuencia de un nucleótido para la detección de modificación de bases. Por ejemplo, Feng et al. utilizaron un modelo jerárquico para analizar IPD para la detección de 4-metilcitosina y 6-metiladenosina en un contexto de secuencia respectivo (Feng et al. 2013). Sin embargo, en su método, solo consideraron la IPD en la base de interés y el contexto de secuencia adyacente a esa base, pero no utilizaron la información de IPD de todas las bases vecinas adyacentes a la base de interés. Además, la PW no se consideró en el algoritmo y no presentaron datos sobre la detección de 5-metilcitosina.

En otro estudio, Schadt et al. desarrollaron un método estadístico, llamado campo aleatorio condicional, para analizar la información IPD de la base de interés y las bases vecinas para determinar si la base de interés era una 5-metilcitosina (Schadt et al., 2012). En este trabajo, también consideraron la interacción IPD entre estas bases al introducirlas en una ecuación. Sin embargo, no ingresaron la secuencia de nucleótidos, a saber, A, T, G o C, en su ecuación. Cuando aplicaron el método para determinar el estado de metilación del plásmido M.Sau3AI, el área bajo la curva ROC fue cercana a 0.5 incluso con una cobertura de secuencia de 800 veces de la secuencia del plásmido. Más aún, en su método, no habían tenido en cuenta la PW en su análisis.

En otro estudio más realizado por Beckman et al., compararon la IPD de todas las secuencias que compartían el mismo motivo de 4 nt o 6 nt en el genoma entre un genoma bacteriano diana y un genoma completamente no metilado, por ejemplo, obtenido mediante amplificación del genoma completo (Beckman et al. 2014). El propósito de dicho análisis era únicamente identificar motivos que se verían afectados con mayor frecuencia por modificaciones de base. En el estudio, solo consideraron la IPD de una base potencialmente modificada pero no la IPD de la base vecina o PW. Su método no proporcionó información sobre el estado de metilación de un nucleótido individual.

En resumen, estos intentos previos de utilizar IPD solo o con combinación de información de secuencia en los nucleótidos vecinos para agrupar datos no pudieron determinar la modificación de bases de 5-metilcitosina con precisión significativa o práctica. En una revisión reciente de Gouil et al., los autores concluyeron que debido a la baja relación señal-ruido, la detección de 5-metilcitosina en una única molécula utilizando secuenciación en tiempo real de única molécula es inexacta (Gouil et al., 2019). En estos estudios previos, aún se desconoce si puede ser factible utilizar las características cinéticas para el análisis metilómico de todo el genoma, especialmente para genomas complejos tales como genomas humanos, genomas de cáncer o genomas fetales.

A diferencia de estudios anteriores, algunas realizaciones de los métodos descritos en esta divulgación se basan en medir y utilizar IPD, PW y contexto de secuencia para cada base dentro de la ventana de medición. Razonamos que si podemos utilizar una combinación de múltiples métricas, por ejemplo, haciendo uso simultáneo de características que incluyen el contexto de secuencia en dirección ascendente y en dirección descendente, información de hebra, IPD, anchuras de pulso e intensidad de pulso, podríamos lograr una medición precisa de metilación (por ejemplo, detección de mC) con resolución de base única. El contexto de secuencia se refiere a las composiciones de bases (A, C, G o T) y los órdenes de bases en un tramo de ADN. Dicho tramo de ADN podría estar rodeando una base que está sujeta o es la diana del análisis de metilación. En una realización, el tramo de ADN podría estar proximal a una base que se somete a análisis de metilación. En otra realización, el tramo de ADN podría estar lejos de una base que se somete a análisis de metilación. El tramo de ADN podría estar en dirección ascendente y/o en dirección descendente de una base que se somete a análisis de metilación.

En una realización, las características del contexto de secuencia en dirección ascendente y en dirección descendente, información de hebra, IPD, anchuras de pulso así como intensidad de pulso, que se utilizan para el análisis de metilación, se denominan características cinéticas.

Las realizaciones presentes en esta divulgación se pueden utilizar para ADN obtenido de, pero no limitado a, estirpes celulares, muestras de un organismo (por ejemplo, órganos sólidos, tejidos sólidos, una muestra obtenida mediante endoscopia, sangre o plasma o suero u orina de una mujer embarazada, biopsia de vellosidades coriónicas, etc.), muestras obtenidas del ambiente (por ejemplo, bacterias, contaminantes celulares), alimentos (por ejemplo, carne). En algunas realizaciones, los métodos presentes en esta divulgación también se pueden aplicar después de una etapa en la que primero se enriquece una fracción del genoma, por ejemplo, utilizando sondas de hibridación (Albert et al., 2007; Okou et al., 2007; Lee et al., 2011), o enfoques en base a la separación física (por ejemplo, de acuerdo con tamaños, etc.) o después de la digestión con enzimas de restricción (por ejemplo, MspI), o enriquecimiento basado en

Cas9 (Watson et al., 2019). Si bien los métodos descritos no requieren conversión enzimática o química para funcionar, en ciertas realizaciones, se puede incluir dicha etapa de conversión para potenciar aún más el rendimiento de los métodos.

Las realizaciones de la presente divulgación permiten una precisión, practicidad o conveniencia mejoradas en la detección de la metilación. La metilación se puede detectar directamente. Las realizaciones pueden evitar la conversión enzimática o química, que puede no conservar toda la información de metilación para la detección. Adicionalmente, determinadas conversiones enzimáticas o químicas pueden no ser compatibles con determinados tipos de metilación. Las realizaciones de la presente divulgación también pueden evitar la amplificación por PCR que puede no transferir información de metilación a los productos de PCR. Adicionalmente, ambas hebras de ADN se pueden secuenciar juntas, permitiendo de esta manera el emparejamiento de la secuencia de una hebra con su secuencia complementaria a la otra hebra. Por el contrario, la amplificación por PCR divide las dos hebras de ADN de hebra doble, por lo que dicho emparejamiento de secuencias es difícil.

Los perfiles de metilación, determinados con o sin conversión enzimática o química, se pueden utilizar para analizar muestras biológicas. En una realización, los perfiles de metilación se pueden utilizar para detectar el origen del ADN celular (por ejemplo, materno o fetal, de tejido, viral o tumoral). La detección de perfiles de metilación aberrantes en tejidos ayuda a la identificación de trastornos del desarrollo en individuos y a la identificación y pronóstico de tumores o neoplasias malignas. Los desequilibrios en los niveles de metilación entre hapltotipos se pueden utilizar para detectar trastornos, que incluyen el cáncer. Los patrones de metilación en una única molécula pueden identificar ADN quimérico (por ejemplo, entre un virus y un humano) e híbrido (por ejemplo, entre dos genes normalmente no fusionados en un genoma natural); o entre dos especies (por ejemplo, a través de manipulación genética o genómica).

El análisis de metilación se puede potenciar mediante un entrenamiento mejorado, que puede incluir la reducción de los datos utilizados en un conjunto de entrenamiento. Se pueden dirigir regiones específicas para el análisis. En las realizaciones, dicho direccionamiento puede implicar una enzima que, sola o en combinación con otro(s) reactivo(s), puede escindir una secuencia de ADN o un genoma en base a su secuencia. En algunas realizaciones, la enzima es una enzima de restricción que reconoce y escinde una(s) secuencia(s) de ADN específica(s). En otras realizaciones, se pueden utilizar en combinación más de una enzima de restricción con diferentes secuencias de reconocimiento. En algunas realizaciones, la enzima de restricción se puede escindir o no en base al estado de metilación de las secuencias de reconocimiento. En algunas realizaciones, la enzima es una de la familia CRISPR/Cas. Por ejemplo, se pueden dirigir regiones genómicas de interés utilizando un sistema CRISPR/Cas9 u otro sistema basado en ARN guía (es decir, secuencias cortas de ARN que se unen a secuencias de ADN diana complementarias y en el proceso guían una enzima para que actúe en una ubicación genómica diana). En algunos casos, el análisis de metilación puede ser posible sin alineación con un genoma de referencia.

I. Detección de metilación con secuenciación en tiempo real de molécula única,

Las realizaciones de la presente invención permiten detectar directamente la metilación, sin conversión enzimática o química. Las características cinéticas (por ejemplo, contexto de secuencia, IPD y PW) obtenidas a través de una secuenciación en tiempo real de única molécula se pueden analizar con aprendizaje automático para desarrollar un modelo para detectar la metilación o la ausencia de una modificación. Los niveles de metilación se pueden utilizar para determinar el origen de las moléculas de ADN o la presencia o el nivel del trastorno.

Utilizando la secuenciación SMRT de Pacific Biosciences como ejemplo de secuenciación en tiempo real de única molécula con fines ilustrativos, se coloca una molécula de ADN polimerasa en el fondo de los pocillos que sirven como guías de ondas de modo cero (ZMW). El ZMW es un dispositivo nanofotónico para confinar la luz a un pequeño volumen de observación, que puede ser un agujero cuyo diámetro es muy pequeño y no permite la propagación de la luz en el rango de longitud de onda utilizado para la detección, de tal manera que sólo la emisión de señales ópticas del nucleótido etiquetado con tinte incorporada por la polimerasa inmovilizada es detectable frente a una señal de fondo baja y constante (Eid et al., 2009). La ADN polimerasa cataliza la incorporación de nucleótidos etiquetados con fluorescencia en hebras de ácidos nucleicos complementarias.

Sólo con fines ilustrativos, la FIG. 1 muestra un ejemplo de moléculas que llevan modificaciones de bases que se secuenciaron mediante secuenciación de consenso circular de una única molécula. Las moléculas 102, 104 y 106 llevan modificaciones de bases. Las moléculas de ADN (por ejemplo, la molécula 106) se pueden ligar con adaptadores de horquilla para formar la molécula ligada 108. La molécula ligada 108 puede entonces formar la molécula circularizada 110. Las moléculas circularizadas se pueden unir a la ADN polimerasa inmovilizada y pueden iniciar la síntesis de ADN. También se pueden secuenciar moléculas que no llevan modificaciones de bases.

La FIG. 2 muestra un ejemplo de moléculas que llevan sitios CpG metilados y/o no metilados que se secuenciaron mediante secuenciación en tiempo real de única molécula. Las moléculas de ADN se ligaron primero con adaptadores de horquilla para formar moléculas circulares que se unirían a la ADN polimerasa inmovilizada e iniciarían la síntesis de ADN. Como se muestra en la FIG. 2, la molécula de ADN 202 se liga con adaptadores de horquilla para formar la molécula ligada 204. La molécula ligada 204 forma luego la molécula circularizada 206. También se pueden secuenciar las moléculas sin sitios CpG. La molécula circularizada 206 incluye un sitio CpG no metilado 208, que aún se puede secuenciar.

Una vez inicializada la síntesis de ADN, la polimerasa inmovilizada incorporaría nucleótidos etiquetados con colorante fluorescente en la hebra recién sintetizada sobre la base de una plantilla de ADN circular, lo que conduciría a la emisión de señales ópticas. Debido a que las plantillas de ADN estaban circularizadas, toda la plantilla de ADN circular pasaría por la polimerasa múltiples veces (es decir, un nucleótido en una plantilla de ADN se secuenciaría múltiples veces). Una secuencia generada a partir del proceso, en el que todas las bases de la plantilla de ADN circularizada pasaron completamente a través de la ADN polimerasa, se denomina sublectura. Una molécula en un ZMW generaría múltiples sublecturas porque la polimerasa puede continuar alrededor de toda la plantilla circular de ADN múltiples veces. En una realización, una sublectura solo puede contener un subconjunto de la secuencia, modificaciones de bases u otra información molecular de la plantilla de ADN circular debido, en una realización, a la existencia de errores de secuenciación.

Como se ilustra en la FIG. 3, los tiempos de llegada y las duraciones de los pulsos de fluorescencia resultantes permitirían medir la cinética de la polimerasa. La duración interpulso (IPD) es una métrica para la duración de un período de tiempo entre dos pulsos de emisión, cada uno de los cuales sugeriría un nucleótido etiquetado fluorescentemente incorporado en una hebra naciente (FIG. 3). Como se muestra en la FIG. 3, la anchura de pulso (PW) es otra métrica que refleja la cinética de la polimerasa, en asociación con la duración de los pulsos relacionados con una llamada de base. PW podría ser la duración del pulso al 0 % de la altura del pico de la señal (es decir, la intensidad fluorescente del nucleótido etiquetado con colorante según se incorpora). En una realización, PW se podría definir, por ejemplo, pero no se limita a, la duración del pulso al 5 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, o el 90 % de la altura del pico de la señal. En algunas realizaciones, la PW puede ser el área bajo el pico dividida por la altura del pico de la señal.

Se ha mostrado que dichas cinéticas de polimerasa, tales como las IPD, se ven afectadas por modificaciones de bases tales como N6-metiladenina (6mA), 5-metilcitosina (5mC) y 5-hidroximetilcitosina (5hmC) en secuencias sintéticas y microbianas (por ejemplo, E. coli) (Flusberg et al., 2010). Flusberg et al. 2010 no utilizó el contexto de secuencia ni la IPD como entradas independientes para detectar una modificación, lo que resultó en un modelo que carecía de una precisión prácticamente significativa para la detección. Flusberg et al. Solo se utilizó el contexto de secuencia para confirmar que ocurrieron 6mA en GATC. Flusberg et al. no dice nada sobre el uso del contexto de secuencia en combinación con IPD como entradas para detectar el estado de metilación.

Las interrupciones débiles conferidas a la incorporación de nuevas bases a la 5-metilcitosina en hebras complementarias hacen que la llamada de metilación sea extremadamente desafiante incluso para genomas microbianos relativamente simples cuando se utilizan señales IPD únicamente, ya que se informó que la detección del motivo de metilación $C^{m}CWGG$ solo varió desde 1.9 % hasta 4.3 % (Clark et al., 2013). Por ejemplo, el paquete de software analítico (SMRT Link v6.0.0) proporcionado por Pacific Biosciences no puede realizar análisis de 5mC. Además, una versión anterior de SMRT Link v5.1.0 requería utilizar la enzima Tet1 para convertir 5mC a 5-carboxilcitosina (ScaC) antes del análisis de metilación, ya que las señales IPD asociadas con ScaC se potenciarían (Clark et al., 2013). Por lo tanto, no es sorprendente que no haya estudios que muestren la viabilidad de utilizar la secuenciación en tiempo real de única molécula para analizar el ADN nativo en todo el genoma humano.

II. Patrones de ventana de medición y modelos de aprendizaje automático

Se desean técnicas para detectar la metilación en bases sin convertir enzimática o químicamente la metilación y/o la base. Como se describe en el presente documento, la metilación en una base diana se puede detectar utilizando datos de características cinéticas obtenidos de una secuenciación en tiempo real de única molécula para las bases que rodean la base diana. Las características cinéticas pueden incluir duración interpulso, anchura de pulso y contexto de secuencia. Estas características cinéticas se pueden obtener para una ventana de medición de un cierto número de nucleótidos en dirección ascendente y en dirección descendente de la base diana. Estas características (por ejemplo, en ubicaciones particulares en la ventana de medición) se pueden utilizar para entrenar un modelo de aprendizaje automático. Como ejemplo de preparación de muestra, las dos hebras de una molécula de ADN se pueden conectar mediante adaptadores de horquilla, formando de esta manera una molécula de ADN circular. La molécula de ADN circular permite obtener características cinéticas para una o ambas hebras de Watson y Crick. Se puede desarrollar un marco de análisis de datos basado en las características cinéticas de las ventanas de medición. Este marco de análisis de datos se puede utilizar para detectar la metilación. La sección describe varias técnicas para detectar la metilación.

A. Utilización de una sola hebra

Como se muestra en la FIG. 4, como un ejemplo de la presente invención, obtuvimos las sublecturas de la hebra de Watson de la secuenciación SMRT de Pacific Biosciences para analizar una base particular con respecto a los estados de metilación. En la Fig. 4, las 3 bases de cada lado de una base que se sometió a análisis de metilación se definirían como una ventana de medición 400. En una realización, el contexto de secuencia, las IPD y las PW para estas 7 bases (es decir, 3 nucleótidos (nt) en dirección ascendente y la secuencia en dirección descendente y un nucleótido para el análisis de metilación) se compilaron en una matriz bidimensional (es decir, 2-D) como una ventana de medición. En el ejemplo mostrado, la ventana de medición 400 es para una sublectura de la hebra de Watson. Otras variaciones se describen en el presente documento.

La primera fila 402 de la matriz indicó la secuencia que se estudió. En la segunda fila 404 de la matriz, la posición 0 representaba la base para el análisis de metilación. Las posiciones relativas de -1, -2 y -3 indicaron la posición 1-nt, 2-nt y 3-nt, respectivamente, en dirección ascendente de la base que se sometió al análisis de metilación. Las posiciones relativas de +1, +2 y +3 indicaron la posición 1-nt, 2-nt y 3-nt, respectivamente, en dirección descendente de la base que se sometió al análisis de metilación. Cada posición incluye 2 columnas, que contienen los valores de IPD y PW correspondientes. Las siguientes 4 filas (filas 408, 412, 416 y 420) correspondieron a 4 tipos de nucleótidos (A, C, G y T) en la hebra (por ejemplo, hebra de Watson), respectivamente. La presencia de valores de IPD y PW en la matriz dependía de qué tipo de nucleótido correspondiente se secuenció en una posición particular. Como se muestra en la FIG. 4, en la posición relativa de 0. los valores de IPD y PW se mostraron en la fila que indica 'G' en la hebra de Watson, lo que sugiere que se llamó una guanina en el resultado de la secuencia en esa posición. Las otras cuadrículas en una columna que no correspondieran a una base secuenciada se codificarían como '0'. Como un ejemplo, la información de secuencia correspondiente a la matriz digital 2-D (FIG. 4) sería 5'-GATGACT-3' para la hebra de Watson.

Como se muestra en una realización de la invención representada en la FIG. 5, la ventana de medición podría aplicarse a los datos de la hebra de Crick. Obtuvimos las sublecturas de la hebra de Crick a partir de una secuenciación en tiempo real de única molécula para analizar una base particular con respecto a los estados de metilación. En la Fig. 5, las 3 bases de cada lado de una base que se sometió a análisis de metilación y la base sometida a análisis de metilación se definirían como una ventana de medición. En una realización, el contexto de secuencia, IPD, PW para estas 7 bases (es decir, secuencia de 3 nucleótidos (nt) en dirección ascendente y en dirección descendente y un nucleótido para el análisis de metilación) se compilaron en una matriz bidimensional (es decir, 2-D) como ventana de medición. La primera fila de la matriz indicó la secuencia que se estudió. En la segunda fila de la matriz, la posición 0 representaba la base para el análisis de metilación. Las posiciones relativas de -1, -2 y -3 indicaron la posición 1-nt, 2-nt y 3-nt, respectivamente, en dirección ascendente de la base que se sometió al análisis de metilación. Las posiciones relativas de +1, +2 y +3 indicaron la posición 1-nt, 2-nt y 3-nt, respectivamente, en dirección descendente de la base que se sometió al análisis de metilación. Cada posición incluye 2 columnas, que contenían los valores de IPD y PW correspondientes. Las siguientes 4 filas correspondieron a 4 tipos de nucleótidos (A, C, G y T) en esta hebra (por ejemplo, la hebra de Crick). La presencia de valores de IPD y PW en la matriz dependía de qué tipo de nucleótido correspondiente se secuenció en una posición particular. Como se muestra en la FIG. 5, en la posición relativa de 0. los valores de IPD y PW se mostraron en la fila que indica 'T' en la hebra de Crick, lo que sugiere que se llamó timina en el resultado de la secuencia en esa posición. Las otras cuadrículas en una columna que no correspondieran a una base secuenciada se codificarían como '0'. Como un ejemplo, la información de secuencia correspondiente a la matriz digital 2-D (FIG. 5) sería 5'-ACTTAGC-3' para la hebra de Crick.

B. Utilización de ambas hebras de Watson y Crick

La FIG. 6 muestra una realización de la invención en la que la ventana de medición se podría implementar de manera que se pudieran combinar datos de la hebra de Watson y su hebra de Crick complementaria. Como se muestra en la FIG. 6, obtuvimos las sublecturas de las hebras de Watson y Crick a partir de una secuenciación en tiempo real de única molécula para analizar una base particular para la metilación. En una realización, la ventana de medición de la hebra de Crick de la plantilla de ADN circular era complementaria a la ventana de medición de la hebra de Watson, que se sometió a análisis de metilación. En la Fig. 6, las 3 bases de cada lado de la primera base en la hebra de Watson que se sometió al análisis de metilación y la primera base se definirían como la primera ventana de medición. Las 3 bases de cada lado de la segunda base en la hebra de Crick y la segunda base se definirían como la segunda ventana de medición. La segunda base fue complementaria a la primera base. En una realización, el contexto de secuencia, las IPD, las PW para estas 7 bases (es decir, secuencia de 3 nucleótidos (nt) en dirección ascendente y en dirección descendente y un nucleótido para el análisis de metilación) de las hebras de Watson y Crick se compilaron en matrices bidimensionales (es decir, 2 -D). Estas ventanas de medición de las hebras de Watson y Crick se consideraron como la primera y segunda ventanas de medición, respectivamente.

La primera fila de la matriz de las hebras de Watson y Crick indicó la secuencia que se estudió. En la segunda fila de la matriz de la hebra de Watson, la posición 0 representaba la primera base para el análisis de metilación. La posición de 0 mostrada en la segunda fila de la matriz de la hebra de Crick representó la segunda base complementaria a la primera base. Las posiciones relativas de -1, -2 y -3 indicaron la posición 1-nt, 2-nt y 3-nt, respectivamente, en dirección ascendente de la primera y segunda bases. Las posiciones relativas de +1, +2 y +3 indicaron la posición 1-nt, 2-nt y 3-nt, respectivamente, en dirección descendente de la primera y segunda bases. Cada posición derivada de las hebras de Watson y Crick correspondería a 2 columnas que contenían los valores de IPD y PW correspondientes. Las siguientes 4 filas en las matrices de las hebras de Watson y Crick correspondieron a 4 tipos de nucleótidos (A, C, G y T) en la hebra específica (por ejemplo, la hebra de Crick), respectivamente. La presencia de valores de IPD y PW en la matriz dependía de qué tipo de nucleótido correspondiente se secuenció en una posición particular.

Como se muestra en la FIG. 6, en la posición relativa de 0. los valores de IPD y PW se mostraron en la fila que indica 'A' en la hebra de Watson y 'T' en la hebra de Crick, lo que sugiere que se llamaron adenina y timina en el resultado de la secuencia en esa posición de las hebras de Watson y Crick, respectivamente. Las otras cuadrículas en una columna que no correspondieran a la base secuenciada se codificarían como '0'. Como un ejemplo, la información de secuencia correspondiente a la matriz digital 2-D de la hebra de Watson (FIG. 6) sería 5'-ATAAGTT-3'. La información de secuencia correspondiente a la matriz digital 2-D de la hebra de Crick (FIG. 6) sería 5'-AACTTAT-3'.

Como se muestra en este ejemplo, los datos de las hebras de Watson y Crick se pueden combinar para formar una nueva matriz, que también se puede considerar como una ventana de medición. Esta nueva matriz se puede utilizar como una muestra única que se utiliza para entrenar un modelo de aprendizaje automático. Por lo tanto, todos los valores de la nueva matriz se pueden tratar como características separadas, aunque la colocación particular en la matriz 2D puede tener un impacto, por ejemplo, cuando se utiliza una red neuronal convolucional (CNN). El contexto de secuencia en las diversas posiciones para las diferentes hebras se puede transmitir a través de entradas distintas de cero en la matriz.

La FIG. 7 muestra que la ventana de medición de la presente invención se podría implementar de manera que los datos de las hebras de Watson y Crick no sean posiciones exactamente complementarias entre sí. Como se muestra en la FIG. 7, la primera ventana de medición fue 5'-ATAAGTT-3'; y la segunda ventana de medición fue 5'-GTAACGC-3'. En algunas realizaciones, las hebras de Watson y Crick se pueden desplazar entre sí de tal manera que las posiciones no sean complementarias.

La FIG. 8 muestra que se podría utilizar una ventana de medición de la presente invención para analizar los estados de metilación en los sitios CpG. La posición 0 corresponde a la citosina del sitio CpG y, por tanto, hay un desplazamiento de una posición entre las dos hebras, de tal manera que C está en la posición 0 para ambas hebras. De acuerdo con lo anterior, sólo una porción de las secuencias incluidas en la ventana de medición de las hebras de Watson y Crick son complementarias entre sí. En otras realizaciones, todas las secuencias en la ventana de medición de las hebras de Watson y Crick pueden ser complementarias entre sí. En todavía otras realizaciones, ninguna de las secuencias en la ventana de medición de las hebras de Watson y Crick es complementarias entre sí.

En una realización, para una ventana de medición, la longitud del tramo de ADN que rodea una base que se sometió a análisis de metilación podría ser asimétrica. Por ejemplo, X-nt en dirección ascendente e Y-nt en dirección descendente de esa base se podrían utilizar para el análisis de modificación de bases. X podría incluir, pero no se limita a, 00, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000, y 10000; Y podría incluir, pero no se limita a, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000, y 10000.

C. Modelos de entrenamiento y detección de metilación.

La FIG. 9 muestra un procedimiento general de la presente invención en cuanto a cómo utilizar la ventana de medición para determinar cualquier metilación. Las muestras de ADN que se sabía que estaban modificadas y no modificadas se sometieron a una secuenciación en tiempo real de única molécula. El ADN modificado (por ejemplo, la molécula 902 modificada) significa que la base (por ejemplo, la base 904) tiene la modificación (metilación) en el sitio. El ADN no modificado (por ejemplo, la molécula 906 no modificada) significa que la base (por ejemplo, la base 908) no tiene la metilación en el sitio. Ambos conjuntos de ADN se pueden crear o procesar artificialmente para formar el ADN modificado o no modificado.

En el estadio 910, las muestras luego se pueden someter a una secuenciación en tiempo real de única molécula. Como parte de la secuenciación SMRT, las moléculas circulares se podrían secuenciar múltiples veces al pasar repetidamente a través de la ADN polimerasa inmovilizada. La información de secuencia obtenida de cada vez se considerará una sublectura. De este modo, una plantilla de ADN circular generaría múltiples sublecturas. Las sublecturas de secuenciación se pueden alinear con un genoma de referencia utilizando, por ejemplo, pero sin limitarse a, BLASR (Mark J Chaisson et al, BMC Bioinformatics. 2012; 13: 238). En varias otras realizaciones, se podrían utilizar BLAST (Altschul SF et al, J Mol Biol. 1990;215(3):403-410), BLAT (Kent WJ, Genome Res. 2002;12(4):656-664), BWA (Li H et al, Bioinformatics. 2010;26(5):589-595), NGMLR (Sedlazeck FJ et al, Nat Methods. 2018;15(6):461-468), LAST (Kielbasa SM et al, Genome Res. 2011;21(3):487-493) y Minimap2 (Li H, Bioinformatics. 2018;34(18):3094-3100) para alinear sublecturas con un genoma de referencia. La alineación puede permitir que los datos de múltiples sublecturas se combinen (por ejemplo, promedien) ya que se pueden identificar los datos en cada sublectura para la misma posición.

En el estadio 912, a partir del resultado de la alineación, se obtuvieron IPD, PW y contexto de secuencia que rodea una base que se sometió a análisis de metilación. En el estadio 914, las IPD, PW y el contexto de secuencia se registraron en una determinada estructura, por ejemplo, pero no se limitan a, una matriz 2-D como se muestra en la FIG. 9.

En el estadio 916, se utilizaron una serie matrices 2-D que contenían los patrones cinéticos de referencia derivados de moléculas con metilaciones conocidas para entrenar el(los) modelo(s) analítico(s), computacional(es), matemático(s) o estadístico(s). En el estadio 918 se desarrolla un modelo estadístico resultante del entrenamiento. Por simplicidad, la FIG. 9 muestra solo un modelo estadístico desarrollado mediante entrenamiento, pero se puede desarrollar cualquier modelo o marco de análisis de datos. Los marcos de análisis de datos de ejemplo incluyen modelos de aprendizaje automático, modelos estadísticos y modelos matemáticos. Los modelos estadísticos podrían incluir, pero no se limitan a, regresión lineal, regresión logística, red neuronal recurrente profunda (por ejemplo, memoria a largo plazo, LSTM), clasificador de Bayes, modelo oculto de Markov (HMM), análisis discriminante lineal

(LDA), agrupamiento k-medias, agrupamiento espacial de aplicaciones con ruido basada en densidad (DBSCAN), algoritmo de bosque aleatorio y máquina de vectores de soporte (SVM). Un tramo de ADN que rodea una base que se sometió a análisis de metilación podría ser X-nt en dirección ascendente e Y-nt en dirección descendente de esa base, a saber, "ventana de medición".

Las estructuras de datos se pueden utilizar en un proceso de entrenamiento, ya que se conocen los resultados correctos (es decir, el estado de metilación). Por ejemplo, las IPD, PW y el contexto de secuencia correspondiente a 3 nt en dirección ascendente y en dirección descendente de una base de la(s) hebra(s) de Watson y/o Crick se pueden utilizar para construir la matriz 2-D que se utilizará para ellos modelo(s) estadístico(s) para clasificar las metilaciones. De esta manera, el entrenamiento puede proporcionar un modelo que pueda clasificar una metilación en una posición de un ácido nucleico con un estado previamente conocido.

La FIG. 10 muestra un procedimiento general de la presente invención sobre cómo el(los) modelo(s) estadístico(s) aprendido(s) de muestras de ADN que llevaban estados conocidos de metilaciones pueden detectar la metilación. Una muestra con estados desconocidos de metilaciones se sometió a secuenciación SMRT. Las sublecturas de secuenciación se alinearon con un genoma de referencia utilizando, por ejemplo, las técnicas mencionadas anteriormente. Además o en su lugar, las sublecturas se pueden alinear entre sí. Todavía otras realizaciones pueden utilizar solo una sublectura o analizarlas de forma independiente de tal manera que no se realice la alineación.

Para una base que se sometió a análisis de metilación, se obtendrían las IPD, las PW y contexto de secuencia de la(s) hebra(s) de Watson y/o Crick en los resultados de alineación utilizando una ventana de medición comparable a la utilizada en la etapa de entrenamiento (FIG. 9), y estaba asociado con esa base. En otra realización, las ventanas de medición entre los procedimientos de entrenamiento y prueba serían diferentes. Por ejemplo, el tamaño de las ventanas de medición entre los procedimientos de entrenamiento y de prueba podría ser diferente. Esas IPD, PW y contexto de secuencia se transformarían en una matriz 2-D. Dicha matriz 2-D de una muestra de prueba se compararía con las características cinéticas de referencia para determinar las metilaciones. Por ejemplo, la matriz 2-D de una muestra de prueba se puede comparar con características cinéticas de referencia a través del(los) modelo(s) estadístico(s) que se aprendieron de las muestras de entrenamiento, de tal manera que se podrían determinar las metilaciones en sitios en las moléculas de ácido nucleico en una muestra de prueba. Los modelos estadísticos podrían incluir, pero no se limitan a, regresión lineal, regresión logística, red neuronal recurrente profunda (por ejemplo, memoria a largo plazo, LSTM), clasificador de Bayes, modelo oculto de Markov (HMM), análisis discriminante lineal (LDA), agrupamiento k-medias, agrupamiento espacial de aplicaciones con ruido basada en densidad (DBSCAN), algoritmo de bosque aleatorio y máquina de vectores de soporte (SVM).

La FIG. 11 muestra un procedimiento general sobre cómo se podría elaborar el método para clasificar estados de metilación en sitios CpG. Las muestras de ADN que se sabía que estaban desmetiladas y metiladas en sitios CpG se sometieron a secuenciación en tiempo real de única molécula. Las sublecturas de secuenciación se alinearon con un genoma de referencia. Se utilizaron datos de la hebra de Watson.

A partir del resultado de la alineación, las IPD, PW y el contexto de secuencia que rodean una citosina en un sitio CpG que se sometió a análisis de metilación se obtuvieron y registraron en una determinada estructura, por ejemplo, pero no limitada a, matriz 2-D como se muestra. en la Fig. 11. Para entrenar el(los) modelo(s) estadístico(s) se utilizaron una serie matrices 2-D que contenían los patrones cinéticos de referencia derivados de moléculas con estados de metilación conocidos. Un tramo de ADN que rodea una base bajo interrogación podría ser X-nt en dirección ascendente e Y-nt en dirección descendente de esa base, a saber, la "ventana de medición". X podría incluir, pero no se limitan a, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000, y 10000; Y podría incluir, pero no se limita a, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000, y 10000. En una realización, las IPD, las PW y el contexto de secuencia correspondiente a 3- nt en dirección ascendente y en dirección descendente de una base de la hebra de Watson se podría utilizar para construir la matriz 2-D que se utilizó para entrenar el(los) modelo(s) estadístico(s) para clasificar las metilaciones.

La FIG. 12 muestra un procedimiento general para clasificar los estados de metilación de una muestra desconocida. Una muestra con estados de metilación desconocidos se sometió a una secuenciación en tiempo real de única molécula. Las sublecturas de secuenciación se alinearon con un genoma de referencia.

Para una citosina de un sitio CG en el resultado de la alineación, se obtendrían las IPD, las PW y contexto de secuencia de la hebra de Watson utilizando una ventana de medición comparable que se aplicó en la etapa de entrenamiento (FIG. 11), asociada con esa base cuya metilación estaba bajo interrogación. Esos IPD, PW y contexto de secuencia se pueden transformar en una matriz 2-D. Dicha matriz 2D de una muestra de prueba se compararía con los patrones cinéticos de referencia ilustrados en la FIG. 11 para determinar los estados de metilación. X11

La FIG. 13 y FIG. 14 muestran que las características cinéticas de la hebra de Crick se podrían utilizar para los procedimientos de entrenamiento y prueba de la invención como se explicó anteriormente, de manera similar a los procedimientos con la hebra de Watson. El(los) modelo(s) estadístico(s) podrían ser iguales o diferentes. Cuando son

modelos diferentes, se podrían utilizar para obtener clasificaciones independientes, que se pueden comparar; por ejemplo, si están de acuerdo entonces se identifica un estado de modificación. Si no están de acuerdo, entonces se podría identificar un estatus de no clasificado. Cuando son el mismo modelo, los datos se pueden combinar en una única estructura de datos, por ejemplo, la matriz en la FIG. 6.

La FIG. 15 y FIG. 16 muestran que las características cinéticas de las hebras de Watson y Crick se podrían utilizar para los procedimientos de entrenamiento y prueba de la invención como se explicó anteriormente. Las muestras de ADN que se sabía que estaban desmetiladas y metiladas en sitios CpG se sometieron a secuenciación en tiempo real de única molécula. Las sublecturas de secuenciación se alinearon con un genoma de referencia, aunque es posible la alineación de las sublecturas entre sí, como se puede hacer con otros métodos descritos en el presente documento.

Para una sublectura en el resultado de la alineación, se obtuvieron IPD, PW y contexto de secuencia que rodean una citosina de un sitio CpG que se sometió a análisis de metilación. Debido a que las moléculas de ADN se circularizaron a través del uso de dos adaptadores de horquilla (por ejemplo, siguiendo un protocolo de preparación de plantilla SMRTBell), las moléculas circulares se pudieron secuenciar más de una vez, generando de esta manera múltiples sublecturas de una molécula. Las sublecturas se pueden utilizar para generar lecturas de secuenciación de consenso circular (CCS). En general, para todos los métodos descritos en el presente documento, un ZMW podría generar múltiples sublecturas pero solo correspondería a una lectura CCS.

En algunas realizaciones, el conjunto de datos completamente no metilado se podría crear mediante PCR en fragmentos de ADN humano. Por ejemplo, el conjunto de datos completamente metilado se podría producir a través de fragmentos de ADN humano tratados con CpG metiltransferasa M.SssI, en los que se suponía que todos los sitios CpG estaban metilados. En otros ejemplos, se podría utilizar otra CpG metiltransferasa, tal como M.MpeI. En otras realizaciones, secuencias sintéticas con estados de metilación conocidos o muestras de ADN preexistentes con diferentes niveles de metilación, o estados metilados híbridos creados mediante corte con enzimas de restricción de moléculas de ADN metiladas y no metiladas seguido de ligación (lo que crearía una proporción de moléculas de ADN quiméricas metiladas/no metiladas) se podrían utilizar para entrenar los modelos o clasificadores de predicción de metilación.

La transformación de patrones cinéticos, que incluye el contexto de secuencia, IPD y anchura de pulso (PW), puede ser una matriz 2-D que comprende características de hebras de Watson y Crick para analizar estados de metilación en sitios CG, como se ilustra en la FIG. 15. Este enfoque nos permitió capturar con precisión los cambios cinéticos sutiles causados por las citosinas metiladas, así como su contexto de secuencia cercano. Como con cualquiera de los diversos métodos descritos en el presente documento, para cada CpG presente en una sublectura, la ventana de medición de (por ejemplo, 3 bases en dirección ascendente y en dirección descendente de una citosina de un sitio CpG) se puede utilizar para análisis posteriores, lo que lleva a un total de 7 nucleótidos (que incluyen la citosina de un sitio CpG) que se analizaron juntos. Se pueden calcular la IPD y la PW para cada base entre esos 7 nucleótidos. Para capturar el contexto de secuencia que se atribuye a los cambios cinéticos, las señales IPD y PW se pueden compilar en una llamada de base particular, posiciones de secuenciación relativas y la información de la hebra como se muestra en la FIG. 15. Dicha estructura de datos se denomina matriz cinética digital 2-D por simplicidad.

Dicha matriz digital 2-D es análoga a una “imagen digital 2-D”. Por ejemplo, la primera fila de la matriz digital 2-D contenía las posiciones relativas que rodean una citosina de un locus CpG que se sometió a análisis de metilación, con 3 nt en dirección ascendente y en dirección descendente de ese sitio de citosina. La posición 0 representaba el sitio de citosina cuya metilación se iba a determinar. Las posiciones relativas de -1 y -2 indicaron el 1 nt y el 2 nt en dirección ascendente de la citosina en cuestión. Las posiciones relativas de +1 y +2 indicaron el 1-nt y el 2-nt en dirección descendente de la citosina que se utilizarían. Cada posición correspondería a 2 columnas que contenían los valores de IPD y PW correspondientes. Cada fila correspondía a los 4 tipos de nucleótidos (A, C, G y T) en las hebras de Watson y Crick. El llenado de los valores de IPD y PW en la matriz dependía de qué tipo de nucleótido correspondiente estaba preestablecido en el resultado secuenciado (es decir, sublectura) en una posición particular.

Como se muestra en la FIG. 15, en la posición relativa de 0. los valores de IPD y PW se mostraron en la fila de ‘C’ en la hebra de Watson, lo que sugiere que se llamó citosina en esa posición. Las otras cuadrículas en una columna que no correspondieran a una base secuenciada se codificarían como ‘0’. Como un ejemplo, la información de secuencia correspondiente a la matriz digital 2-D (FIG. 15) sería 5’-ATACGTT-3’ y 5’-TAACGTA-3’ para las hebras de Watson y Crick, respectivamente. En este contexto, las secuencias en dirección ascendente y en dirección descendente que flanquean una citosina de un sitio CpG en las hebras de Watson y Crick serían diferentes. Dado que la metilación en los sitios CpG sería simétrica entre las hebras de Watson y Crick (Lister et al., 2009), se utilizó la cinética en ambas hebras para entrenar el modelo de predicción de la metilación en una realización preferida. En otra realización, las hebras de Watson y Crick se podrían utilizar para entrenar el modelo de predicción de metilación por separado.

Teniendo en cuenta el alto rendimiento de datos de la secuenciación en tiempo real de única molécula, en una realización, un algoritmo de aprendizaje profundo (por ejemplo, redes neuronales convolucionales (CNN)) (LeCun et al., 1989) puede ser adecuado para distinguir los CpG metilados. de CpG no metilados. También se podrían utilizar otros algoritmos además o en su lugar, por ejemplo, pero no se limitan a, regresión lineal, regresión logística, red neuronal recurrente profunda (por ejemplo, memoria a largo plazo, LSTM), clasificador de Bayes, modelo oculto de Markov (HMM)., análisis discriminante lineal (LDA), agrupamiento de k-medias, agrupamiento espacial de aplicaciones

con ruido basada en densidad (DBSCAN), algoritmo de bosque aleatorio y máquina de vectores de soporte (SVM), etc. El entrenamiento puede utilizar las hebras de Watson y Crick por separado o en una nueva matriz combinada, como se describe en las FIG. 6-8.

Otra transformación de patrones cinéticos podría ser una matriz N-dimensional. N podría ser, por ejemplo, 1, 3, 4, 5, 6 y 7. Por ejemplo, la matriz 3-D sería una pila de matrices 2-D estratificadas de acuerdo con el número de sitios CG en tándem para un tramo de ADN. que se va a analizar, en el que la 3ª dimensión sería el número de sitios CG en tándem en ese tramo de ADN. La intensidad de pulso o la magnitud del pulso (por ejemplo, medida por la altura máxima de un pulso o por el área bajo la señal del pulso) también se podrían incorporar en una matriz en algunas realizaciones. La intensidad de pulso (una métrica para la amplitud del pico del pulso, FIG. 3) se podría agregar a una columna adicional adyacente a las columnas en asociación con los valores de PW e IPD en la parte superior de la matriz 2-D original, o agregar a una 3ª dimensión para formar una matriz 3-D.

Como ejemplos adicionales, una matriz 2D de 8 (fila) x 21 (columna) se puede transformar en una matriz 1-D (es decir, vector) que comprende 168 elementos. Y podemos escanear esta matriz 1-D, por ejemplo, para realizar CNN u otro modelado. Como otro ejemplo, los métodos pueden dividir una matriz 2D de 8 x 21 en múltiples matrices más pequeñas, por ejemplo, dos matrices 2D de 4 x 21. Al juntar estas dos matrices más pequeñas en dirección vertical se obtiene una matriz tridimensional (es decir, x=21, y=4, z=2). Los métodos pueden escanear la 1ª matriz 2D y luego la 2ª matriz 2D, para formar la presentación de datos para el aprendizaje automático. Los datos se pueden dividir aún más para formar una matriz de dimensiones superiores. Adicionalmente, se puede agregar información de estructura secundaria a la estructura de datos, por ejemplo, una matriz adicional (matriz 1-D) encima de la matriz 2-D. Dicha matriz adicional puede codificar si cada base dentro de la ventana de medición está involucrada en una estructura secundaria (por ejemplo, estructura de tallo-bucle), por ejemplo, la base que involucra el “tallo” se codifica como 0 y la base que involucra el “bucle” se codifica como 1.

En una realización, el estado de metilación de un sitio CpG dentro de una única molécula de ADN se puede expresar como una probabilidad de ser metilado basándose en un modelo estadístico, en lugar de dar un resultado cualitativo de “metilado” o “no metilado”. Una probabilidad de 1 indica que, en base al modelo estadístico, un sitio CpG se puede considerar metilado. Una probabilidad de 0 indica que, en base al modelo estadístico, un sitio CpG se puede considerar no metilado. En análisis en dirección descendente posteriores, se puede utilizar un valor de corte para clasificar si un sitio CpG particular se clasifica como “metilado” o “no metilado” en base a la probabilidad. Los posibles valores del valor de corte incluyen 5 %, 10 %, 15 %, 20 %, 25 %, 30 %, 35 %, 40 %, 45 %, 50 %, 55 %, 60 %, 65 %, 70 %, 75 %, 80 %, 85 %, 90 % o 95 %. La probabilidad prevista de ser metilado para un sitio CpG mayor que un valor de corte predefinido se puede clasificar como “metilado”, mientras que la probabilidad de ser metilado para un sitio CpG no mayor que un valor de corte predefinido se puede clasificar como “no metilado”. Se obtendría un valor de corte deseado a partir del conjunto de datos de entrenamiento utilizando, por ejemplo, el análisis de la curva de características operativas del receptor (ROC).

La FIG. 16 muestra un procedimiento general para clasificar los estados de metilación de una muestra desconocida de las hebras de Watson y Crick. la muestra con estados de metilación desconocidos se sometió a secuenciación en tiempo real de única molécula. Las sublecturas de secuenciación se pueden alinear con un genoma de referencia o entre sí, como ocurre con otros métodos, para determinar valores de consenso (por ejemplo, promedio, mediana, moda u otro valor estadístico) para una posición determinada. Como se muestra, los valores medidos para las dos hebras se pueden combinar en una única matriz 2D.

Para una citosina de un sitio CG en el resultado de la alineación, se obtendrían IPD, PW y contexto de secuencia de la hebra de Watson utilizando una ventana de medición comparable (3 nt en dirección ascendente y en dirección descendente de una citosina de un sitio CpG) como aplicado en la etapa de entrenamiento (FIG. 16), asociada con esa base cuya metilación estaba bajo interrogación, aunque se pueden utilizar ventanas de diferentes tamaños. Dicha matriz 2D de una muestra de prueba se puede comparar con los patrones cinéticos de referencia ilustrados en la FIG. 16 para determinar los estados de metilación.

III. Ejemplo de modelo de entrenamiento para la detección de metilación

Para probar la viabilidad y validez de los enfoques propuestos de la invención, preparamos una biblioteca de ADN de placenta con tratamiento con M.SssI (biblioteca metilada) y amplificación por PCR (biblioteca no metilada) antes de la secuenciación en tiempo real de única molécula. Obtuvimos 44,799,736 y 43,580,452 sublecturas para bibliotecas metiladas y no metiladas, respectivamente, correspondientes a 421,614 y 446,285 secuencias de consenso circulares (CCS). Como resultado, cada molécula se secuenció con una mediana de 34 y 32 veces en bibliotecas metiladas y no metiladas. El conjunto de datos se generó a partir de ADN preparado con el Kit de Secuenciación Sequel 3.0 de Pacific Biosciences. Este kit fue desarrollado para utilizarse con el secuenciador Sequel original de Pacific Biosciences. Para diferenciar el Sequel de su sucesor, Sequel II, en el presente documento nos referimos a la Sequel original como Sequel I. Por lo tanto, al Kit de secuenciación Sequel 3.0 se le denominará en el presente documento Kit de Secuenciación Sequel I 3.0. Los kits de secuenciación diseñados para el secuenciador Sequel II incluyen el Kit de Secuenciación Sequel II 1.0 y el Kit de Secuenciación Sequel II 2.0 que también se describen en esta divulgación.

Utilizamos el 50 % de las moléculas secuenciadas generadas a partir de bibliotecas metiladas y no metiladas para entrenar un modelo estadístico (y utilizamos el 50 % restante para la validación), que en este caso es un modelo de red neuronal convolucional (CNN). Como un ejemplo, el modelo CNN puede tener una o más capas convolucionales (por ejemplo, capas 1D o 2D). Una capa convolucional puede utilizar uno o más filtros diferentes, y cada filtro utiliza un núcleo que opera con valores de matriz locales (por ejemplo, vecinos o circundantes) de un elemento de matriz particular, proporcionando de esta manera un nuevo valor al elemento de matriz particular. Una implementación utilizó dos capas de convolución 1D (cada una con 100 filtros con un tamaño de núcleo de 4). Los filtros se pueden aplicar por separado y luego combinar (por ejemplo, en un promedio ponderado). Una matriz resultante puede ser más pequeña que la matriz de entrada.

Las capas convolucionales pueden ir seguidas de una capa ReLU (unidad lineal rectificada), a la que puede seguir una capa de exclusión con una tasa de exclusión de 0.5. la ReLU es un ejemplo de una función de activación que puede operar en los valores individuales que dan como resultado la nueva matriz (imagen) de la(s) capa(s) convolucional(es). También se pueden utilizar otras funciones de activación (por ejemplo, sigmoide, softmax, etc.). Se pueden utilizar una o más de dichas capas. La capa de exclusión se puede utilizar en la capa ReLU o en una capa de agrupación máxima y actuar como una regularización para evitar el sobreajuste. La capa de exclusión se puede utilizar durante el proceso de entrenamiento para ignorar valores diferentes (por ejemplo, aleatorios) durante diferentes iteraciones de un proceso de optimización (por ejemplo, para reducir una función de coste/pérdida) que se realiza como parte del entrenamiento.

Se puede utilizar una capa de agrupación máxima (por ejemplo, un tamaño de agrupación de 2) después de la capa ReLU. La capa de agrupación máxima puede actuar de manera similar a la capa de convolución, pero en lugar de tomar un producto de punto entre la entrada y el núcleo, se puede tomar el máximo de la región de la entrada superpuesta por el núcleo. Se pueden utilizar más capa(s) convolucional(es). Por ejemplo, los datos de una capa de agrupación se pueden ingresar a otras dos capas de convolución 1D (por ejemplo, cada una con 128 filtros con un tamaño de núcleo de 2 seguidas de una capa ReLU), utilizando además una capa de exclusión con una tasa de exclusión de 0.5. Se utilizó una capa de agrupación máxima con un tamaño de agrupación de 2. Finalmente, se puede utilizar una capa completamente conectada (por ejemplo, con 10 neuronas seguidas de una capa ReLU). Una capa de salida con una neurona puede ir seguida de una capa sigmoidea, lo que produce la probabilidad de metilación. Se pueden adaptar varias configuraciones de capas, filtros y tamaños de núcleo. En este conjunto de datos de entrenamiento, utilizamos 468,596 y 432,761 sitios CpG de bibliotecas metiladas y no metiladas.

A. Resultados de conjuntos de datos de entrenamiento y prueba

La FIG. 17A muestra la probabilidad de ser metilado para cada sitio CpG en cada molécula de ADN en el conjunto de datos de entrenamiento. La probabilidad de metilación fue mucho mayor en la biblioteca metilada que en la biblioteca no metilada. Para un valor de corte de 0.5 para la probabilidad de estar metilados, se predijo correctamente que el 94.7 % de los sitios CpG no metilados estaban desmetilados y que el 84.7 % de los sitios CpG metilados estaban metilados.

La FIG. 17B muestra el rendimiento del conjunto de datos de prueba. Utilizamos un modelo entrenado por el conjunto de datos de entrenamiento para predecir los estados de metilación de 469,729 y 432,024 sitios CpG en un conjunto de datos de prueba independiente de bibliotecas metiladas y no metiladas. Para un valor de corte de 0.5 para la probabilidad de estar metilados, se predijo correctamente que el 94.0 % de los sitios CpG no metilados estaban desmetilados y que el 84.1 % de los sitios CpG metilados estaban metilados. Estos resultados sugirieron que el uso de una nueva transformación de la cinética junto con el contexto de la secuencia podría permitir la determinación de los estados de metilación en el ADN (por ejemplo, de sujetos humanos).

Evaluamos el poder de cada característica (contexto de secuencia, IPD y PW) para predecir el estado de metilación de CpG al incluir un subconjunto de las características en el modelo. En el conjunto de datos de entrenamiento, los modelos con (i) solo contexto de secuencia, (ii) solo la IPD y (iii) solo la PW dieron valores de área bajo la curva (AUC) de 0.5, 0.74 y 0.86, respectivamente. Mientras que la combinación de IPD y contexto de secuencia mejoró el rendimiento con un AUC de 0.86. El análisis combinado del contexto de secuencia (“Seq”), IPD y PW mejoró sustancialmente el rendimiento con un AUC de 0.94 (FIG. 18A). El rendimiento de un conjunto de datos de prueba independiente fue comparable al conjunto de datos de entrenamiento (FIG. 18B).

Definimos la profundidad de sublectura de un sitio CpG como el número promedio de sublecturas que lo cubren y sus 10 bp circundantes. Como se muestra en la FIG. 19A y FIG. 19B, cuanto mayor sea la profundidad de la sublectura de un sitio CpG, mayor será la precisión de detección de metilación que lograríamos. Por ejemplo, como se muestra en el conjunto de datos de prueba (FIG. 19B), si la profundidad de cada sitio CpG fuera al menos 10, el AUC para predecir estados de metilación sería 0.93. Sin embargo, si la profundidad de la sublectura de cada sitio CpG es al menos 300, el AUC para predecir los estados de metilación sería 0.98. Por otro lado, incluso para la profundidad de 1, podríamos lograr un AUC de 0.9, lo que sugiere que nuestro enfoque podría lograr la predicción de la metilación con el uso de una profundidad de secuenciación baja.

Para probar el efecto de la información de la hebra sobre el rendimiento del análisis de metilación, se utilizaron el contexto de secuencia, IPD y PW que se originan de las hebras de Watson y Crick para entrenar de acuerdo con las

realizaciones presentes en esta divulgación, respectivamente. Las FIG. 20A y FIG. 20B mostraron que es factible utilizar una sola hebra, concretamente la hebra de Watson o Crick, para el entrenamiento y las pruebas, ya que el AUC podría alcanzar hasta 0.91 y 0.87 en conjuntos de datos de entrenamiento y prueba. El uso de ambas hebras (por ejemplo, como se describe en la FIGS. 6-8), que incluyen las hebras de Watson y Crick, daría lugar al mejor rendimiento (AUC: 0.94 y 0.90 en conjuntos de datos de entrenamiento y prueba, respectivamente), lo que sugiere que la información de la hebra sería ser importante para lograr el rendimiento óptimo.

Probamos además el número diferente de nucleótidos en dirección ascendente y en dirección descendente de un sitio CpG, para estudiar cómo este parámetro afectó el rendimiento de acuerdo con las realizaciones presentes en esta divulgación desarrolladas en esta divulgación. Las FIG. 21A y FIG. 21B muestran que el número de nucleótidos en dirección ascendente y en dirección descendente de una citosina en el contexto de CpG afectaría la precisión de la predicción de la metilación. Por ejemplo, a modo de ilustración, considerando, pero sin limitarse a, 2 nucleótidos (nt), 3 nt, 4 nt, 6 nt, 8 nt, 10 nt, 15 nt y 20 nt en dirección ascendente y en dirección descendente de una citosina que se está analizando, el AUC de un método que utiliza 2 nt en dirección ascendente y descendente de la citosina que se interroga sería solo 0.50 en ambos conjuntos de datos de entrenamiento y de prueba, mientras que el AUC de un método que utiliza 15 nt en dirección ascendente y descendente de una citosina que se interroga aumentaría a 0.95 y 0.92 en el conjuntos de datos de entrenamiento y prueba. Estos resultados sugirieron que variar la longitud de las regiones en dirección ascendente y descendente que flanquean las citosinas que se analizan permitiría determinar el rendimiento óptimo. En una realización, como se muestra en la FIG. 21B, se utilizarían 3 nt en dirección ascendente y en dirección descendente de una citosina para determinar los estados de metilación, lo que podría lograr un AUC de 0.89.

En una realización, se podrían utilizar secuencias asimétricas que flanquean una citosina que se está interrogando para realizar el análisis de acuerdo con las realizaciones presentes en esta divulgación. Por ejemplo, se podría utilizar 2 nt en dirección ascendente combinado con 1 nt, 3 nt, 4 nt, 5 nt, 6 nt, 7 nt, 8 nt, 9 nt, 10 nt, 11 nt, 12 nt, 13 nt, 14 nt, 15 nt, 16 nt, 17 nt, 18 nt, 19 nt, 20 nt, 25 nt, 30 nt, 35 nt, y 40 nt en dirección descendente de una citosina; se podría utilizar 3 nt en dirección ascendente combinado con 1 nt, 2 nt, 4 nt, 5 nt, 6 nt, 7 nt, 8 nt, 9 nt, 10 nt, 11 nt, 12 nt, 13 nt, 14 nt, 15 nt, 16 nt, 17 nt, 18 nt, 19 nt, 20 nt, 25 nt, 30 nt, 35 nt, y 40 nt en dirección descendente de una citosina; se podría utilizar 4 nt en dirección ascendente combinado con 1 nt, 2 nt, 3 nt, 5 nt, 6 nt, 7 nt, 8 nt, 9 nt, 10 nt, 11 nt, 12 nt, 13 nt, 14 nt, 15 nt, 16 nt, 17 nt, 18 nt, 19 nt, 20 nt, 25 nt, 30 nt, 35 nt, y 40 nt en dirección descendente de una citosina. Como otro ejemplo, se podría utilizar 2 nt en dirección descendente combinado con 1 nt, 3 nt, 4 nt, 5 nt, 6 nt, 7 nt, 8 nt, 9 nt, 10 nt, 11 nt, 12 nt, 13 nt, 14 nt, 15 nt, 16 nt, 17 nt, 18 nt, 19 nt, 20 nt, 25 nt, 30 nt, 35 nt, y 40 nt en dirección ascendente de una citosina; se podría utilizar 3 nt en dirección descendente combinado con 1 nt, 2 nt, 4 nt, 5 nt, 6 nt, 7 nt, 8 nt, 9 nt, 10 nt, 11 nt, 12 nt, 13 nt, 14 nt, 15 nt, 16 nt, 17 nt, 18 nt, 19 nt, 20 nt, 25 nt, 30 nt, 35 nt, y 40 nt en dirección ascendente de una citosina; se podría utilizar 4 nt en dirección descendente combinado con 1 nt, 2 nt, 3 nt, 5 nt, 6 nt, 7 nt, 8 nt, 9 nt, 10 nt, 11 nt, 12 nt, 13 nt, 14 nt, 15 nt, 16 nt, 17 nt, 18 nt, 19 nt, 20 nt, 25 nt, 30 nt, 35 nt, y 40 nt en dirección ascendente de una citosina. Al aprovechar las IPD, las PW, la información de la hebra y el contexto de secuencia en asociación con n-nt en dirección ascendente y m-nt en dirección descendente de una citosina podría proporcionar una precisión mejorada en la determinación de los estados de metilación en ciertas realizaciones. Dichas ventanas de medición variables se podrían aplicar a otros tipos de análisis de metilación, como 5hmC, 6mA, 4mC y oxoG, o cualquier modificación divulgada en el presente documento. Dichas ventanas de medición variables podrían incluir el análisis de la estructura secundaria del ADN, como la estructura G-quadruplex y la estructura de tallo-bucle. Dicho ejemplo se explicó anteriormente. Esta información de estructura secundaria también se podría agregar como otra columna en una matriz.

Las FIG. 22A y FIG. 22B muestran que es factible determinar los estados de metilación utilizando patrones cinéticos asociados solo con bases en dirección descendente de al menos 3 bases. De acuerdo con las realizaciones presentes en esta divulgación, con el uso de características asociadas con la citosina y sus 3, 4, 6, 8 y 10 bases en dirección descendente, las AUC de la determinación de los estados de metilación en el conjunto de datos de entrenamiento fueron 0.91, 0.92, 0.94. 0.94 y 0.94, respectivamente, en el conjunto de datos de entrenamiento; las AUC fueron 0.87, 0.88, 0.90, 0.90 y 0.90, respectivamente, en el conjunto de datos de prueba.

Las FIG. 23A y FIG. 23B muestran, sin embargo, si solo se utilizan las características asociadas con las bases en dirección ascendente, el poder de clasificación parece disminuir en la capacidad de distinguir los estados de metilación. Las AUC en el conjunto de datos de entrenamiento y el conjunto de datos de prueba fueron todas de 0.50 para 2 a 10 bases en dirección ascendente.

Las FIG. 24 y FIG. 25 muestran que diferentes combinaciones de bases en dirección ascendente y en dirección descendente permitirían lograr un poder de clasificación óptimo para determinar los estados de metilación. Por ejemplo, las características asociadas con 8 bases en dirección ascendente y 8 bases en dirección descendente de una citosina lograrían un mejor rendimiento en este conjunto de datos, con un AUC de 0.94 y 0.91 en los conjuntos de datos de entrenamiento y prueba, respectivamente.

La FIG. 26 muestra la importancia relativa de las características con respecto a la clasificación de los estados de metilación en los sitios CpG. 'W' y 'C' entre paréntesis indican la información de la hebra, 'W' para la hebra de Watson y 'C' para la hebra de Crick. La importancia de cada característica, que incluye el contexto de secuencia, IPD y PW, se determinó utilizando bosque aleatorio. El análisis de árboles de bosque aleatorio mostró que la importancia de las

características de las IPD y las PW alcanzó su punto máximo en la dirección descendente de una citosina que estaba bajo interrogatorio, revelando que las principales contribuciones al poder de clasificación fueron las IPD y las PW en dirección descendente de una citosina que estaba bajo interrogatorio.

El bosque aleatorio estaba compuesto de múltiples árboles de decisión. Durante la construcción del árbol de decisión, se utilizó la impureza de Gini para determinar qué lógica de decisión para los nodos de decisión se debería tomar. Las características importantes que tienen más influencia en el resultado final de la clasificación probablemente se encuentran en los nodos más cercanos a la raíz del árbol de decisión, mientras que las características sin importancia que tienen menos influencia en el resultado final de la clasificación probablemente se encuentran en los nodos más alejados de la raíz. Por tanto, la importancia de la característica se podría estimar al calcular la distancia promedio relativa a las raíces de todos los árboles de decisión en el bosque aleatorio.

En algunas realizaciones, el consenso de llamadas de metilación en sitios CpG entre las hebras de Watson y Crick se podría utilizar además para mejorar la especificidad. Por ejemplo, podría ser necesario que ambas hebras que se muestran metiladas se denominen en estado metilado, y ambas hebras que se muestran sin metilar se denominen en estado no metilado. Dado que se sabía que la metilación en los sitios CpG era normalmente simétrica, la confirmación de cada hebra puede mejorar la especificidad.

En diversas realizaciones, las características cinéticas generales de una molécula completa se podrían utilizar para la determinación de los estados de metilación. Por ejemplo, la metilación en una molécula completa afectaría la cinética de la molécula completa durante la secuenciación en tiempo real de única molécula. Al modelar la cinética de secuenciación de toda la molécula de ADN de plantilla, que incluye las IPD, PW, tamaños de fragmentos, información de hebra y contexto de secuencia, se puede mejorar la precisión de la clasificación sobre si una molécula está metilada o no. Como un ejemplo, las ventanas de medición pueden ser la molécula plantilla completa. Se pueden utilizar valores estadísticos (por ejemplo, media, mediana, moda, percentil, etc.) de IPD, PW u otras características cinéticas para determinar la metilación de una molécula completa.

B. Limitaciones de otras técnicas de análisis

Se informó que la detección de metilación basada en IPD para una C particular en un motivo de secuencia particular era muy baja, por ejemplo, una sensibilidad de solo el 1.9 % (Clark et al., 2013). También intentamos reproducir dicho análisis al combinar diferentes motivos de secuencia con IPD sin utilizar la métrica PW y simplemente utilizando un valor de corte para la IPD y no las estructuras de datos como se describe en el presente documento. Por ejemplo, se extrajeron 3 nt en dirección ascendente y en dirección descendente que flanquean un CpG que se está interrogando. Las IPD de ese CpG se estratificaron en diferentes grupos (4096 grupos para las 6 posiciones) dependiendo del contexto de las secuencias flanqueantes de 6 nt (es decir, 3 nt en dirección ascendente y en dirección descendente, respectivamente) que estaban centradas en ese CpG. Las IPD entre CpG metilados y no metilados dentro del mismo motivo de secuencia se estudiaron utilizando ROC. Por ejemplo, se compararon las IPD de CpG en el motivo "AATCGGAC" no metilado y el motivo "AAT^{m}CGGAC" metilado, mostrando un AUC de 0.48. Por lo tanto, el uso de los valores de corte en un grupo de secuencia particular tuvo un desempeño pobre en relación con la realización que utiliza varios

Sólo con fines ilustrativos, la FIG. 27 muestra el rendimiento del análisis IPD basado en motivos anterior (Beckmann et al. BMC Bioinformatics. 2014) para la detección de metilación sin utilizar la señal de anchura de pulso. Los gráficos de barras verticales representan las AUC promediadas en diferentes motivos k-mer que flanquean los sitios CpG que se están estudiando (es decir, el número de bases que rodean los sitios CpG que se están interrogando). La FIG. 27 mostró que las AUC promediadas para los poderes discriminativos basados en IPD entre citosinas metiladas y no metiladas en diferentes motivos k-mer (por ejemplo, 2-mero, 3-mero, 4-mero, 6-mero, 8-mero, 10-mero, 15-mero, 20-mero que rodean los sitios CpG en cuestión) fue menor al 60 %. Estos resultados sugirieron que la consideración de la IPD del nucleótido candidato en un contexto de motivo determinado sin tener en cuenta las IPD de los nucleótidos vecinos (Flusberg et al., 2010) sería inferior a los métodos divulgados en el presente documento para la determinación de la metilación de CpG.

También probamos el método presente en Flusberg et al. estudio (Flusberg et al., 2010). Analizamos un total de 5,948,348 segmentos de ADN que estaban 2 nt en dirección ascendente y 6 nt en dirección descendente de una citosina que se sometió a análisis de metilación. Había 2,828,848 segmentos metilados y 3,119,500 segmentos no metilados. Como se muestra a modo de ilustración en la FIG. 28, se encontró que las señales deducidas del análisis de componentes principales con el uso de IPD y PW se superponían en gran medida entre fragmentos con citosinas metiladas (mC) y citosinas no metiladas (C), lo que sugiere que el método descrito por Flusberg et al carece prácticamente de precisión significativa. Estos resultados sugirieron que el análisis de componentes principales, que combinaba linealmente los valores de PW e IPD en las bases y bases vecinas, como se utilizó en el estudio de Flusberg et al. (Flusberg et al., 2010), no pudo diferenciar de manera confiable o significativa la 5-metilcitosina y citosinas no metiladas.

La FIG. 29 muestra que el AUC del método basado en el análisis de componentes principales para el cual se utilizaron dos componentes principales en el estudio de Flusberg et al (Flusberg et al., 2010) que involucraba IPD y PW fue

mucho menos preciso (AUC: 0.55) que el enfoque basado en una red neuronal convolucional que involucra IPD y PW, así como el contexto de secuencia como se muestra en nuestra divulgación de la invención (AUC: 0.94).

C. Otros modelos matemáticos/estadísticos

En otra realización, se podrían entrenar otros modelos matemáticos/estadísticos, por ejemplo que incluyen, pero no se limitan a, un bosque aleatorio y una regresión logística, adaptando las características desarrolladas anteriormente. En cuanto al modelo CNN, los conjuntos de datos de entrenamiento y prueba se construyeron a partir del ADN con tratamiento M.SssI (metilado) y amplificación por PCR (no metilado), que se utilizaron para entrenar un bosque aleatorio (Breiman, 2001). En este análisis de bosque aleatorio, describimos cada nucleótido con 6 características: IPD, PW y un vector binario de 4 componentes que codifica la identidad de la base. En dicho vector binario, A, C, G y T se codificaron con [1,0,0,0], [0,1,0,0], [0,0,1,0] y [0,0,0,1], respectivamente. Para cada sitio CpG que se analiza, incorporamos la información de sus 10 nt en dirección ascendente y en dirección descendente en ambas hebras, formando un vector de 252 dimensiones (252-D), donde cada característica representa una dimensión. El conjunto de datos de entrenamiento descrito anteriormente con los vectores 252-D se utilizó para entrenar un modelo de bosque aleatorio, así como el modelo de regresión logística. El modelo entrenado se utilizó para predecir los estados de metilación en un conjunto de datos de prueba independiente. El bosque aleatorio estaba compuesto por 100 árboles de decisión. Durante la construcción del árbol, se utilizaron muestras de arranque. Al dividir el nodo de cada árbol de decisión, se empleó la impureza de Gini para determinar la mejor división, y se consideraría un máximo de 15 características en cada división. Además, cada hoja del árbol de decisión debía contener al menos 60 muestras.

Las FIG. 30A y FIG. 30B muestran el rendimiento de un método de acuerdo con la invención que utiliza un bosque aleatorio y regresión logística para la predicción de la metilación. La FIG. 30A muestra los valores de AUC en el conjunto de datos de entrenamiento para CNN, bosque aleatorio y regresión logística. La FIG. 30B muestra los valores de AUC en el conjunto de datos de prueba para CNN, bosque aleatorio y regresión logística. El AUC de un método que utiliza bosque aleatorio alcanzó 0.93 y 0.86 en el conjunto de datos de entrenamiento y prueba, respectivamente.

El conjunto de datos de entrenamiento descrito con los mismos vectores 252-D se utilizó para entrenar un modelo de regresión logística. El modelo entrenado se utilizó para predecir los estados de metilación en un conjunto de datos de prueba independiente. Se ajustó un modelo de regresión logística con regularización L2 (Ng and Y., 2004) al conjunto de datos de entrenamiento. Como se muestra en la FIG. 30A y FIG. 30B, el AUC de un método que utiliza regresión logística alcanzaría 0.87 y 0.83 en el conjunto de datos de entrenamiento y prueba, respectivamente.

Por lo tanto, estos resultados sugirieron que ciertos modelos (por ejemplo, pero no limitados a, el bosque aleatorio y la regresión logística) distintos de CNN se podrían utilizar para el análisis de metilación utilizando las características y protocolos analíticos que desarrollamos en esta divulgación. Estos resultados también sugirieron que CNN implementado de acuerdo con las realizaciones en esta divulgación con un AUC de 0.90 en el conjunto de datos de prueba (FIG. 30B) fue superior tanto al bosque aleatorio (AUC: 0.86) como a la regresión logística (AUC: 0.83).

D. Determinación de modificaciones de 6mA de ácidos nucleicos.

Además del CpG metilado, los métodos descritos en el presente documento también pueden detectar otras modificaciones de bases de ADN. Por ejemplo, se puede detectar adenina metilada, incluso en forma de 6mA.

1. Detección de 6mA utilizando características cinéticas y contexto de secuenciación

Para evaluar el rendimiento y la utilidad de las realizaciones de la invención divulgadas para la determinación de la metilación de ácidos nucleicos, analizamos adicionalmente la metilación de N6-adenina (6mA). En una realización, se amplificó aproximadamente 1 ng de ADN humano (por ejemplo, extraído de tejidos placentarios) para obtener 100 ng de producto de ADN a través de la amplificación del genoma completo con adenina no metilada (uA), citosina no metilada (C), guanina no metilada (G) y timina (T).

La FIG. 31A muestra un ejemplo de un enfoque para generar moléculas con adeninas no metiladas mediante amplificación del genoma completo. En la figura, “uA” indica una adenina no metilada y “mA” indica una adenina metilada. La amplificación del genoma completo se realizó utilizando hexámeros aleatorios modificados con tiofosfato resistentes a exonucleasa como cebadores, que se unen aleatoriamente a un genoma, permitiendo que la polimerasa (por ejemplo, ADN polimerasa Phi29) amplifique el ADN (por ejemplo, mediante amplificación lineal isotérmica). En el estadio 3102, se desnaturaliza el ADN de hebra doble. En el estadio 3106, la reacción de amplificación se inicia cuando se hibridan varios hexámeros aleatorios (por ejemplo, 3110) con el ADN de plantilla desnaturalizado (es decir, ADN de hebra sencilla). Como se muestra en 3114, cuando la síntesis de ADN mediada por hexámero de la hebra 3118 avanzó en la dirección 5’ a 3’ y llegó al siguiente sitio de síntesis de ADN mediada por hexámero, la polimerasa desplazó la hebra de ADN recién sintetizada (3122) y continuó la extensión de la hebra. Las hebras desplazadas se convirtieron en plantillas de ADN de hebra sencilla para unirse nuevamente a hexámeros aleatorios y podrían iniciar una nueva síntesis de ADN. La hibridación repetida de hexámeros y el desplazamiento de hebras en un proceso isotérmico darían como resultado un alto rendimiento de productos de ADN amplificados. Esta amplificación descrita en el presente documento se puede incluir en la técnica de amplificación por desplazamiento múltiple (MDA).

Los productos de ADN amplificados se fragmentaron adicionalmente en, por ejemplo, pero no se limitan a, fragmentos con tamaños de 0 bp, 200 bp, 300 bp, 400 bp, 500 bp, 600 bp, 700 bp, 800 bp, 900 bp, 1 kb, 5 kb, 10 kb, 20 kb, 30 kb, 40 kb, 50 kb, 60 kb, 70 kb, 80 kb, 90 kb, 100 kb u otros rangos de tamaño deseados. El proceso de fragmentación puede incluir digestión enzimática, nebulización, cizallamiento hidrodinámico y sonicación, etc. Como resultado, la metilación original, tal como la de 6mA, se puede eliminar casi por completo mediante la amplificación del genoma completo con A no metilado (uA). La FIG. 31A muestra posibles fragmentos (3126, 3130 y 3134) de los productos de ADN, teniendo ambas hebras A no metilado. Dichos productos de ADN amplificados de genoma completo sin mA se sometieron a secuenciación en tiempo real de única molécula para generar un conjunto de datos de uA.

La FIG. 31B muestra un ejemplo de un enfoque para generar moléculas con adeninas metiladas mediante amplificación del genoma completo. En la figura, "uA" indica una adenina no metilada y "mA" indica una adenina metilada. Se amplificó aproximadamente 1 ng de ADN humano para obtener 10 ng de producto de ADN a través de la amplificación del genoma completo con 6mA y C, G y T no metilados. Las adeninas metiladas se pueden producir mediante una serie de reacciones químicas (JD Engel et al. J Biol Chem. 1978;253:927-34). Como se ilustra en la FIG. 31B, la amplificación del genoma completo se realizó utilizando hexámeros aleatorios modificados con tiofosfato resistentes a exonucleasa como cebadores que se unen aleatoriamente sobre un genoma, permitiendo que la polimerasa (por ejemplo, ADN polimerasa Phi29) amplifique el ADN (por ejemplo, mediante amplificación lineal isotérmica), similar a la FIG. 31A. Los hexámeros aleatorios modificados con tiofosfato resistentes a exonucleasas son resistentes a la actividad exonucleasa 3'→5' de las ADN polimerasas de corrección. Por tanto, durante la amplificación, los hexámeros aleatorios estarán protegidos de la degradación.

La reacción de amplificación se inició cuando se hibridaron varios hexámeros aleatorios con el ADN de plantilla desnaturalizado (es decir, ADN de hebra sencilla). Cuando la síntesis de ADN mediada por hexámero avanzó en la dirección 5' a 3' y llegó al siguiente sitio de síntesis de ADN mediada por hexámero, la polimerasa desplazó la hebra de ADN recién sintetizada y continuó la extensión de la hebra. Las hebras desplazadas se convirtieron en plantillas de ADN de hebra sencilla para unirse nuevamente a hexámeros aleatorios e iniciar una nueva síntesis de ADN. La hibridación repetido de hexámeros y el desplazamiento de hebras en un proceso isotérmico darían como resultado un alto rendimiento de productos de ADN amplificados.

Los productos de ADN amplificados se fragmentaron adicionalmente en, por ejemplo, pero no se limita a, fragmentos con tamaños de 100 bp, 200 bp, 300 bp, 400 bp, 500 bp, 600 bp, 700 bp, 800 bp, 900 bp, 1 kb, 5 kb, 10 kb, 20 kb, 30 kb, 40 kb, 50 kb, 60 kb, 70 kb, 80 kb, 90 kb, 100 kb u otras combinaciones de longitud. Como se muestra en la FIG. 31B, los productos de ADN amplificados incluirían diferentes formas de patrones de metilación en los sitios de adenina en cada hebra. Por ejemplo, ambas hebras de una molécula de hebra doble pueden estar metiladas con respecto a las adeninas (Molécula I), que se generarían cuando dos hebras derivan de la síntesis de ADN durante la amplificación del genoma completo.

Como otro ejemplo, una hebra de una molécula de hebra doble puede contener patrones de metilación entrelazados a través de sitios de adenina (Molécula II). Un patrón de metilación entrelazado se define como aquel que incluye una mezcla de bases metiladas y no metiladas presentes en una hebra de ADN. En los siguientes ejemplos, utilizamos un patrón de metilación de adenina entrelazado que incluye una mezcla de adeninas metiladas y no metiladas presentes en una hebra de ADN. Este tipo de molécula de hebra doble (Molécula II) posiblemente se generaría porque un hexámero no metilado que contenía adeninas no metiladas se unió a una hebra de ADN e inició la extensión del ADN. Se secuenciaría dicho producto de ADN amplificado que contiene el hexámero con adeninas no metiladas. Alternativamente, este tipo de molécula de hebra doble (Molécula II) se iniciaría mediante ADN fragmentado a partir de ADN de plantilla original que contiene adeninas no metiladas, ya que dicho ADN fragmentado se podría unir a una hebra de ADN como cebador. Se secuenciaría dicho producto de ADN amplificado que contuviera parte del ADN original con adeninas no metiladas en una hebra. Como los cebadores hexámeros no metilados son sólo una pequeña porción de las hebras de ADN resultantes, la mayoría de los fragmentos seguirán conteniendo 6mA.

Como otro ejemplo, una hebra de una molécula de ADN de hebra doble puede estar metilada a través de sitios de adenina, pero la otra hebra puede estar sin metilar (Molécula III). Este tipo de molécula de hebra doble se puede generar cuando se proporciona una hebra de ADN original sin adeninas metiladas como molécula de ADN de plantilla para producir una nueva hebra con adeninas metiladas.

Ambas hebras pueden estar sin metilar (Molécula IV). Este tipo de molécula de hebra doble se puede deber a la rehibridación de dos hebras de ADN originales sin adeninas metiladas.

El proceso de fragmentación puede incluir digestión enzimática, nebulización, cizallamiento hidrodinámico y sonicación, etc. Dichos productos de ADN amplificados del genoma completo pueden estar predominantemente metilados en términos de sitios A. Este ADN con mA se sometió a secuenciación en tiempo real de única molécula para generar un conjunto de datos de mA.

Para el conjunto de datos de uA, secuenciamos 262.608 moléculas con una mediana de 964 bp de longitud utilizando secuenciación en tiempo real de única molécula. La profundidad media de la sublectura fue de 103 x. De las sublecturas, el 48 % se podría alinear con un genoma de referencia humano utilizando el alineador BWA (Li H et al. Bioinformatics. 2009;25:1754-60). Como un ejemplo, se podría emplear el sistema Sequel II (Pacific Biosciences) para

llevar a cabo la secuenciación en tiempo real de única molécula. Las moléculas de ADN fragmentadas se sometieron a la construcción de una plantilla de secuenciación en tiempo real de única molécula (SMRT) utilizando un Kit SMRTbell Express Template Prep 2.0 (Pacific Biosciences). Las condiciones de hibridación del cebador de secuenciación y de unión de la polimerasa se calcularon con el software SMRT Link v8.0 (Pacific Biosciences). Brevemente, el cebador de secuenciación v2 se hibridó con la plantilla de secuenciación y luego se unió una polimerasa a las plantillas utilizando un Kit de Control Interno y Unión Sequel II 2.0 (Pacific Biosciences). La secuenciación se realizó en un Sequel II SMRT Cell 8M. Las películas de secuenciación se recopilaron en el sistema Sequel II durante 30 horas con un Kit de Secuenciación Sequel II 2.0 (Pacific Biosciences).

Para el conjunto de datos de mA, secuenciamos 804,469 moléculas con una mediana de 826 bp de longitud utilizando secuenciación en tiempo real de única molécula. La profundidad media de la sublectura fue de 34 x. De las sublecturas, el 27 % se podría alinear con un genoma de referencia humano utilizando el alineador BWA (Li H et al. Bioinformatics. 2009;25:1754-60).

En una realización, las características cinéticas que incluyen, pero no se limitan a, IPD y PW se analizaron de una manera específica de hebra. Para los resultados de secuenciación derivados de la hebra de Watson, se utilizaron 644,318 sitios A sin metilación seleccionados aleatoriamente del conjunto de datos uA y 718,586 sitios A con metilación seleccionados aleatoriamente del conjunto de datos mA para constituir un conjunto de datos de entrenamiento. Dicho conjunto de datos de entrenamiento se utilizó para establecer los modelos de clasificación y/o umbrales para diferenciar entre adeninas metiladas y no metiladas. Se constituyó un conjunto de datos de prueba a partir de 639,702 sitios A sin metilación y 723,320 sitios A con metilación. Dicho conjunto de datos de prueba se utilizó para validar el rendimiento de un modelo/umbral deducido de un conjunto de datos de entrenamiento.

Analizamos los resultados de secuenciación que se originan a partir de las hebras de Watson. La FIG. 32A muestra valores de duración de interpulso (IPD) en todo el conjunto de datos de entrenamiento de los conjuntos de datos de uA y mA. Para el conjunto de datos de entrenamiento, se observó que los valores de IPD en los sitios A secuenciados eran más altos en el conjunto de datos de mA (mediana: 1.09; rango: 0 - 9.52) que en el conjunto de datos de uA (mediana: 0.20; rango: 0 - 9.52) (valor de $P < 0.0001$; prueba U de Mann Whitney).

La FIG. 32B muestra IPD para el conjunto de datos de prueba de los conjuntos de datos de uA y mA. Cuando estudiamos los valores de IPD en los sitios A secuenciados en el conjunto de datos de prueba, observamos que los valores de IPD eran más altos en el conjunto de datos de mA que en el conjunto de datos de uA (mediana 1.10 frente a 0.19; valor de $P < 0.0001$; prueba U de Mann Whitney).

La FIG. 32C muestra el área bajo la curva característica operativa del receptor (ROC) utilizando el valor de corte de IPD. La tasa positiva verdadera está en el eje y y la tasa positiva falsa está en el eje x. El área bajo la curva característica operativa del receptor (AUC) para diferenciar bases A secuenciadas en moléculas de ADN de plantilla con y sin metilación utilizando los valores de IPD correspondientes fue de 0.86 tanto para los conjuntos de datos de entrenamiento como para los de prueba.

Además de los resultados de la hebra de Watson, analizamos los resultados de secuenciación que se originan en las hebras de Crick. La FIG. 33A muestra valores de IPD en todo el conjunto de datos de entrenamiento de conjuntos de datos de uA y mA. Para el conjunto de datos de entrenamiento, se observó que los valores de IPD en los sitios A secuenciados eran más altos en el conjunto de datos de mA (mediana: 1.10 rango: 0 - 9.52) que en el conjunto de datos de uA (mediana: 0.19; rango: 0 - 9.52) (valor de $P < 0.0001$; prueba U de Mann Whitney).

La FIG. 34B muestra valores de IPD para el conjunto de datos de prueba de conjuntos de datos de uA y mA. Los valores de IPD más altos en los sitios A secuenciados también se observaron en el conjunto de datos de mA para el conjunto de datos de prueba, en comparación con el conjunto de datos de uA (mediana 1.10 versus 0.19; valor de $P < 0.0001$; prueba U de Mann Whitney).

La FIG. 33C muestra el área bajo la curva ROC. La tasa positiva verdadera está en el eje y y la tasa positiva falsa está en el eje x. El área bajo el valor de la curva ROC (AUC) para diferenciar bases A secuenciadas en moléculas de ADN de plantilla con y sin metilación utilizando los valores de IPD correspondientes fue de 0.86 y 0.87 para los conjuntos de datos de entrenamiento y prueba, respectivamente.

La FIG. 34 muestra una ilustración para la determinación de 6mA de la hebra de Watson utilizando una ventana de medición de acuerdo con realizaciones de la presente invención. Dicha ventana de medición puede incluir características cinéticas como IPD y PW y un contexto de secuencia cercano. La determinación de 6mA se puede realizar de manera similar a la determinación de CpG metilado.

La FIG. 35 muestra una ilustración para la determinación de 6mA de la hebra de Crick utilizando una ventana de medición de acuerdo con realizaciones de la presente invención. Dicha ventana de medición puede incluir características cinéticas como IPD y PW y un contexto de secuencia cercano.

Como un ejemplo, se utilizaron 10 bases de cada lado de la base A secuenciada en un ADN de plantilla que estaba siendo interrogado para construir una ventana de medición. Los valores de las características, que incluyen las IPD, PW y contexto de secuencia, se utilizaron para entrenar un modelo utilizando una red neuronal convolucional (CNN)

de acuerdo con los métodos divulgados en el presente documento. En otras realizaciones, los modelos estadísticos pueden incluir, pero no se limitan a, regresión lineal, regresión logística, red neuronal recurrente profunda (por ejemplo, memoria a largo plazo, LSTM), clasificador de Bayes, modelo oculto de Markov (HMM), análisis discriminante lineal (LDA), agrupamiento k-medias, agrupamiento espacial de aplicaciones con ruido basada en densidad (DBSCAN), algoritmo de bosque aleatorio y máquina de vectores de soporte (SVM), etc.

Las FIG. 36A y FIG. 36B muestran la probabilidad determinada de ser metilado para bases A secuenciadas de la hebra de Watson entre conjuntos de datos de uA y mA utilizando un modelo CNN basado en ventana de medición de acuerdo con la invención. La FIG. 36A muestra que se aprendió de un modelo CNN a partir de un conjunto de datos de entrenamiento. Como un ejemplo, el modelo CNN hizo uso de dos capas de convolución 1D (cada una con 64 filtros con un tamaño de núcleo de 4 seguidas de una capa ReLU (unidad lineal rectificada)), seguida de una capa de exclusión con una tasa de exclusión de 0.5. Se utilizó una capa de agrupación máxima con un tamaño de agrupación de 2. Luego fluyó en dos capas de convolución 1D (cada una con 128 filtros con un tamaño de núcleo de 2 seguidas de una capa ReLU), utilizando además una capa de exclusión con una tasa de exclusión de 0.5. Se utilizó una capa de agrupación máxima con un tamaño de agrupación de 2. Finalmente, una capa completamente conectada con 10 neuronas seguida de una capa ReLU, con una capa de salida con una neurona seguida de una capa sigmoidea, produjo de esta manera la probabilidad de metilación. Las otras configuraciones de capas, filtros y tamaños de granos se podrían adaptar, por ejemplo, como se describe en el presente documento para otras metilaciones (por ejemplo, CpG). En este conjunto de datos de entrenamiento sobre los resultados de secuenciación de la hebra de Watson, utilizamos 644,318 y 718,586 bases A de bibliotecas metiladas y no metiladas.

Basado en el modelo CNN, para los datos relacionados con la hebra de Watson, las bases A secuenciadas en moléculas de ADN plantilla de la base de datos mA dieron lugar a una probabilidad mucho mayor de metilación tanto en los conjuntos de datos de entrenamiento como en los de prueba, en comparación con aquellas bases A presentes. en el conjunto de datos uA (valor de p <0.0001; prueba U de Mann Whitney). Para el conjunto de datos de entrenamiento, la probabilidad mediana de metilación en los sitios A en el conjunto de datos uA fue 0.13 (rango intercuartil, IQR: 0.09 - 0.15), mientras que ese valor en el conjunto de datos mA fue 1.000 (IQR: 0.998 - 1.000).

La FIG. 36A muestra la probabilidad de metilación determinada para el conjunto de datos de prueba. Para el conjunto de datos de prueba, la probabilidad mediana de metilación en los sitios A en el conjunto de datos uA fue 0.13 (IQR: 0.10 - 0.15), mientras que ese valor en el conjunto de datos mA fue 1.000 (IQR: 0.997 - 1.000). Las FIG. 36A y 36B muestran que se puede entrenar un modelo CNN basado en una ventana de medición para detectar la metilación en un conjunto de datos de prueba.

La FIG. 37 es una curva ROC para la detección de 6mA utilizando un modelo CNN basado en ventana de medición para bases A secuenciadas de la hebra de Watson de acuerdo con la invención. La tasa positiva verdadera está en el eje y y la tasa positiva falsa está en el eje x. La figura muestra que el valor de AUC para diferenciar sitios A secuenciados con y sin metilación utilizando un modelo CNN fue de 0.94 y 0.93 para conjuntos de datos de entrenamiento y prueba que consistían en los resultados de secuenciación de la hebra de Watson, respectivamente. Sugirió que era factible utilizar la divulgación en el presente documento para determinar los estados de metilación en los sitios A utilizando datos de la hebra de Watson. Si utilizamos la probabilidad de metilación determinada de 0.5 como valor de corte, se podría lograr una especificidad del 99.3 % y una sensibilidad del 82.6 % para la detección de 6mA. La FIG. 37 muestra que se puede utilizar un modelo CNN basado en ventana de medición para detectar 6mA con alta especificidad y sensibilidad. La precisión del modelo se puede comparar con una técnica que utiliza solo una métrica IPD.

La FIG. 38 muestra una comparación de rendimiento entre la detección de 6mA basada en métrica IPD y una detección de 6mA basada en una ventana de medición. La sensibilidad se grafica en el eje y y la especificidad se grafica en el eje x. La FIG. 38 muestra que el rendimiento utilizando la clasificación de 6mA basada en ventana de medición de acuerdo con la divulgación de la invención en el presente documento (AUC: 0.94) fue superior al método convencional que utiliza solo la métrica IPD (AUC: 0.87) (valor de P < 0.0001; prueba de DeLong). El modelo CNN basado en ventana de medición superó a la detección basada en métricas IPD.

Las FIG. 39A y 39B muestran la probabilidad determinada de ser metilados para aquellas bases A secuenciadas de la hebra de Crick entre conjuntos de datos de uA y mA utilizando el modelo CNN basado en ventana de medición de la invención. La FIG. 39A muestra el conjunto de datos de entrenamiento y la FIG. 39B muestra el conjunto de datos de prueba. Ambas figuras representan la probabilidad de metilación en el eje y. Las FIG. 39A y 39B muestran que, sobre la base del modelo CNN, para los datos relacionados con la hebra de Crick, las bases A secuenciadas en moléculas de ADN plantilla de la base de datos mA dieron lugar a una probabilidad mucho mayor de metilación tanto en el conjunto de datos de entrenamiento como en el de prueba, en comparación con aquellas bases A presentes en la base de datos uA (valor de P <0.0001; prueba U de Mann-Whitney).

La FIG. 40 muestra el rendimiento de la detección de 6mA utilizando el modelo CNN basado en ventana de medición en bases A secuenciadas de la hebra de Crick de acuerdo con la invención. La tasa positiva verdadera está en el eje y. La tasa positiva falsa está en el eje x. La FIG. 40 muestra que el valor de AUC para diferenciar sitios A secuenciados con y sin metilación utilizando un modelo CNN fue de 0.95 y 0.94 para conjuntos de datos de entrenamiento y prueba que consistían en los resultados de secuenciación de la hebra de Crick, respectivamente. También se mostró que el

rendimiento utilizando el enfoque CNN divulgado en el presente documento (AUC: 0.94) es superior al que utiliza únicamente la métrica IPD (0.87) (valor de P <0.0001). Los resultados sugirieron que era factible utilizar la divulgación del presente documento para determinar los estados de metilación en los sitios A utilizando datos de la hebra de Crick. Si utilizamos la probabilidad de metilación determinada de 0.5 como valor de corte, se podría lograr una especificidad del 99.3 % y una sensibilidad del 83.0 % para la detección de 6mA. La FIG. 40 muestra que se puede utilizar un modelo CNN basado en ventana de medición para detectar 6mA con alta especificidad y sensibilidad.

La FIG. 41 muestra ejemplos de estados de metilación en bases A en una molécula que incluye las hebras de Watson y Crick. Los puntos blancos representan adeninas no metiladas. Los puntos rellenos negros representan adeninas metiladas. Las líneas horizontales con puntos representan una hebra de una molécula de ADN de hebra doble. La molécula 1 muestra que se determina que tanto las hebras de Watson como las de Crick no están metiladas en las bases A. La molécula 2 muestra que la hebra de Watson estaba casi completamente desmetilada, mientras que la hebra de Crick estaba casi toda metilada. La molécula 3 muestra que se determinó que tanto las hebras de Watson como las de Crick estaban casi todas metiladas en las bases A.

2. Entrenamiento potenciado utilizando un conjunto de datos selectivo

Como se muestra en la FIGS. 36A, 36B, 39A y 39B, hubo una distribución bimodal de probabilidad de metilación entre bases A secuenciadas en moléculas de ADN de plantilla en el conjunto de datos mA. En otras palabras, existían algunas moléculas con señales de uA en el conjunto de datos de mA. Esto se evidenció aún más por la existencia de moléculas completamente no metiladas y moléculas hemimetiladas en el conjunto de datos de mA (FIG. 41). Una posible razón puede ser que las moléculas con uA en las plantillas de ADN seguirían representando una porción considerable del conjunto de datos de mA después de la amplificación del genoma completo, ya que las moléculas con 6mA conducirían a una menor eficiencia de la amplificación del ADN durante la etapa de amplificación del genoma completo. Esta explicación fue respaldada por el hecho de que 1 ng de ADN genómico amplificado con 6mA solo daría lugar a 10 ng de productos de ADN, mientras que 1 ng de ADN genómico amplificado con A no metilado daría lugar a 100 ng de productos de ADN en las mismas condiciones de amplificación. Por lo tanto, para el conjunto de datos mA, las moléculas de ADN de plantilla originales cuyas adeninas usualmente no están metiladas (por ejemplo, 0.051 %) (Xiao CL et al. Mol Cell. 2018;71:306-318) representarían aproximadamente el 10 % del total de adeninas.

En una realización, cuando se intenta entrenar un modelo CNN para diferenciar entre mA y uA, se utilizarían selectivamente aquellas bases A con valores de IPD relativamente más altos en el conjunto de datos de mA para reducir la influencia de los datos de uA en el entrenamiento del modelo para la detección de mA. Sólo se pueden utilizar bases A con valores de IPD superiores a un determinado valor de corte. El valor de corte puede corresponder a un percentil. En una realización, se utilizarían aquellas bases A en un conjunto de datos de mA con valores de IPD mayores que el valor en el percentil 10. En algunas realizaciones, se utilizarían aquellos A con valores de IPD mayores que el valor en el percentil 1, 5, 15, 20, 30, 40, 50, 60, 70, 80, 90 o 95. El percentil se puede basar en datos de todas las moléculas de ácido nucleico en una muestra de referencia o en múltiples muestras de referencia.

La FIG. 42 muestra el rendimiento con entrenamiento mejorado mediante el uso selectivo de bases A en un conjunto de datos de mA con valores de IPD superiores a su percentil 10. La FIG. 42 muestra la tasa positiva verdadera en el eje y y la tasa positiva falsa en el eje x. La figura muestra que con el uso de bases A en el conjunto de datos de mA con valores de IPD superiores al percentil 10 para entrenar un modelo CNN, el AUC para diferenciar entre bases de mA y uA aumentaría a 0.98, lo que era superior al modelo (AUC : 0.94) entrenado por datos sin la selección de acuerdo con los valores de IPD antes del entrenamiento. Sugirió que la selección de sitios de mA utilizando valores de IPD para crear un conjunto de datos de entrenamiento ayudaría a mejorar el poder discriminativo.

Para confirmar aún más la existencia de moléculas con bases uA en el conjunto de datos mA, planteamos la hipótesis de que el porcentaje de uA en el conjunto de datos mA se enriquecería en aquellos pocillos con más sublecturas, ya que los 6mA presentes en una molécula ralentizarían el alargamiento de la polimerasa cuando generando una nueva hebra, en comparación con una molécula sin 6mA.

La FIG. 43 muestra un gráfico de los porcentajes de adeninas no metiladas en el conjunto de datos de mA frente al número de sublecturas en cada pocillo. El eje y muestra el porcentaje de uA en el conjunto de datos de mA. El eje x muestra el número de sublecturas en cada pocillo. El conjunto de datos de prueba se volvió a analizar utilizando el modelo potenciado que se entrenó al utilizar sitios mA después de eliminar los sitios A cuyos valores de IPD estaban por debajo del percentil 10. Se observó un aumento gradual de uA (es decir, que se eleva desde 14.6 hasta 55.05 %) a medida que aumentó el número de sublecturas por pocillo, que incluyen desde 1 hasta 10 sublecturas por pocillo de secuenciación, de 10 a 20 sublecturas por pocillo y de 40 a 50 sublecturas por pocillo , 60 a 70 sublecturas por pocillo y más de 70. Por lo tanto, los pocillos que tienen una gran cantidad de sublecturas tienden a tener mA bajos. La metilación de A puede retrasar la progresión de la reacción de secuenciación. Por lo tanto, es más probable que los pocillos de secuenciación con una profundidad de sublectura alta no estén metilados con respecto a A. Este comportamiento se puede aprovechar para la detección de moléculas no metiladas utilizando un valor de corte para el número de sublecturas asociadas con la molécula, por ejemplo, se pueden identificar más de 70 sublecturas como mayoritariamente no metiladas.

La FIG. 44 muestra patrones de metiladenina entre las hebras de Watson y Crick de una molécula de ADN de hebra doble en un conjunto de datos de prueba. La metilación de A es asimétrica y, por tanto, el comportamiento es diferente entre las dos hebras. La mayoría de las moléculas estaban metiladas debido a la incorporación de mA, con algo de A no metilado residual. El eje y muestra el nivel de metiladenina de la hebra de Crick. El eje x muestra el nivel de metiladenina de la hebra de Watson. Cada punto representa una molécula de hebra doble. Utilizando el modelo potenciado que fue entrenado por sitios mA seleccionados, las moléculas de hebra doble se pueden clasificar en diferentes grupos de acuerdo con el nivel de metilación de cada hebra de la siguiente manera:

(a) Para una molécula de ADN de hebra doble, los niveles de metiladenina de las hebras de Watson y Crick fueron mayores que 0.8. Dicha molécula de hebra doble se definió como una molécula completamente metilada con respecto a los sitios de adenina (FIG. 44, área A). El nivel de metiladenina de una hebra se definió como el porcentaje de sitios A que se determinó que estaban metilados entre el total de sitios A en esa hebra.

(b) Para una molécula de ADN de hebra doble, el nivel de metiladenina de una hebra era mayor que 0.8 mientras que la otra hebra era menor que 0.2. Dicha molécula se definió como molécula hemimetilada con respecto a los sitios de adenina (FIG. 44, áreas B1 y B2).

(c) Para una molécula de ADN de hebra doble, los niveles de metiladenina de las hebras de Watson y Crick eran menores a 0.2. Dicha molécula de hebra doble se definió como una molécula completamente no metilada con respecto a los sitios de adenina (FIG. 44, área C).

(d) Para una molécula de ADN de hebra doble, los niveles de metiladenina de las hebras de Watson y Crick no pertenecían a los grupos a, b y c. Dicha molécula de hebra doble se definió como una molécula con patrones de metilación entrelazados con respecto a los sitios de adenina (FIG. 44, área D). Los patrones de metilación entrelazados se definieron como una mezcla de adeninas metiladas y no metiladas presentes en una hebra de ADN.

En algunas otras realizaciones, los valores de corte de los niveles de metiladenina para definir la hebra no metilada pueden ser, pero no se limitan a, menos de 0.01, 0.05, 0.1, 0.2, 0.3, 0.4 y 0.5. Los valores de corte de los niveles de metiladenina para definir la hebra metilada serían, pero no se limitan a, más de 0.5, 0.6, 0.7, 0.8, 0.9, 0.95 y 0.99.

La FIG. 45 es una tabla que muestra el porcentaje de moléculas completamente no metiladas, moléculas hemimetiladas, moléculas completamente metiladas y moléculas con patrones de metiladenina entrelazados en conjuntos de datos de entrenamiento y prueba. Las moléculas en el conjunto de datos de prueba se pueden clasificar en moléculas completamente no metiladas (7.0 %) con respecto a los sitios de adenina, moléculas hemimetiladas (9.8 %), moléculas completamente metiladas (79.4 %) y moléculas con patrones de metiladenina entrelazados (3.7 %). Estos resultados fueron comparables a los resultados mostrados en el conjunto de datos de entrenamiento, para los cuales había moléculas completamente no metiladas (7.0 %) con respecto a los sitios de adenina, moléculas hemimetiladas (10.0 %), moléculas completamente metiladas (79.4 %) y moléculas con patrones de metiladenina entrelazados (3.6 %).

La FIG. 46 ilustra ejemplos representativos de moléculas con moléculas completamente no metiladas con respecto a sitios de adenina, moléculas hemimetiladas, moléculas completamente metiladas y moléculas con patrones de metiladenina entrelazados. Los puntos blancos representan adeninas no metiladas. Los puntos rellenos negros representan adeninas metiladas. Las líneas horizontales con puntos representan una hebra de una molécula de ADN de hebra doble.

En realizaciones, se puede mejorar el rendimiento en la diferenciación entre adeninas metiladas y no metiladas al aumentar la pureza de las bases de 6mA que se utilizaron para entrenar un modelo CNN. Con este fin, se puede aumentar la duración de la reacción de amplificación del ADN de tal manera que el aumento de productos de ADN recién producidos pueda diluir el efecto de las adeninas no metiladas aportadas a partir de plantillas de ADN originales. En otras realizaciones, se pueden incorporar bases biotiniladas durante la amplificación del ADN con 6mA. Los productos de ADN recién producidos con 6mA se pueden extraer y enriquecer utilizando perlas magnéticas recubiertas de estreptavidina.

3. Usos de perfiles de metilación de 6 mA

La modificación del ADN 6mA está presente en los genomas de bacterias, arqueas, protistas y hongos (Didier W et al. Nat Rev Micorbiol. 2009;4: 183-192). También se informó que existían 6mA en el genoma humano, lo que representa el 0.051 % del total de adeninas (Xiao CL et al. Mol Cell. 2018;71:306-318). Teniendo en cuenta el bajo contenido de 6mA en un genoma humano, en una realización, se puede crear un conjunto de datos de entrenamiento al ajustar la relación de 6mA en la mezcla de dNTP (N representa A, C, G y T no modificados) en la etapa de amplificación del genoma completo. Por ejemplo, se podría utilizar la relación de 6mA a dNTP de 1:10, 1:100, 1:1000, 1:10000, 1:100000 o 1:1000000. En otra realización, se puede utilizar adenina ADN metiltransferasa M. EcoGII para crear un conjunto de datos de entrenamiento de 6mA.

La cantidad de 6mA fue menor en los tejidos de cáncer gástrico y de hígado, y esta regulación a la baja de 6mA se correlacionó con una mayor tumorigénesis (Xiao CL et al. Mol Cell. 2018;71:306-318). Por otro lado, se informó que estaban presentes niveles más altos de 6mA en el glioblastoma (Xie et al. Cell. 2018;175:1228-1243). Por lo tanto, el

enfoque para 6mA como se divulga en el presente documento sería útil para estudiar la genómica del cáncer (Xiao CL et al. Mol Cell. 2018;71:306-318; Xie et al. Cell. 2018;175:1228-1243). Además, se encontró que 6mA era más prevalente y abundante en el ADN mitocondrial de los mamíferos, lo que se muestra en asociación con la hipoxia (Hao Z et al. Mol Cell. 2020; doi:10.1016/j.molcel.2020.02.018). Por lo tanto, el enfoque para la detección de 6mA en esta divulgación sería útil para estudiar la respuesta al estrés mitocondrial en diferentes condiciones clínicas tales como el embarazo, cáncer y enfermedades autoinmunitarias.

IV. Resultados y aplicaciones

A. Detección de metilación

La detección de metilación en sitios CpG utilizando métodos de la invención descritos anteriormente se realizó para diferentes muestras biológicas y regiones genómicas. Como un ejemplo, la determinación de la metilación con ADN libre de células en el plasma de mujeres embarazadas utilizando una secuenciación en tiempo real de única molécula se verificó frente a la determinación de la metilación utilizando la secuenciación con bisulfito. Los resultados de la metilación se pueden utilizar para diferentes aplicaciones, que incluye la determinación del número de copias y el diagnóstico de trastornos. Los métodos descritos a continuación no se limitan a sitios CpG y también se pueden aplicar a cualquier modificación descrita en el presente documento y, por lo tanto, se describen con fines ilustrativos.

1. Detección de metilación de moléculas largas de ADN en tejido placentario.

La secuenciación en tiempo real de única molécula podría secuenciar moléculas de ADN de kilobases de longitud (Nattestad et al., 2018). El descifrado de los estados de metilación para los sitios CpG utilizando la invención descrita en el presente documento permitiría inferir la información del haplotipo de los estados de metilación al hacer uso sinérgico de la información de lectura larga de la secuenciación en tiempo real de única molécula. Para demostrar la viabilidad de inferir los estados de metilación de lectura larga, así como su información de haplotipo, secuenciamos un ADN de tejido de placenta con 478,739 moléculas que estaban cubiertas por 28,913,838 sublecturas. Había 7 moléculas de más de 5 kb de tamaño. Cada uno estaba cubierto en promedio por 3 sublecturas.

La FIG. 47 muestra los estados de metilación a lo largo de la molécula larga de ADN con un tamaño de 6,265 bp (es decir, un bloque de haplotipo), que se secuenció en un ZMW con el número de agujero ZMW m54276_180626_162240/40763503 y se mapeó a la ubicación genómica de chr1:113246546-113252811 en el genoma humano. '-' representaba el nucleótido no CpG; 'U' representa el estado no metilado en un sitio CpG; y 'M' representa el estado metilado en un sitio CpG. La región 4710 resaltada en amarillo indicó una región isla CpG que se sabía que no estaba metilada en general (FIG. 47). Se dedujo que la mayoría de los sitios CpG en esa isla CpG no estaban metilados (96 %). Por el contrario, se dedujo que el 75 % de los sitios CpG fuera de la isla CpG no estaban metilados. Estos resultados sugirieron que el nivel de metilación fuera de la isla CpG (por ejemplo, orilla/plataforma de la isla CpG) era mayor que el de la isla CpG. La mezcla de estados metilados y no metilados en una disposición de haplotipos en las regiones fuera de esa isla CpG indicaría la variabilidad de los patrones de metilación. En general, dichas observaciones coincidieron con los conocimientos actuales (Zhang et al., 2015; Feinberg e Irizarry, 2010). Por lo tanto, esta divulgación ha permitido llamar a diferentes estados de metilación a lo largo de una molécula larga, que incluyen los estados de metilación y desmetilación, lo que implica que la información del haplotipo de los estados de metilación podría estar en fase. La información de haplotipo se refiere a la vinculación de los estados de metilación de los sitios CpG en un tramo contiguo de ADN.

Describimos cómo podríamos utilizar este enfoque en el presente documento para analizar estados de metilación a lo largo de un haplotipo para detectar y analizar las regiones impresas. Las regiones impresas están sujetas a una regulación epigenética que provoca estados de metilación en una forma de progenitor de origen. Por ejemplo, una región impresa importante se encuentra en el cromosoma humano 11p15.5 y contiene los genes impresos IGF2, H19 y CDKN1C ($P57^{kip2}$), que son fuertes reguladores del crecimiento fetal (Brioude et al, Nat Rev Endocrinol. 2018;14:229-249). Las aberraciones genéticas y epigenéticas en las regiones impresas estarían asociadas con enfermedades. El síndrome de Beckwith-Wiedemann (BWS) es un síndrome de sobrecrecimiento, en el que los pacientes a menudo presentan macroglosia, defectos de la pared abdominal, hemihiperplasia, agrandamiento de los órganos abdominales y un mayor riesgo de tumores embrionarios durante la primera infancia. Se considera que el BWS es causado por defectos genéticos o epigenéticos dentro de las regiones 11p15.5 (Brioude et al, Nat Rev Endocrinol. 2018;14:229-249). Una región llamada ICR1 (región de control de impronta 1) que se encuentra entre H19 e IGF2 está metilada diferencialmente en el alelo paterno. ICR1 dirige la expresión específica del progenitor de origen de IGF2. Por lo tanto, las aberraciones genéticas y epigenéticas en ICR1 conducirían a una expresión aberrante de IGF2, que es una de las posibles razones que causan el BWS. Por tanto, la detección de estados de metilación a lo largo de las regiones impresas sería de importancia clínica.

Descargamos datos para 92 genes impresos de una base de datos pública que selecciona genes impresos informados actualmente (http://www.geneimprint.org/). Las regiones de 5 kb en dirección ascendente y en dirección descendente de estos genes impresos se utilizaron para análisis adicionales. Entre estas regiones, 160 islas CpG se asocian con estos genes impresos. Obtuvimos 324,248 secuencias consenso circulares de una muestra de placenta. Después de eliminar las secuencias de consenso circulares con baja calidad y regiones cortas superpuestas con las islas CpG (por

ejemplo, más pequeñas que el 50 % de la longitud de esa isla CpG relevante), obtuvimos 9 secuencias de consenso circulares que se superponían con 9 islas CpG que correspondían a 8 genes impresos.

La FIG. 48 es una tabla que muestra que las 9 moléculas de ADN se secuenciaron mediante una secuenciación en tiempo real de única molécula y superposición con regiones impresas, que incluyen H19, WT1-AS, WT1, DLK1, MEG3, ATP10A, LRRTM1 y MAGI2. La 6ª columna contenía los tramos de ADN que se superponían con islas CpG que involucraban las regiones impresas. 'U' representa una citosina no metilada en el contexto CpG; 'M' representa una citosina metilada en el contexto CpG. '*' representa un sitio CpG que no estaba cubierto en el resultado de la secuenciación; '-' representa un nucleótido de sitios no CpG; el genotipo se indica entre paréntesis si una molécula se superpone con un polimorfismo de un solo nucleótido (SNP). La 7ª columna indica los estados de metilación de una molécula completa. Una molécula se puede considerar metilada si se demostró que la mayoría de los sitios CpG (por ejemplo, más del 50 %) están metilados de acuerdo con las realizaciones presentes en esta divulgación; de lo contrario se llamaría no metilado.

Entre 9 moléculas de ADN, 5 moléculas de ADN (55.6 %) se llamaron metiladas, lo que no se desvió significativamente de la expectativa en la que el 50 % de las moléculas de ADN estarían metiladas. Como se muestra en la 6ª columna de la tabla de la FIG. 48, se mostró que la mayoría de los sitios CpG estaban metilados o no metilados de manera concertada, es decir, como un haplotipo de metilación. Una molécula se denominaría metilada si se demostrara que la mayoría de los sitios CpG (por ejemplo, más del 50 %) están metilados de acuerdo con las realizaciones presentes en esta divulgación; de lo contrario, se denominaría no metilada. Se podrían utilizar otros valores de corte, para determinar si una molécula está metilada o no, por ejemplo, pero no se limita a, al menos 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 % y 100 % de los sitios CpG en una molécula se analizaron para considerarlos metilados.

También describimos que podríamos utilizar moléculas que comprendan simultáneamente al menos un SNP y al menos un análisis de sitio CpG para determinar si una región podría estar asociada con una región impresa o si un gen impreso conocido podría ser aberrante (por ejemplo, pérdida de impronta). Con fines ilustrativos, la FIG. 49 muestra que la primera molécula de una región de impronta portaba el alelo 'A'; y la segunda molécula de esa región de impronta llevaba el alelo 'G'. Suponiendo que la región de impronta se imprimiera de forma paterna, la primera molécula del haplotipo materno estaba completamente desmetilada; y la segunda molécula del haplotipo paterno estaba completamente metilada. Dicha suposición proporcionaría la verdad absoluta de los estados de metilación, permitiendo probar el rendimiento de la detección de metilación de acuerdo con las realizaciones presentes en esta divulgación.

La FIG. 49 muestra un ejemplo para la determinación de patrones de metilación en una región impresa. Se extrajo el ADN de una muestra biológica y se ligó con adaptadores de horquilla para formar moléculas circulares de ADN. Se desconocía la información de secuencia y las modificaciones de bases (por ejemplo, estados de metilación en sitios CpG) con respecto a esas moléculas circulares de ADN. Esas moléculas circulares de ADN se sometieron a una secuenciación en tiempo real de única molécula. Las IPD, PW y el contexto de secuencia para las bases en cada sublectura que se originan a partir de esas moléculas circulares de ADN se determinaron después de que las sublecturas se mapearan en el genoma de referencia. Además, se determinaron los genotipos de esas moléculas. Las IPD, PW y el contexto de secuencia en una ventana de medición asociada con los sitios CG se compararían con los patrones cinéticos de referencia para determinar los estados de metilación para cada CpG. Si dos moléculas con alelos diferentes mostraran diferentes patrones de metilación de manera que una estuviera completamente desmetilada y la otra completamente metilada, la región genómica asociada con estas dos moléculas sería una región impresa. Si dicha región genómica fuera una región impresa conocida, por ejemplo, como se ilustra en la FIG. 49, los patrones de metilación para estas dos moléculas estaban en línea con los patrones de metilación esperados (es decir, la verdad absoluta) en una situación normal. Puede sugerir la exactitud de los métodos para la clasificación de estados de metilación de acuerdo con esta divulgación. La derivación entre los patrones de metilación medidos descritos en esta divulgación y los patrones de metilación esperados indicaría las aberraciones de la impronta, por ejemplo, la pérdida de la impronta.

La FIG. 50 muestra un ejemplo para la determinación de patrones de metilación en una región impresa. El patrón de impronta se podría determinar aún más a través del análisis de los patrones de metilación de esa región en un determinado árbol genealógico. Por ejemplo, se podría realizar el análisis de patrones de metilación e información alélica en los genomas paterno y materno y en la descendencia. Dicho árbol genealógico podría incluir además los genomas del abuelo paterno o materno, de la abuela paterna o materna u otros genomas relevantes. Dicho análisis se podría ampliar a conjuntos de datos de tríos familiares (madre, padre e hijo) en una población determinada, por ejemplo, obteniendo información de metilación y genotipo para cada individuo como se describe en el presente documento.

Como se muestra después de la clasificación, se pueden determinar tanto el genotipo (alelo en el cuadro) como el estado de metilación. Para cada una de las moléculas, se puede proporcionar un patrón de metilación en cada sitio (por ejemplo, todas metiladas o todas no metiladas) para identificar de qué progenitor se hereda la molécula. O bien, se puede determinar una densidad de metilación, y uno o más valores de corte pueden clasificar si la molécula está hipermetilada (por ejemplo, > 80 % u otro % y de un progenitor) o hipometilada (por ejemplo, < 20 % u otro % y del otro progenitor).

2. Detección de metilación de moléculas de ADNlc

Como otro ejemplo ilustrativo, la metilación del ADN libre de células (ADNlc) también se ha reconocido cada vez más como señales moleculares importantes para las pruebas prenatales no invasivas. Por ejemplo, hemos mostrado que las moléculas de ADNlc de regiones que llevan metilación específica de tejido se pueden utilizar para determinar las contribuciones proporcionales de diferentes tejidos tales como neutrófilos, células T, células B, hígado y placenta en el plasma de mujeres embarazadas (Sun et al., 2015). También se ha demostrado la viabilidad de utilizar la metilación del ADN plasmático de mujeres embarazadas para detectar la trisomía 21 (Lun et al., 2013). Las moléculas de ADNlc en el plasma materno se fragmentaron con un tamaño medio de 166 bp, que es mucho más corto que el ADN de E. coli fragmentado artificialmente con aproximadamente 500 bp de tamaño. Se ha informado que el ADNlc está fragmentado de manera no aleatoria, por ejemplo, motivos terminales del ADN plasmático en asociación con los orígenes del tejido, tales como por ejemplo de la placenta. Dichas propiedades características del ADN libre de células dan un contexto de secuencia extremadamente diferente al del ADN de E. coli fragmentado artificialmente. Por lo tanto, aún se desconoce si dicha cinética de la polimerasa permitiría deducir cuantitativamente los niveles de metilación, normalmente para moléculas de ADN libres de células. Las divulgaciones en esta solicitud de patente serían aplicables, pero no se limitan a, al análisis de metilación del ADN libre de células en el plasma de mujeres embarazadas, por ejemplo al utilizar el modelo de predicción de la metilación entrenado a partir de las moléculas de ADN de tejido mencionadas anteriormente.

Utilizando secuenciación en tiempo real de única molécula, se secuenciaron seis muestras de ADN plasmático de mujeres embarazadas con un feto masculino con una mediana de 30,738,399 sublecturas (rango: 1,431,215-105,835,846), correspondiente a una mediana de 111.834 CCS (rango: 61,010-503,582). Cada ADN plasmático se secuenció con una mediana de 262 veces (rango: 173-320). El conjunto de datos se generó a partir de ADN preparado con el Kit de Secuenciación Sequel I 3.0.

Para evaluar la detección de metilación de moléculas de ADNlc, utilizamos secuenciación con bisulfito (Jiang et al., 2014) para analizar la metilación de las 6 muestras de ADN plasmático de mujeres embarazadas mencionadas anteriormente. Obtuvimos una mediana de 66 millones de lecturas de pares (58-82 millones de lecturas de pares). Se encontró que la mediana de metilación general era del 69.6 % (67.1 %-72.0 %).

La FIG. 51 muestra la comparación de los niveles de metilación deducidos por el nuevo enfoque y la secuenciación con bisulfito convencional. El eje y son los niveles de metilación predichos como se describe en esta solicitud de patente. El eje x son los niveles de metilación deducidos mediante la secuenciación con bisulfito. Se analizó una mediana de 314,675 sitios CpG (rango: 144,546-1,382,568) para obtener resultados de ADN plasmático generados a partir de secuenciación en tiempo real de única molécula. La proporción de la mediana de sitios CpG que se predijo que estaban metilados fue del 64.7 % (rango: 60.8-68.5 %), lo que parecía ser comparable con los resultados deducidos de la secuenciación con bisulfito. Como se muestra en la FIG. 51, hubo una buena correlación (r: 0.96, valor de p = 0.0023) entre los niveles generales de metilación deducidos por una secuenciación en tiempo real de única molécula con el enfoque actual de predicción de la metilación y la secuenciación con bisulfito.

Debido a la poca profundidad de la secuenciación con bisulfito, podría no ser sólida para deducir los niveles de metilación (es decir, la fracción de CpG secuenciado que se metila) para cada CpG en el genoma humano. En cambio, calculamos los niveles de metilación en algunas regiones con múltiples sitios CpG, al agregar señales de lectura que cubren sitios CpG de una región genómica en la que dos sitios CpG consecutivos estaban dentro de 50 nt y el número de sitios CpG era al menos 10. El porcentaje de citosina secuenciada entre la suma de citosinas y timinas secuenciadas en sitios CpG en una región indicó los niveles de metilación de esa región. Las regiones se dividieron en diferentes grupos de acuerdo con los niveles regionales de metilación. La probabilidad de metilación predicha por el modelo aprendido de los conjuntos de datos de entrenamiento anteriores (es decir, ADN de tejido) se elevó de acuerdo con lo anterior a medida que aumentaron los niveles de metilación (FIG. 52A). Estos resultados sugirieron además la viabilidad y validez del uso de secuenciación en tiempo real de única molécula para predecir los estados de metilación de las moléculas de ADNlc en mujeres embarazadas. La FIG. 52B mostró que el nivel de metilación en una ventana genómica de 10 Mb estimado utilizando secuenciación en tiempo real de única molécula de acuerdo con las realizaciones presentes en esta divulgación se corrigió bien con el de secuenciación con bisulfito (r = 0.74; valor de p < 0.0001).

La FIG. 53 mostraron que las representaciones genómicas (GR) del cromosoma Y en el plasma materno de mujeres embarazadas medidas mediante secuenciación en tiempo real de única molécula estaban bien correlacionadas con las medidas mediante BS-seq (r = 0.97; valor de P = 0.007). Estos resultados sugirieron que la secuenciación en tiempo real de única molécula también permitiría la cuantificación precisa de moléculas de ADN originadas en tejidos no hematopoyéticos tales como la placenta, cuyo ADN aportado generalmente representaba una minoría. En otras palabras, esta divulgación demostró la viabilidad de analizar simultáneamente las aberraciones del número de copias y los estados de metilación de moléculas nativas sin conversiones ni amplificaciones de bases antes de la secuenciación.

3. Método basado en bloques CpG

También describimos la realización de un análisis de metilación en una serie regiones genómicas que albergan múltiples sitios CpG, por ejemplo, pero no se limitan a, 2, 3, 4, 5, 10, 20, 30, 40, 50, 100 sitios CpG, etc. El tamaño de dicha región genómica puede ser, por ejemplo, pero no se limitan a, 50, 100, 200, 300 y 500 nt, etc. La distancia entre sitios CpG en esta región podría ser, por ejemplo, pero no se limita a 10, 20, 30, 40, 50, 100, 200, 300 nt, etc. Podríamos fusionar dos sitios CpG consecutivos dentro de 50 nt para formar un bloque CpG de tal manera que el número de sitios CpG en este bloque fuera más de 10. En dicho método basado en bloques, se pueden combinar múltiples regiones en una ventana representada como una única matriz, tratando efectivamente las regiones juntas.

Como un ejemplo ilustrativo, como se muestra en la FIG. 54, la cinética de todas las sublecturas asociadas con un bloque CpG se utilizó para el análisis de metilación. Los perfiles IPD proyectados de los 10 nt que flanquean en dirección ascendente y en dirección descendente en cada CpG en ese bloque se alinearon artificialmente con respecto a los sitios CpG para calcular el perfil IPD promedio (FIG. 54). La palabra "proyectada" significa que habíamos alineado las señales cinéticas de sublecturas con cada sitio CpG correspondiente en cuestión. Los perfiles IPD promedio para un bloque CpG se utilizaron para entrenar un modelo (por ejemplo, utilizando una red neuronal artificial, ANN para abreviar) para identificar los estados de metilación para cada bloque. El análisis de ANN incluyó una capa de entrada, dos capas ocultas y una capa de salida. Cada bloque CpG se caracterizó por un vector de características de 21 valores de IPD que se ingresarían a la ANN. La primera capa oculta incluía 10 neuronas con ReLu como función de activación. La segunda capa oculta incluía 5 neuronas con ReLu como función de activación. Finalmente, la capa de salida incluía 1 neurona con sigmoide como función de activación que generaría la probabilidad de metilación. Un sitio CpG que muestra una probabilidad de metilación> 0.5 se consideró metilación; de lo contrario, se consideró desmetilación. El perfil IPD promedio se puede utilizar para analizar el estado de metilación de una molécula completa. La molécula completa se puede considerar metilada si un cierto número de sitios por encima de un umbral (por ejemplo, 0, 1, 2, 3, etc.) están metilados o si la molécula tiene una cierta densidad de metilación.

Había 9,678 y 9,020 bloques CpG en bibliotecas metiladas y no metiladas, cada una de las cuales albergaba al menos 10 sitios CpG. Esos bloques CpG cubrieron 176,048 y 162,943 sitios CpG para bibliotecas metiladas y no metiladas. Como se muestra en la FIG. 55A y FIG. 55B, podríamos lograr una precisión general superior al 90 % en la predicción de estados de metilación tanto en el conjunto de datos de entrenamiento como en el conjunto de datos de prueba. Sin embargo, dicha realización basada en bloques CpG reduciría en gran medida la cantidad de CpG que se pudieron evaluar. Por definición, el requisito del menor número de sitios CpG restringiría el análisis de metilación a algunas regiones genómicas particulares (por ejemplo, analizando preferentemente islas CpG).

B. Determinación del origen o trastorno

Los perfiles de metilación se pueden utilizar para detectar el origen del tejido o determinar la clasificación de un trastorno. El análisis del perfil de metilación se puede utilizar junto con otros datos clínicos, que incluyen formación de imágenes, paneles de sangre convencionales y otra información de diagnóstico médico. Los perfiles de metilación se pueden determinar utilizando cualquier método descrito en el presente documento.

1. Determinación de la aberración del número de copias.

Esta sección muestra que SMRT es precisa para determinar el número de copias y, por lo tanto, el perfil de metilación y el perfil del número de copias se pueden analizar simultáneamente.

Se ha demostrado que las aberraciones en el número de copias pueden revelarse mediante la secuenciación de los tejidos tumorales (Chan (2013)). En el presente documento, mostramos que las aberraciones en el número de copias asociadas al cáncer se pueden identificar mediante la secuenciación de tejidos tumorales utilizando secuenciación en tiempo real de única molécula. Por ejemplo, para el caso TBR3033, obtuvimos 589,435 y 1,495,225 secuencias consenso (el requisito mínimo de sublecturas utilizadas para construir cada secuencia consenso fue 5) para el ADN tumoral y su ADN de tejido hepático no tumoral adyacente emparejado, respectivamente. El conjunto de datos se generó a partir de ADN preparado con el Kit de Secuenciación Sequel II 1.0. El genoma se dividió, in silico, en ventanas de 2 Mb. Se calculó el porcentaje de secuencias consenso mapeadas a cada ventana, lo que dio como resultado una representación genómica (GR) con una resolución de 2 Mb. La GR se puede determinar mediante una serie de lecturas en una posición normalizada por las lecturas de secuencia total en todo el genoma.

La FIG. 56A muestra la relación de GR entre el tumor y su ADN de tejido no tumoral adyacente emparejado utilizando secuenciación en tiempo real de única molécula. La relación del número de copias entre el ADN del tumor y el ADN del tejido normal adyacente emparejado se muestra en el eje y, y el índice de intervalo genómico para cada ventana de 2 Mb, que incluye los cromosomas 1 a 22, se muestra en el eje x. Para esta figura, una región con una relación de GR superior al percentil 95 de todas las ventanas de 2 Mb se clasificó como con ganancia en el número de copias, mientras que una región con una proporción de GR inferior al percentil 5 de todas las ventanas de 2 Mb se clasificó como que tiene pérdida del número de copia. Observamos que el cromosoma 13 albergaba pérdidas en el número de copias, mientras que el cromosoma 20 albergaba ganancias en el número de copias. Dichas ganancias y pérdidas son el resultado correcto.

La FIG. 56B muestra la relación de GR entre el tumor y su tejido no tumoral adyacente emparejado utilizando secuenciación con bisulfito. La relación del número de copias entre el ADN del tumor y el ADN del tejido normal

adyacente emparejado se muestra en el eje y, y el índice del intervalo genómico para cada ventana de 2 Mb, que incluye los cromosomas 1 a 22, se muestra en el eje x. Los cambios en el número de copias identificados mediante secuenciación en tiempo real de única molécula en la FIG. 56A se verificaron en los resultados de secuenciación de bisulfito coincidentes en la FIG. 56B.

Para el caso TBR3032, obtuvimos 413,982 y 2,396,054 secuencias consenso (el requisito mínimo de sublecturas utilizadas para construir cada secuencia consenso fue 5) para el ADN tumoral y su ADN de tejido no tumoral adyacente emparejado, respectivamente. El genoma se dividió, in silico, en ventanas de 2 Mb. Se calculó el porcentaje de secuencias de consenso mapeadas a cada ventana, a saber, representación genómica (GR) de 2 Mb.

La FIG. 57A muestra la relación de GR entre el tumor y su ADN de tejido no tumoral adyacente emparejado utilizando secuenciación en tiempo real de única molécula. La relación del número de copias entre el ADN del tumor y el ADN del tejido normal adyacente emparejado se muestra en el eje y, y el índice de intervalo genómico para cada ventana de 2 Mb, que incluye los cromosomas 1 a 22, se muestra en el eje x. Para esta figura, una región con una proporción de GR superior al percentil 95 de todas las ventanas de 2 Mb se clasificó como con ganancia en el número de copias, mientras que una región con una proporción de GR inferior al percentil 5 de todas las ventanas de 2 Mb se clasificó como tener pérdida del número de copia. Observamos que los cromosomas 4, 6, 11, 13, 16 y 17 albergaban pérdidas en el número de copias, mientras que los cromosomas 5 y 7 albergaban ganancias en el número de copias.

La FIG. 57B muestra la relación de GR entre el tumor y su tejido no tumoral adyacente emparejado utilizando secuenciación con bisulfito. La relación del número de copias entre el ADN del tumor y el ADN del tejido normal adyacente emparejado se muestra en el eje y, y el índice de intervalo genómico para cada ventana de 2 Mb, que incluye los cromosomas 1 a 22, se muestra en el eje x. Los cambios en el número de copias identificados mediante secuenciación en tiempo real de única molécula en la FIG. 57A se verificaron en los resultados de secuenciación de bisulfito coincidentes en la FIG. 57B.

De acuerdo con lo anterior, el perfil de metilación y el perfil del número de copias se pueden analizar simultáneamente. En este ejemplo, dado que la pureza tumoral de un tejido tumoral generalmente no siempre es del 100 %, las regiones amplificadas aumentarían relativamente la contribución del ADN tumoral, mientras que las regiones suprimidas disminuirían relativamente la contribución del ADN tumoral. Debido a que el genoma del tumor se caracteriza por una hipometilación global, las regiones amplificadas disminuirían aún más los niveles de metilación en comparación con las regiones suprimidas. A modo de ilustración, para el caso TBR3033, el nivel de metilación del cromosoma 22 (ganancias del número de copias) medido utilizando la presente invención fue del 48.2 %, que fue inferior al del cromosoma 3 (pérdidas del número de copias) (nivel de metilación: 54.0 %). Para el caso TBR3032, el nivel de metilación del brazo del cromosoma 5p (ganancias del número de copias) medido utilizando la presente invención fue del 46.5 %, que fue inferior al del brazo del cromosoma 5q (pérdidas del número de copias) (nivel de metilación: 54.9 %).

2. Mapeo de tejido del ADN plasmático en mujeres embarazadas.

Como se muestra en la FIG. 58, razonamos que la precisión del análisis de metilación nos permitiría comparar los perfiles de metilación del ADN plasmático de una mujer embarazada con los perfiles de metilación de diferentes tejidos de referencia (por ejemplo, hígado, neutrófilos, linfocitos, placenta, células T, células B, corazón, cerebro, etc.). Por lo tanto, las contribuciones de ADN en la agrupación de ADN plasmático de una mujer embarazada a partir de diferentes tipos de células podrían deducirse mediante los siguientes procedimientos. Los niveles de metilación de CpG de una mezcla de ADN (por ejemplo, ADN plasmático) determinados como se describe en esta divulgación se registraron en un vector (X) y los niveles de metilación de referencia recuperados en diferentes tejidos se registraron en una matriz (M) que se podría cuantificar mediante, pero sin limitarse a, secuenciación con bisulfito. Las contribuciones proporcionales (p) de diferentes tejidos a una mezcla de ADN podrían resolverse mediante programación cuadrática, entre otras. En el presente documento, utilizamos ecuaciones matemáticas para ilustrar la deducción de la contribución proporcional de diferentes órganos a una mezcla de ADN que se está analizando. La relación matemática entre las densidades de metilación de los diferentes sitios en una mezcla de ADN y las densidades de metilación de los sitios correspondientes en diferentes tejidos se puede expresar como:

$$\bar{X}_i = \sum_k (p_k \times M_{ik}),$$

donde $\bar{X}_i$ representa la densidad de metilación de un sitio CpG i en una mezcla de ADN; p_k representa la contribución proporcional del tipo celular k a una mezcla de ADN; M_{ik} representa la densidad de metilación del sitio CpG *i* en el tipo de célula *k*. Cuando el número de sitios es igual o mayor que el número de órganos, se podrían determinar los valores de p_k individuales. Para mejorar la información, los sitios CpG mostraron una pequeña variabilidad de los niveles de metilación en todos los tipos de tejidos de referencia y se descartaron. Utilizamos un conjunto específico de sitios CpG para realizar el análisis. Por ejemplo, esos sitios CpG se caracterizaron con un coeficiente de variación (CV) de los niveles de metilación en diferentes tejidos mayor al 30 % y una diferencia entre los niveles de metilación máximo y mínimo entre los tejidos mayor al 25 %. En algunas otras realizaciones, también se podría utilizar un CV de 5 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 80 %, 90 %, 100 %, 110 %, 200 %, 300 %, etc.; y se podrían utilizar una diferencia

entre los niveles de metilación máximo y mínimo entre tejidos superior al 5 %, 10 %, 15 %, 20 %, 25 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 100 %, etc.

Se pueden incluir criterios adicionales en el algoritmo para mejorar la precisión. Por ejemplo, la contribución agregada de todos los tipos de células se limitaría a ser del 100 %, es decir,

$$\sum_k p_k = 100\%.$$

Además, se requeriría que todas las contribuciones de los órganos no fueran negativas:

$$p_k \geq 0, \forall k$$

Debido a variaciones biológicas, el patrón de metilación general observado puede no ser completamente idéntico al patrón de metilación deducido de la metilación de los tejidos. En dicha circunstancia, sería necesario el análisis matemático para determinar la contribución proporcional más probable de los tejidos individuales. En este sentido, la diferencia entre el patrón de metilación observado en el ADN y el patrón de metilación deducido de los tejidos se denota por W:

$$\mathrm{W} = \bar{X}_i - \sum_k (p_k \times M_{ik})$$

El valor más probable de cada p_k se puede determinar al minimizar W, que es la diferencia entre los patrones de metilación observados y deducidos. Esta ecuación se puede resolver utilizando algoritmos matemáticos, por ejemplo, pero no se limitan a, utilizar programación cuadrática, regresión lineal/no lineal, algoritmo de maximización de expectativas (EM), algoritmo de máxima verosimilitud, estimación máxima a posteriori y método de mínimos cuadrados.

Como se muestra en la FIG. 59, observamos que la contribución del ADN placentario al plasma materno de mujeres embarazadas que llevan fetos masculinos, utilizando el método de mapeo de tejido del ADN plasmático presente en la FIG. 58, estaba bien correlacionado con las fracciones de ADN fetal que se estimaron mediante lecturas del cromosoma Y. Este resultado sugirió la viabilidad de utilizar la cinética para rastrear los tejidos de origen del ADN plasmático en mujeres embarazadas.

3. Cuantificación del nivel de metilación regional.

Esta sección describe técnicas para determinar un nivel representativo de metilación para regiones genómicas seleccionadas, que se puede realizar utilizando un nivel relativamente bajo de secuenciación. Los niveles de metilación se pueden determinar en una base por hebra o por molécula, o por región, utilizando el número de sitios metilados y un número total de sitios metilados. También se analizan los niveles de metilación de varios tejidos.

Secuenciamos 11 muestras de ADN de tejido humano hasta una mediana de 30.7 millones de sublecturas (rango: 9.1 – 88.6 millones) por muestra que se podrían alinear con un genoma de referencia humano (hg19). Las sublecturas de cada muestra se generaron a partir de una mediana de 3.8 millones de pocillos de Secuenciación en Tiempo Real de Molécula Única (SMRT) de Pacific Biosciences (rango: 1.1 – 11.5 millones), cada uno de los cuales contenía al menos una sublectura que se podría alinear con un genoma humano de referencia. En promedio, cada molécula en un pocillo de SMRT se secuenció un promedio de 9.9 veces (rango: 6.5 – 13.4 veces). Las muestras de ADN de tejido humano incluyeron 1 muestra de capa leucocitaria materna de una mujer embarazada, 1 muestra de placenta, 2 tejidos tumorales de carcinoma hepatocelular (HCC), 2 tejidos no tumorales adyacentes emparejados con los 2 tejidos de HCC mencionados anteriormente, 4 muestras de capa leucocitaria de sujetos de control sanos (M1 y M2 eran de sujetos masculinos; F1 y F2 eran de sujetos femeninos), 1 estirpe celular HCC (HepG2). Los detalles del resumen de datos de secuenciación se muestran en la FIG. 60.

La FIG. 60 muestra los diferentes grupos de tejidos en la primera columna y los nombres de las muestras en la segunda columna. “Sublecturas totales” indica el número total de secuencias generadas a partir de pocillos de SMRT, que incluyen las de las hebras de Watson y Crick. Las “sublecturas mapeadas” enumeran el número de sublecturas que se podrían alinear con un genoma de referencia humano. La “mapeabilidad de sublecturas” se refiere a la proporción de sublecturas que se podrían alinear con un genoma de referencia humano. La “profundidad media de sublecturas por pocillo de SMRT” indica el número medio de sublecturas generadas a partir de cada pocillo de SMRT. “Número de pocillos de SMRT” se refiere al número de pocillos de SMRT que produjeron sublecturas detectables. “Pocillos mapeables” indica el número de pocillos que contienen al menos una sublectura alineable. “Tasa de pocillos mapeables (%)” es el porcentaje de pocillos que contenían al menos una sublectura alineable.

a) Técnicas de análisis de patrones y niveles de metilación.

Describimos cómo se puede medir la densidad de metilación de una única hebra de ácido nucleico (por ejemplo, ADN o ARN), que se define como el número de bases metiladas dentro de la hebra dividido por el número total de bases metilables dentro de esa hebra. Esta medición también se denomina "nivel de metilación de hebra sencilla". Esta medición de una sola hebra es particularmente factible en el contexto de la divulgación actual porque la plataforma de secuenciación en tiempo real de única molécula puede obtener información de secuenciación de cada una de las dos hebras de una molécula de ADN de hebra doble. Esto se facilita con el uso de adaptadores de horquilla en la preparación de las bibliotecas de secuenciación de tal manera que las hebras de Watson y Crick de una molécula de ADN de hebra doble se conecten en un formato circular y se secuencian juntas. De hecho, esta estructura también permite secuenciar las hebras de Watson y Crick asociadas de la misma molécula de ADN de hebra doble en la misma reacción, de tal manera que el estado de metilación de los sitios complementarios correspondientes en las hebras de Watson y Crick de cualesquier moléculas de ADN de hebra doble se podrían determinar individualmente y comparar directamente (por ejemplo, las FIG. 20A y 20B).

Estos análisis de metilación basados en hebras no podrían lograrse fácilmente con otras tecnologías. Debido a que sin el uso del método de análisis de metilación directa como se divulga en esta solicitud, sería necesario aplicar otros medios para diferenciar las bases metiladas de las no metiladas, por ejemplo, mediante conversión de bisulfito. La conversión con bisulfito requiere que el ADN se trate con bisulfito de sodio para que las citosinas metiladas y las citosinas no metiladas puedan distinguirse como citosinas y timinas, respectivamente. Bajo las condiciones desnaturalizantes de muchos protocolos de conversión de bisulfito, las dos hebras de una molécula de ADN de hebra doble se disocian entre sí. En muchas aplicaciones de secuenciación, utilizando, por ejemplo, la plataforma Illumina, el ADN convertido con bisulfito se amplifica mediante la reacción en cadena de la polimerasa (PCR), que implica la disociación del ADN de hebra doble en hebras simples.

Con la secuenciación de Illumina, se pueden preparar bibliotecas de secuenciación sin PCR utilizando adaptadores metilados antes de la conversión con bisulfito. Incluso con el uso de esta estrategia, cada hebra de ADN de una molécula de ADN de hebra doble se elegiría aleatoriamente para la amplificación del puente en la celda de flujo. Debido a la naturaleza aleatoria de la secuenciación, es poco probable que cada hebra de la misma molécula de ADN se secuencia en la misma reacción. Incluso si se analiza más de una secuencia leída del mismo locus en la misma ejecución, no hay manera fácil de determinar si las dos lecturas son de cada una de las hebras de Watson y Crick asociadas de una molécula de ADN de hebra doble o son de dos diferentes moléculas de ADN de hebra doble. Dichas consideraciones son importantes porque describimos cómo las dos hebras de una molécula de ADN de hebra doble pueden exhibir diferentes patrones de metilación. Cuando se miden las densidades de metilación de una sola hebra de múltiples hebras de ácido nucleico (por ejemplo, ADN o ARN), también se puede determinar un "nivel de metilación de múltiples hebras" en base a los conceptos y la ecuación con respecto al "nivel de metilación de una región genómica de interés" en la FIG. 61.

La FIG. 61 muestra varias formas de analizar patrones de metilación. Una molécula de ADN de hebra doble (X) con secuencia e información de metilación desconocidas se liga con adaptadores, lo que forma, en un ejemplo, una estructura en forma de horquilla. Como resultado, las dos únicas hebras de la molécula de ADN, que incluyen las hebras de Watson X(a) y Crick X(b), están físicamente asociadas entre sí en forma circular en este ejemplo. Los estados de metilación de los sitios en las hebras de Watson y Crick se pueden obtener utilizando métodos descritos en esta divulgación (por ejemplo, utilizando señales cinéticas, electrónicas, electromagnéticas, ópticas u otro tipo de señales físicas del secuenciador). Las hebras de Watson y Crick en la molécula de ADN circularizada se pueden interrogar en la misma reacción. Después de la secuenciación, se recortan las secuencias adaptadoras.

Se pueden determinar diferentes niveles de metilación a partir del análisis. En (I) de la FIG. 61, se puede analizar el patrón de metilación de sólo una molécula de hebra sencilla, como X(a) o X(b). Este análisis se puede denominar análisis del patrón de metilación de hebra sencilla. El análisis puede incluir, pero no se limitan a, determinar el estado de metilación de los sitios o el patrón de metilación. En la Fig. 61, la molécula de hebra sencilla X(a) muestra un patrón de metilación 5'-UMMUU-3' en el que "U" indica un sitio no metilado y "M" indica un sitio metilado mientras que la molécula de hebra sencilla complementaria X(b) muestra una metilación. patrón 3'-UMUUU-5'. Por tanto, X(b) tiene un patrón de metilación diferente al de X(a). Los niveles correspondientes de metilación de hebra sencilla de X(a) y X(b) son 40 % y 20 %, respectivamente.

Por el contrario, como se muestra en (II), se pueden analizar los patrones de metilación en un nivel de molécula de ADN de hebra doble único (es decir, tener en cuenta los patrones de metilación de las hebras de Watson y Crick. Este análisis se puede denominar como un análisis del patrón de metilación del ADN de hebra doble de una única molécula. El nivel de metilación de una única molécula y hebra doble del ADN de esta molécula ejemplar X es del 30 %. Una variante de este análisis, las señales cinéticas de las hebras de Watson y Crick serían combinados para analizar la metilación. En particular, como la metilación en los sitios CpG es generalmente simétrica, las señales cinéticas de las hebras de Watson y Crick se podrían combinar para un sitio antes de determinar los estados de metilación de los sitios. En algunas situaciones, el rendimiento de la determinación de metilaciones utilizando señales cinéticas combinadas de las hebras de Watson y Crick de una molécula sería superior a una que utiliza independientemente señales cinéticas de única hebra. Por ejemplo, como se muestra en la FIG. 20B, el uso combinado de señales cinéticas de ambas hebras, que incluyen las hebras de Watson y Crick, daría lugar a un AUC mayor (0.90) en un conjunto de datos de prueba en comparación con el uso independiente de una única hebra (AUC: 0.85).

En (III) de la FIG. 61, se determina el nivel de metilación de una región genómica de interés, donde diferentes moléculas de ADN, que llevan diferentes tamaños moleculares y diferente número de sitios metilables (por ejemplo, sitios CpG), pueden contribuir a la región genómica de interés. Este análisis se puede denominar análisis del nivel de metilación de múltiples hebras. El término "hebra múltiple" se puede referir a múltiples moléculas de ADN de hebra sencilla, o múltiples moléculas de ADN de hebra doble, o cualquier combinación de las mismas. En este ejemplo, hay tres moléculas de ADN de hebra doble que cubren una región genómica de interés: las moléculas "X", "Y" y "Z", cada una tiene hebras "a" y "b". El nivel de metilación correspondiente de esta región es 9/28, es decir, 32 %. El tamaño de la región genómica que se va a analizar puede tener un tamaño de 1 nt, 10 nt, 20 nt, 30 nt, 40 nt, 50 nt, 100 nt, 1 knt (kilonucleótidos, es decir, mil nucleótidos), 2 knt, 3 knt, 4 knt, 5 knt, 10 knt, 20 knt, 30 knt, 40 knt, 50 knt, 100 knt, 200 knt, 300 knt, 400 knt, 500 knt, 1 Mnt (meganucleótidos, es decir, 1 millón de nucleótidos), 2 Mnt, 3 Mnt, 4 Mnt, 5 Mnt, 10 Mnt, 20 Mnt, 30 Mnt, 40 Mnt, 50 Mnt, 100 Mnt, o 200 Mnt. La región genómica puede ser un brazo cromosómico o el genoma completo.

También se puede determinar un patrón de metilación después de determinar los estados de metilación para sitios en una molécula. Por ejemplo, en un escenario donde hay tres sitios CpG secuenciales en una única molécula de ADN de hebra doble, el patrón de metilación en cada una de las hebras de Watson y Crick se puede divulgar como metilado (M), no metilado (N) y metilado (M) para los tres sitios. Este patrón, MNM, por ejemplo, para la hebra de Watson, se puede denominar "haplotipo de metilación" para la hebra de Watson para esta región. Debido a la presencia de actividad de mantenimiento de la metilación del ADN, el patrón de metilación de las hebras de Watson y Crick de una molécula de ADN de hebra doble puede ser complementario entre sí. Por ejemplo, si un sitio CpG está metilado en la hebra de Watson, el sitio CpG complementario en la hebra de Crick también puede estar metilado. De manera similar, un sitio CpG no metilado en la hebra de Watson puede ser complementario de un sitio CpG no metilado en la hebra de Crick.

Se puede medir el nivel de metilación de una única molécula de ADN, que se define como el número de bases o nucleótidos metilados dentro de la molécula dividido por el número total de bases o nucleótidos metilables dentro de esa molécula. Esta medición también se denomina "nivel de metilación de única molécula". Esta medición de única molécula puede ser particularmente útil en el contexto de la divulgación actual debido a la larga longitud de lectura posible con la plataforma de secuenciación en tiempo real de única molécula. Cuando se miden los niveles de metilación de única molécula de múltiples moléculas de ADN, también se puede determinar un "nivel de metilación de múltiples moléculas" en base a los conceptos y la ecuación de la FIG. 61. Por ejemplo, el "nivel de metilación de múltiples moléculas" puede ser una media o una mediana de los niveles de metilación de única molécula.

También describimos cómo uno o más polimorfismos genéticos (por ejemplo, polimorfismos de un solo nucleótido (SNP)) se pueden analizar en la molécula de ADN junto con el estado de metilación de un sitio en la molécula, revelando de esta manera información tanto genética como epigenética de esa molécula. Dicho análisis revelaría el "haplotipo de metilación en fases" para la molécula de ADN analizada. El análisis de haplotipos de metilación en fases es útil, por ejemplo, en el estudio de la impronta genómica y los ácidos nucleicos libres de células en el plasma materno (que contiene una mezcla de moléculas de ADN libres de células que llevan firmas genéticas y epigenéticas maternas y fetales).

b) Comparación de los resultados de la metilación.

Las densidades de metilación a nivel de genoma completo de los tejidos en la tabla en la FIG. 60 se determinan utilizando secuenciación con bisulfito y utilizando secuenciación en tiempo real de única molécula como se describe en esta divulgación. Con fines ilustrativos, la FIG. 62A muestra la densidad de metilación cuantificada mediante secuenciación con bisulfito en el eje y y el tipo de tejido en el eje x. La FIG. 62B muestra la densidad de metilación cuantificada por secuenciación en tiempo real de molécula única, como se describe en esta divulgación en el eje y y el tipo de tejido en el eje x.

La FIG. 62A muestra las densidades de metilación en diferentes tejidos utilizando secuenciación con bisulfito (es decir, las muestras se convirtieron con bisulfito y luego se sometieron a secuenciación de Illumina) (Lister et al. Nature. 2009;462:315-322), que incluye HepG2, tejidos tumorales de HCC, tejidos de hígado normales compatibles adyacentes al tumor HCC (es decir, tejidos normales adyacentes), tejido placentario y muestras de capa leucocitaria. HepG2 exhibió el nivel de metilación más bajo, con un nivel de metilación del 40.4 %. Las muestras de capa leucocitaria exhibieron el nivel de metilación más alto, con un nivel de metilación del 76.5 %. Se encontró que la densidad de metilación media de los tejidos tumorales de HCC (51.2 %) era menor que la de los tejidos normales adyacentes compatibles (71.0 %). Esto es consistente con la expectativa de que los tumores de HCC estén hipometilados a nivel de todo el genoma en comparación con los tejidos normales adyacentes (Ross et al. Epigenomics. 2010;2:245-69). El conjunto de datos se generó a partir de ADN preparado con el Kit de Secuenciación Sequel II 1.0.

Se sometieron porciones de los mismos tejidos a análisis de metilación utilizando una secuenciación en tiempo real de única molécula y los métodos de acuerdo con esta divulgación. Los resultados se muestran en la FIG. 62B. El análisis de metilación utilizando los métodos de secuenciación en tiempo real de única molécula de esta divulgación pudo mostrar que la estirpe celular HepG2 era la más hipometilada, seguida por el tejido tumoral de HCC analizado y

luego seguido por el tejido placentario. La muestra de tejido hepático no tumoral adyacente estaba más metilada que los otros tejidos, que incluyen el HCC y los tejidos placentarios, siendo la capa leucocitaria la más hipermetilada.

Las FIG. 63A, 63B y 63C muestran la correlación de los niveles de metilación generales cuantificados mediante secuenciación con bisulfito y secuenciación en tiempo real de única molécula de acuerdo con los métodos descritos en el presente documento. La FIG. 63A muestra el nivel de metilación cuantificado mediante secuenciación con bisulfito en el eje x y el nivel de metilación cuantificado mediante secuenciación en tiempo real de única molécula utilizando los métodos descritos en el presente documento en el eje y. La línea negra continua es una línea de regresión ajustada. La línea discontinua es donde las dos medidas son iguales.

Hubo una correlación muy alta de los niveles de metilación entre la secuenciación con bisulfito y la secuenciación en tiempo real de única molécula de acuerdo con la invención divulgada en el presente documento ($r = 0.99$; valor de $P < 0.0001$). Estos datos indicaron que el análisis de metilación utilizando los métodos de secuenciación en tiempo real de molécula única divulgados en la presente fueron medios eficaces para determinar los niveles de metilación entre tejidos y permitieron la comparación de los estados y perfiles de metilación entre estos tejidos. Para dos medidas de niveles de metilación, observamos que la pendiente de la línea de regresión en la FIG. 63A se desvió de uno. Estos resultados sugirieron que existe una desviación entre las dos mediciones (en algún contexto, esta desviación se puede denominar sesgo) que podría estar presente en la determinación de los niveles de metilación utilizando secuenciación en tiempo real de única molécula de acuerdo con la divulgación en comparación con la Secuenciación convencional masivamente paralela con bisulfito.

Describimos cómo podríamos cuantificar el sesgo utilizando regresión lineal o LOESS (suavización ponderada localmente). Como un ejemplo, si consideráramos la secuenciación masiva de bisulfito paralela (Illumina) como una referencia, los resultados determinados por la secuenciación en tiempo real de única molécula de acuerdo con la divulgación se podrían transformar utilizando los coeficientes de regresión, conciliando de esta manera las lecturas entre diferentes plataformas. En la Fig. 63A, la fórmula de regresión lineal era $Y=aX+b$, donde "Y" representaba los niveles de metilación determinados por secuenciación en tiempo real de única molécula de acuerdo con la divulgación; "X" representó los niveles de metilación determinados mediante secuenciación con bisulfito; "a" representaba la pendiente de la línea de regresión (por ejemplo, $a= 0.62$); "b" representaba la intersección en el eje y (por ejemplo, $b= 17.72$). En esta situación, los valores de metilación reconciliados determinados por una secuenciación en tiempo real de única molécula se calcularían mediante (Y-b)/a. Alternativamente, se podría utilizar la relación de la desviación entre dos mediciones (ΔM) y el promedio correspondiente de las dos mediciones ($\bar{M}$), que se definieron mediante las fórmulas (1) y (2) a continuación:

$$\Delta M = S - Metilación\ basada\ en\ bisulfito, \quad (1)$$

$$\bar{M} = \frac{S + Metilación\ basada\ en\ bisulfito}{2}, \quad (2)$$

donde "S" representa el nivel de metilación determinado por una secuenciación en tiempo real de única molécula como se describió anteriormente y "Metilación basada en bisulfito" representa el nivel de metilación determinado por secuenciación con bisulfito.

Con fines ilustrativos, la FIG. 63B muestra la relación entre ΔM y $\bar{M}$. El promedio de las dos mediciones ($\bar{M}$) se grafica en el eje x, y la desviación entre las dos mediciones (ΔM) se grafica en el eje y. La línea discontinua representa una línea horizontal que cruza el cero en la que un punto de datos sugiere que no hay diferencia entre dos mediciones. Estos resultados sugirieron que la desviación varió dependiendo de los valores promediados. Cuanto mayor sea el promedio de las dos mediciones, mayor será la magnitud de la desviación. La mediana de los valores de ΔM fue -8.5 % (rango: -12.6 % a +2.5 %), lo que sugiere que existía discrepancia entre los métodos.

La FIG. 63C muestra el promedio de las dos mediciones ($\bar{M}$) en el eje x y la desviación relativa (RD) en el eje y. La desviación relativa se define mediante la siguiente fórmula:

$$RD = \frac{\Delta M}{\bar{M}} \times 100\%, \quad (3).$$

La línea discontinua representa una línea horizontal que cruza el cero en la que un punto de datos sugiere que no hay diferencia entre dos mediciones. Estos resultados sugirieron que la desviación relativa varió dependiendo de los valores promediados. Cuanto mayor sea el promedio de las dos medidas, mayor será la magnitud de la derivación relativa. La mediana de los valores de RD fue -12.5 % (rango: -18.1 % a +6.0 %).

Se informó que la secuenciación convencional con bisulfito del genoma completo (Illumina) introdujo una producción de secuencia sesgada significativa y sobrestimó la metilación global, con variaciones sustanciales en la cuantificación de los niveles de metilación entre métodos en regiones genómicas específicas (Olova et al. Genome Biol. 2018;19:33). Los métodos divulgados en el presente documento se pueden realizar sin conversión de bisulfito que degradaría drásticamente el ADN y se pueden realizar sin amplificación por PCR que puede complicar el proceso o puede introducir errores adicionales en la determinación de los niveles de metilación.

Con fines ilustrativos, las FIG. 64A y 64B muestran patrones de metilación con una resolución de 1 Mb. La FIG. 64A muestra el patrón de metilación para una estirpe celular HCC (HepG2). La FIG. 64B muestra el patrón de metilación para una muestra de capa leucocitaria de un sujeto de control sano. Los ideogramas de los cromosomas (anillo más externo en cada figura) están organizados desde pter hasta qter en el sentido de las agujas del reloj. El segundo anillo desde el exterior (también descrito como un anillo medio) muestra los niveles de metilación determinados mediante secuenciación con bisulfito. El anillo más interno muestra los niveles de metilación determinados por una secuenciación en tiempo real de única molécula de acuerdo con la divulgación. Los niveles de metilación se clasifican en 5 grados, a saber, 0-20 % (verde claro), 20-40 % (verde), 40-60 % (azul), 60-80 % (rojo claro) y 80-100 %. (rojo). Como se muestra en las FIG. 64A y 64B, los perfiles de metilación con una resolución de 1 Mb fueron consistentes entre la secuenciación con bisulfito (pista intermedia) y la secuenciación en tiempo real de única molécula (pista más interna) de acuerdo con la presente divulgación. Se mostró que el nivel de metilación de la muestra de capa leucocitaria materna era mayor que el de la estirpe celular HCC (HepG2).

Con fines ilustrativos, la FIG. 65A y 65B muestran diagramas de dispersión de los niveles de metilación medidos con una resolución de 1 Mb. La FIG. 65A muestra los niveles de metilación para la estirpe celular HCC (HepG2). La FIG. 65B muestra los niveles de metilación para una muestra de capa leucocitaria de un sujeto de control sano. Para ambas FIG. 65A y FIG. 65B, los niveles de metilación cuantificados mediante secuenciación con bisulfito están en el eje x, y los niveles de metilación medidos mediante secuenciación en tiempo real de única molécula de acuerdo con la presente divulgación están en el eje y. La línea continua es una línea de regresión ajustada. La línea discontinua es donde las dos técnicas de medición son iguales. Para la estirpe celular HCC, el nivel de metilación determinado mediante secuenciación en tiempo real de única molécula con una resolución de 1 Mb se correlacionó bien con el medido mediante secuenciación con bisulfito ($r = 0.99$; $P < 0.0001$) (Figura 65A). También se observó correlación para los datos de la muestra de capa leucocitaria ($r = 0.87$, $P < 0.0001$) (FIG. 65B).

Con fines ilustrativos, las FIG. 66A y 66B muestran diagramas de dispersión de los niveles de metilación medidos con una resolución de 100 kb. La FIG. 66A muestra los niveles de metilación para la estirpe celular HCC (HepG2). La FIG. 66B muestra los niveles de metilación para una muestra de capa leucocitaria de un sujeto de control sano. Para ambas FIG. 66A y FIG. 66, los niveles de metilación cuantificados mediante secuenciación con bisulfito están en el eje x, y los niveles de metilación medidos mediante secuenciación en tiempo real de única molécula de acuerdo con la presente divulgación están en el eje y. La línea continua es una línea de regresión ajustada. La línea discontinua es donde las dos técnicas de medición son iguales. El alto grado de correlación entre las mediciones cuantitativas de metilación entre los dos métodos con una resolución de 1 Mb (o 1 Mnt) también se observó cuando la resolución del análisis aumentó a cada ventana de 100 kb (o 100 knt). Todos estos datos indican que el enfoque en tiempo real de única molécula de esta divulgación es una herramienta eficaz para cuantificar los niveles de metilación o las densidades de metilación dentro de regiones genómicas, que varían en diferentes grados de resolución, por ejemplo a 1 Mb (o 1 Mnt). o a 100 kb (o 100 knt). Los datos también indican que el presente método descrito en el presente documento es una herramienta eficaz para evaluar los perfiles de metilación o los patrones de metilación entre regiones o entre muestras.

Con fines ilustrativos, la FIG. 67A y 67B muestran patrones de metilación con una resolución de 1 Mb. La FIG. 67A muestra el patrón de metilación para un tejido tumoral de HCC (TBR3033T). La FIG. 67B muestra el patrón de metilación para tejido normal adyacente (TBR3033N). Los ideogramas de los cromosomas (anillo más externo en cada figura) están organizados desde pter hasta qter en el sentido de las agujas del reloj. El segundo anillo desde el exterior (también descrito como un anillo medio) muestra los niveles de metilación determinados mediante secuenciación con bisulfito. El anillo más interno muestra los niveles de metilación determinados por una secuenciación en tiempo real de única molécula de acuerdo con la divulgación. Los niveles de metilación se clasifican en 5 grados, a saber, 0-20 % (verde claro), 20-40 % (verde), 40-60 % (azul), 60-80 % (rojo claro) y 80-100 %. (rojo). Como se muestra en la FIG. 67A, pudimos detectar la hipometilación en el ADN del tejido tumoral del HCC (TBR3033T), que se podría diferenciar del ADN del tejido hepático normal adyacente (TBR3033N) en la FIG. 67B. Los niveles y patrones de metilación determinados mediante secuenciación con bisulfito (pista intermedia) y secuenciación en tiempo real de única molécula (pista más interna) de acuerdo con la divulgación fueron consistentes. Se mostró que el nivel de metilación del ADN del tejido normal adyacente es mayor que el del ADN del tejido tumoral del HCC.

Con fines ilustrativos, las FIG. 68A y 68B muestran diagramas de dispersión de los niveles de metilación medidos con una resolución de 1 Mb. La FIG. 68A muestra los niveles de metilación para el tejido tumoral HCC (TBR3033T). La FIG. 68B muestra los niveles de metilación para el tejido normal adyacente. Para ambas FIG. 68A y FIG. 68B, los niveles de metilación cuantificados mediante secuenciación con bisulfito están en el eje x, y los niveles de metilación medidos mediante secuenciación en tiempo real de única molécula de acuerdo con la presente divulgación están en el eje y. La línea continua es una línea de regresión ajustada. La línea discontinua es donde las dos técnicas de medición son iguales. Para el ADN del tejido tumoral de HCC, el nivel de metilación medido mediante secuenciación en tiempo real de única molécula con una resolución de 1 Mb se correlacionó bien con el determinado mediante secuenciación con bisulfito ($r = 0.96$; valor de $P < 0.0001$) (Figura 68A). Los datos de la muestra de tejido hepático normal adyacente también se correlacionaron ($r = 0.83$, valor de $P < 0.0001$) (FIG. 68B).

Con fines ilustrativos, las FIG. 69A y 69B muestran diagramas de dispersión de los niveles de metilación medidos con una resolución de 100 kb. La FIG. 69A muestra los niveles de metilación para el tejido tumoral HCC (TBR3033T). La FIG. 69B muestra niveles de metilación para tejido normal adyacente (TBR3033N). Para ambas FIG. 69A y FIG. 69B,

los niveles de metilación cuantificados mediante secuenciación con bisulfito están en el eje x, y los niveles de metilación medidos mediante secuenciación en tiempo real de única molécula de acuerdo con la presente divulgación están en el eje y. La línea continua es una línea de regresión ajustada. La línea discontinua es donde las dos técnicas de medición son iguales. También se observó un grado tan alto de correlación de los datos cuantitativos de metilación entre los dos métodos con una resolución de 1 Mb cuando la medición de los niveles de metilación se realizó con una resolución más alta, por ejemplo, en ventanas de 100 kb.

Con fines ilustrativos, las FIG. 70A y 70B muestran patrones de metilación con una resolución de 1 Mb para otros tejidos tumorales y tejidos normales. La FIG. 70A muestra el patrón de metilación para un tejido tumoral de HCC (TBR3032T). La FIG. 70B muestra el patrón de metilación para tejido normal adyacente (TBR3032N). Los ideogramas de los cromosomas (anillo más externo en cada figura) están organizados desde pter hasta qter en el sentido de las agujas del reloj. El segundo anillo desde el exterior (también descrito como un anillo medio) muestra los niveles de metilación determinados mediante secuenciación con bisulfito. El anillo más interno muestra los niveles de metilación determinados por una secuenciación en tiempo real de única molécula de acuerdo con la divulgación. Los niveles de metilación se clasifican en 5 grados, a saber, 0-20 % (verde claro), 20-40 % (verde), 40-60 % (azul), 60-80 % (rojo claro) y 80-100 %. (rojo). Como se muestra en la FIG. 70A, pudimos detectar la hipometilación en el ADN del tejido tumoral del HCC (TBR3032T), que se podría diferenciar del ADN del tejido hepático normal adyacente (TBR3032N) en la FIG. 70B. Los niveles y patrones de metilación determinados mediante secuenciación con bisulfito (pista intermedia) y secuenciación en tiempo real de única molécula utilizando la presente invención (pista más interna) fueron consistentes. Se mostró que el nivel de metilación del ADN del tejido normal adyacente es mayor que el del ADN del tejido tumoral del HCC.

Con fines ilustrativos, la FIG. 71A y 71B muestran diagramas de dispersión de los niveles de metilación medidos con una resolución de 1 Mb. La FIG. 71A muestra los niveles de metilación para el tejido tumoral HCC (TBR3032T). La FIG. 71B muestra los niveles de metilación para el tejido normal adyacente. Para ambas FIG. 71A y FIG. 71B, los niveles de metilación cuantificados mediante secuenciación con bisulfito están en el eje x, y los niveles de metilación medidos mediante secuenciación en tiempo real de única molécula de acuerdo con la presente divulgación están en el eje y. La línea continua es una línea de regresión ajustada. La línea discontinua es donde las dos técnicas de medición son iguales. Para el ADN del tejido tumoral de HCC, el nivel de metilación medido mediante secuenciación en tiempo real de única molécula con una resolución de 1 Mb se correlacionó bien con el determinado mediante secuenciación con bisulfito ($r = 0.98$; $P < 0.0001$) (Figura 71A). Los datos de la muestra de tejido hepático normal adyacente también se correlacionaron ($r = 0.87$, $P < 0.0001$) (FIG. 71B).

Con fines ilustrativos, las FIG. 72A y 72B muestran diagramas de dispersión de los niveles de metilación medidos con una resolución de 100 kb. La FIG. 72A muestra los niveles de metilación para el tejido tumoral HCC (TBR3032T). La FIG. 72B muestra niveles de metilación para tejido normal adyacente (TBR3032N). Para ambas FIG. 72A y FIG. 72B, los niveles de metilación cuantificados mediante secuenciación con bisulfito están en el eje x, y los niveles de metilación medidos mediante secuenciación en tiempo real de única molécula de acuerdo con la presente divulgación están en el eje y. La línea continua es una línea de regresión ajustada. La línea discontinua es donde las dos técnicas de medición son iguales. También se observó un grado tan alto de correlación de los datos cuantitativos de metilación entre los dos métodos con una resolución de 1 Mb cuando la medición de los niveles de metilación se realizó con una resolución más alta, por ejemplo, en ventanas de 100 kb.

4. Regiones de metilación diferencial entre el tumor y los tejidos normales adyacentes.

Las aberraciones metilómicas se encuentran a menudo en regiones de genomas de cáncer. Un ejemplo de dichas aberraciones es la hipometilación e hipermetilación de regiones genómicas seleccionadas (Cadieux et al. Cancer Res. 2006;66:8469-76; Graff et al. Cancer Res. 1995;55:5195-9; Costello et al. Nat Genet.2000;24:132-8). Otro ejemplo es el patrón aberrante de bases metiladas y no metiladas en regiones genómicas seleccionadas. Esta sección muestra que las técnicas para determinar la metilación se pueden utilizar para realizar análisis y diagnósticos cuantitativos en el análisis de tumores.

La FIG. 73 muestra un ejemplo del patrón aberrante de metilación cerca del gen supresor de tumores CDKN2A. Las coordenadas resaltadas en azul y subrayadas indican islas CpG. Los puntos rellenos negros indican sitios metilados. Los puntos sin relleno indican sitios no metilados. Los números entre paréntesis a la derecha de cada línea horizontal con puntos indican el tamaño del fragmento, la densidad de metilación molecular única y el número de sitios CpG. Por ejemplo, (3.3 kb, MD: 17.9 %, CG:39) significa que el tamaño de este fragmento es 3.3 kb, el nivel de metilación de este fragmento es 17.9 % y el número de sitios CpG es 39. MD representa la densidad de metilación.

Como se muestra en la FIG. 73, el gen CDKN2A (inhibidor de la quinasa dependiente de ciclina 2A) codifica dos proteínas, que incluyen INK4A (p16) y ARF (p14), que actúan como supresores de tumores. Había dos moléculas (molécula 7301 y molécula 7302) que cubrían la región que se superponía al gen CDKN2A en el tejido no tumoral adyacente al tejido tumoral. Se mostró que los niveles de metilación de la molécula de ADN de hebra doble única para la molécula 7301 y la molécula 7302 eran del 17.9 % y el 7.6 %, respectivamente. Por el contrario, se encontró que el nivel de metilación de la molécula de ADN de hebra doble simple para la molécula 7303 presente en el tejido tumoral era del 93.9 %, que era mucho más alto que el de las moléculas presentes en los tejidos no tumorales adyacentes emparejados. Por otro lado, también se podría calcular el nivel de metilación de múltiples hebras utilizando las

moléculas 7301 y 7302 presentes en el tejido no tumoral adyacente al tejido tumoral. Como resultado, el nivel de metilación de múltiples hebras fue del 9.7 %, inferior al del tejido tumoral (93.9 %). Los diferentes niveles de metilación sugieren que se podría utilizar el nivel de metilación de única molécula de hebra doble y/o el nivel de metilación de múltiples hebras para detectar o controlar enfermedades tales como el cáncer.

Las FIG. 74A y FIG. 74B muestran regiones de metilación diferencial detectadas por secuenciación en tiempo real de única molécula de acuerdo con realizaciones de la presente invención. La FIG. 74A muestra hipometilación en el genoma del cáncer. La FIG. 74B muestra hipermetilación en el genoma del cáncer. El eje x indica las coordenadas de los sitios CpG. Las coordenadas resaltadas en azul y subrayadas indican islas CpG. Los puntos rellenos negros indican sitios metilados. Los puntos sin relleno indican sitios no metilados. Los números entre paréntesis a la derecha de cada línea horizontal con puntos indican el tamaño del fragmento, la densidad de metilación a nivel de fragmento y el número de sitios CpG. Por ejemplo, (3.1 kb, MD: 88.9 %, CG: 180) significa que el tamaño de este fragmento es 3.1 kb, la densidad de metilación de este fragmento es 88.9 % y el número de sitios CpG es 180.

La FIG. 74A muestra una región cercana al gen GNAS que muestra más fragmentos hipometilados en el tejido tumoral de HCC en comparación con el tejido hepático normal adyacente. La FIG. 74B muestra una región cercana al gen ESR1 que exhibe un fragmento hipermetilado en el tejido de HCC, pero un fragmento de ADN del tejido no tumoral adyacente emparejado que se alinea con la región correspondiente mostró en su lugar hipometilación. Como se muestra en la FIG. 74B, los perfiles de metilación o los haplotipos de metilación de moléculas de ADN individuales fueron adecuados para revelar el estado de metilación aberrante de esas regiones genómicas, a saber, GNAS y ESR1, cuando se comparan muestras cancerosas con muestras no cancerosas.

Estos datos indican que el análisis de metilación de secuenciación en tiempo real de única molécula divulgado en el presente documento podría determinar el estado de metilación en cada sitio CpG (ya sea metilado o no metilado) en fragmentos de ADN individuales. La longitud de lectura de la secuenciación en tiempo real de única molécula es mucho más larga (del orden de kilobases de longitud) que la de la secuenciación de Illumina, que normalmente podría abarcar entre 100-300 nt de longitud por lectura (De Maio et al. Micob Genom. 2019; 5(9)). Combinando la propiedad de longitud de lectura larga de una secuenciación en tiempo real de única molécula con el método de análisis de metilación que hemos divulgado en la presente, se podría determinar fácilmente el haplotipo de metilación de múltiples sitios CpG que están presentes a lo largo de cualquier molécula de ADN. El perfil de metilación se refiere al estado de metilación de los sitios CpG desde una coordenada del genoma a otra coordenada dentro de un tramo contiguo de ADN (por ejemplo, en el mismo cromosoma, o dentro de un plásmido bacteriano, o dentro de un único tramo de ADN en un genoma de virus).

Debido a que la secuenciación en tiempo real de única molécula analiza cada molécula de ADN individualmente sin necesidad de amplificación previa, el perfil de metilación determinado para cualquier molécula de ADN individual es de hecho un haplotipo de metilación, es decir, el estado de metilación de los sitios CpG de un extremo a otro extremo de la misma molécula de ADN. Si se secuencian una o más moléculas de la misma región genómica, el % de metilación (a saber, nivel de metilación o densidad de metilación) de cada sitio CpG en todos los sitios CpG secuenciados en la región genómica se podría agregar a partir de los datos de los múltiples fragmentos de ADN utilizando la misma fórmula que se muestra en la FIG. 61. Se podría informar el % de metilación de cada sitio CpG para todos los sitios CpG secuenciados, que proporcionan el perfil de metilación de la región genómica secuenciada. Alternativamente, los datos se podrían agregar de todas las lecturas y todos los sitios dentro de la región genómica secuenciada para proporcionar un % de valor de metilación de la región, a saber de la misma manera en que se calcularon los niveles de metilación para las regiones de 1 Mb o 1 kb como se muestra en las FIG. 64 a 72.

5. Análisis de metilación del ADN viral.

Con fines ilustrativos, esta sección muestra que las técnicas de metilación de esta divulgación se pueden utilizar para determinar con precisión los niveles de metilación en el ADN viral.

La FIG. 75 muestra patrones de metilación del ADN del virus de la hepatitis B entre dos pares de muestras de tejido de HCC y muestras de tejido no tumoral adyacentes utilizando secuenciación en tiempo real de única molécula. Cada flecha representa una anotación genética en un genoma del HBV. Las flechas con 'P', 'S', 'X' y 'C' indican la anotación genética alrededor de un genoma del HBV: codifica la polimerasa, el antígeno de superficie, la proteína X y la proteína central, respectivamente. Identificamos un fragmento (molécula I) con un tamaño de 1,183 bp que se origina en tejidos no tumorales adyacentes, que abarca un genoma del HBV de 2,278 a 3,141 resaltado en un rectángulo discontinuo, que muestra un nivel de metilación del 12 %. También identificamos tres fragmentos (molécula II, III y IV) con 3,215 bp, 2,961 bp y 3,105 bp procedentes de tejidos tumorales. Entre ellos, dos fragmentos (molécula III y IV) en tumores de HCC se superpusieron con las regiones genómicas del HBV abarcadas por la molécula I en tejidos no tumorales. En contraste con el bajo nivel de metilación (12 %) en la región del HBV resaltada en un rectángulo discontinuo (ubicaciones genómicas del HBV: 2,278 – 3,141), los niveles de metilación fueron más altos para esos fragmentos (moléculas III y IV) en los tejidos del HCC (es decir, 24 % y 30 %). Estos resultados sugirieron que el enfoque que utiliza la secuenciación en tiempo real de única molécula era factible para determinar los patrones de metilación en el genoma viral y podía identificar la región metilada diferencialmente (DMR) del HBV entre los tejidos con y sin HCC. Por lo tanto, la determinación de los estados de metilación en genomas virales utilizando una secuenciación en tiempo

real de única molécula de acuerdo con la divulgación proporcionaría una nueva herramienta para estudiar la relevancia clínica utilizando biopsias de tejido.

Esta región DMR se superpuso con los genes P, C y S. Se informó que esta región también demostró estar hipermetilada en tejidos de HCC en comparación con la de tejidos hepáticos con infección por HBV pero sin cáncer (Jain et al. Sci Rep. 2015;5:10478; Fernández et al. Genome Res. 2009;19:438-51).

Agrupamos los resultados de la secuenciación con bisulfito de tejidos hepáticos de cuatro pacientes con cirrosis pero sin HCC, obteniendo 1,156 fragmentos de HBV para el análisis de metilación. La FIG. 76A muestra los niveles de metilación del ADN del virus de la hepatitis B en tejidos hepáticos de pacientes con cirrosis pero sin HCC. Además, agrupamos los resultados de la secuenciación con bisulfito de tejidos tumorales de HCC de 15 pacientes, obteniendo 736 fragmentos de HBV para el análisis de metilación. La FIG. 76B muestra niveles de metilación para el ADN del virus de la hepatitis B en tejido tumoral de HCC. Como se muestra en la FIG. 76A y FIG. 76B, también observamos una región DMR del HBV (ubicaciones genómicas del HBV: 1,982 – 2,435) que tenía un nivel de metilación más alto en los tejidos de HCC que en los tejidos hepáticos cirróticos mediante secuenciación masiva paralela con bisulfito. Estos resultados sugirieron que el enfoque para determinar el estado de metilación de los genomas virales sería válido.

6. Análisis de metilación asociado a variantes.

Se pueden asociar diferentes alelos con diferentes perfiles de metilación. Por ejemplo, los genes impresos pueden tener un alelo con un nivel de metilación más alto que el otro alelo. Con fines ilustrativos, esta sección muestra que los perfiles de metilación se pueden utilizar para distinguir alelos en determinadas regiones genómicas.

Un pocillo de secuenciación en tiempo real de única molécula que contenga una única plantilla de ADN generaría una serie de sublecturas. Las sublecturas incluyen características cinéticas [por ejemplo duración interpulso (IPD) y anchura de pulso (PW)] y composiciones de nucleótidos. Describimos cómo las sublecturas de una secuenciación en tiempo real de única molécula se pueden utilizar para generar una secuencia de consenso (también llamada secuencia de consenso circular, CCS) que puede reducir drásticamente los errores de secuenciación (por ejemplo, emparejamientos incorrectos, inserciones o supresiones). Detalles adicionales de CCS se describen en el presente documento. Describimos cómo se puede construir la secuencia consenso utilizando esas sublecturas alineadas con un genoma de referencia humano. Describimos cómo se podría construir la secuencia de consenso al mapear las sublecturas a la sublectura más larga en el mismo pocillo de secuenciación en tiempo real de única molécula.

Con fines ilustrativos, la FIG. 77 ilustra el principio del análisis de haplotipos de metilación en fases. Las paletas rellenas representan los sitios CpG que se clasifican como metilados. Las paletas sin rellenar representan los sitios CpG que se clasifican como no metilados.

Como se muestra en una realización en la FIG. 77, las sublecturas se alinearon con un genoma de referencia humano. Las sublecturas alineadas de un pocillo secuenciación en tiempo real de única molécula , se colapsaron para formar una secuencia de consenso. La secuencia consenso generalmente se podría determinar utilizando los nucleótidos más frecuentes presentes en sublecturas en cada posición alineada. Por lo tanto, las variantes de nucleótidos, que incluyen, pero no se limitan a, variantes, inserciones y supresiones de un solo nucleótido, se podrían identificar a partir de secuencias consenso. Las IPD y PW promediadas en la misma molécula marcada por una variante de nucleótido se podrían utilizar para determinar los patrones de metilación de acuerdo con la divulgación. Por lo tanto, podríamos determinar aún más los patrones de metilación asociados a variantes. Los estados de metilación en la misma molécula se podrían considerar un haplotipo de metilación. Es posible que el haplotipo de metilación no se construya fácil y directamente a partir de dos o más moléculas de ADN cortas porque puede que no exista un marcador molecular que permita diferenciar si dos o más moléculas de ADN cortas fragmentadas se derivan de una única molécula original o son aportadas por dos o más moléculas originales diferentes. Las tecnologías sintéticas de lectura larga (como la secuenciación de lectura vinculada desarrollada por 10X Genomics) ofrecen la posibilidad de distribuir una única molécula de ADN larga en una partición (tal como una gota) y etiquetar moléculas de ADN cortas, que se originan a partir de esa molécula de ADN larga, con la mismas secuencias de códigos de barras moleculares. Sin embargo, esta etapa del código de barras implica una amplificación por PCR que no preservaría los estados de metilación originales.

Además, si se intenta utilizar bisulfito para tratar las moléculas largas de ADN, la primera etapa antes del tratamiento con bisulfito implica la desnaturalización del ADN bajo condiciones destructivas, cambiando el ADN de hebra doble por ADN de hebra sencilla, ya que el bisulfito solo podría actuar sobre moléculas de ADN de hebra sencilla en determinadas condiciones químicas. Esta etapa de desnaturalización del ADN degradaría las moléculas largas de ADN en fragmentos cortos, lo que provocaría la pérdida de la información del haplotipo de metilación original. El segundo inconveniente del análisis de metilación basado en bisulfito desnaturalizaría el ADN de hebra doble en ADN de hebra sencilla en la etapa de conversión con bisulfito, a saber, las hebras de Watson y Crick. Para una molécula, hay un 50 % de posibilidades de secuenciar la hebra de Watson y un 50 % de posibilidades de secuenciar la hebra de Crick. Entre millones de hebras de Watson y Crick, existe una posibilidad extremadamente baja de secuenciar simultáneamente las hebras de Watson y Crick de una molécula. Aunque se supone que las hebras de Watson y Crick de una molécula están secuenciadas, aún es imposible determinar definitivamente si dichas hebras de Watson y Crick se derivan de un único fragmento original o si son aportadas por dos o más fragmentos originales diferentes. Liu et al introdujeron recientemente un método de secuenciación sin bisulfito para detectar citosinas metiladas e

hidroximetilcitosina (Liu et al. Nat Biotechnol. 2019;37:424-429) utilizando la conversión basada en enzimas de translocación diez-once (TET) en condiciones suaves, lo que lleva a una menor degradación del ADN. Sin embargo, implica dos etapas secuenciales de reacciones enzimáticas. Una tasa de conversión baja de cualquiera de las etapas de la reacción enzimática afectaría dramáticamente la tasa de conversión general. Además, incluso para este método de secuenciación sin bisulfito para detectar citosinas metiladas, aún existe la dificultad de distinguir las hebras de Watson y Crick de una molécula en los resultados de la secuenciación.

Por el contrario, como se describe en esta divulgación, las hebras de Watson y Crick de una molécula se ligan covalentemente mediante adaptadores en forma de campana para formar moléculas de ADN circulares. Como resultado, tanto las hebras de Watson como de Crick de una molécula se secuencian en el mismo pocillo de reacción y se pueden determinar los estados de metilación de cada hebra.

Una ventaja de lo que describimos es la capacidad de determinar la metilación y la información genética (es decir, secuencia) en una molécula de ADN contigua larga (por ejemplo, kilobases o kilonucleótidos de longitud). Es más difícil generar dicha información utilizando tecnologías de secuenciación de lectura corta. Para las tecnologías de secuenciación de lectura corta, es necesario combinar información de secuenciación en múltiples lecturas cortas utilizando andamios de firmas genéticas o epigenéticas, de tal manera que se pueda deducir una gran cantidad de información genética y de metilación. Sin embargo, esto podría resultar desafiante en muchos escenarios debido a las distancias entre dichos anclajes genéticos o epigenéticos. Por ejemplo, en promedio hay un SNP por 1 kb, mientras que las tecnologías actuales de secuenciación de lectura corta normalmente podrían secuenciar hasta 300 nt por lectura, lo que da como resultado 600 nt incluso en un formato de extremo emparejado.

Describimos cómo el análisis de haplotipos de metilación asociado a variantes se podría utilizar para estudiar los patrones de metilación en genes impresos. Las regiones impresas están sujetas a regulaciones epigenéticas (por ejemplo, metilación de CpG) de manera similar a la del progenitor de origen. Por ejemplo, una muestra de ADN de capa leucocitaria (M2) en la tabla de la FIG. 60 fue secuenciado para obtener alrededor de 152 millones de sublecturas. Para esta muestra, el 53 % de los pocillos de secuenciación en tiempo real de única molécula generaron al menos una sublectura que se podría alinear con un genoma de referencia humano. La profundidad media de la sublectura para cada pocillo de SMRT fue de 7.7x. En total, obtuvimos alrededor de 3 millones de secuencias consenso. Aproximadamente el 91 % del genoma de referencia estaba cubierto por secuencias consenso al menos una vez. Para las regiones cubiertas, la profundidad de secuenciación fue de 7.9 x. El conjunto de datos se generó a partir de ADN preparado con el Kit de Secuenciación Sequel II 1.0.

Con fines ilustrativos, la FIG. 78 muestra la distribución de tamaño de las moléculas secuenciadas determinadas a partir de secuencias consenso, con un tamaño de mediana de 6,289 bp (rango: 66 – 198,109 bp). El tamaño del fragmento (bp) se muestra en el eje x y la frecuencia (%) asociada con el tamaño del fragmento se muestra en el eje y.

Con fines ilustrativos, las FIG. 79A, 79B, 79C y 79D muestran ejemplos de patrones de metilación alélica en las regiones impresas. El eje x indica las coordenadas de los sitios CpG. Las coordenadas resaltadas en azul y subrayadas indican islas CpG. Los puntos rellenos de negro indican sitios CpG metilados. Los puntos sin relleno indican sitios CpG no metilados. El alfabeto incorporado entre cada serie horizontal de puntos rellenos y sin relleno (es decir, sitios CpG) indica el alelo en el sitio SNP. Los números entre paréntesis a la derecha de cada serie horizontal de puntos indican el tamaño del fragmento, la densidad de metilación a nivel de fragmento y el número de sitios CpG. Por ejemplo, (10.0 kb, MD:79.1 %, CG:139) sugirió que el tamaño del fragmento correspondiente era 10.0 kb, la densidad de metilación del fragmento era 79.1 % y el número de sitios CpG era 139. Los rectángulos discontinuos delinean el regiones más diferencialmente metiladas dentro de cada gen.

Con fines ilustrativos, la FIG. 79A muestra 11 fragmentos secuenciados con un tamaño de mediana de 11.2 kb (rango: 1.3 - 25 kb), que se originan desde el gen SNURF. El gen SNURF tenía impronta materna, lo que significa que la copia del gen que un individuo ha heredado de la madre está metilada y transcripcionalmente silenciosa. Como se muestra en la FIG. 79A, en el rectángulo discontinuo, los fragmentos asociados al alelo C estaban altamente metilados, mientras que los fragmentos asociados al alelo T estaban altamente desmetilados. Altamente metilado puede indicar que más del 70 %, 80 %, 90 %, 95 % o 99 % de los sitios están metilados. Los patrones de metilación específicos de alelo se pudieron observar en otros genes impresos, que incluyen PLAGL1 (FIG. 79B), NAP1L5 (FIG. 79C) y ZIM2 (FIG. 79D). La FIG. 79B muestra que con PLAGL1 los fragmentos asociados al alelo T estaban altamente desmetilados mientras que los fragmentos asociados al alelo C estaban altamente metilados. La FIG. 79C muestra que con NAP1L5 los fragmentos asociados al alelo C estaban altamente desmetilados y que los fragmentos asociados al alelo T estaban altamente metilados. La FIG. 79D muestra que con ZIM2 los fragmentos asociados al alelo C estaban altamente desmetilados y que los fragmentos asociados al alelo T estaban altamente metilados.

Con fines ilustrativos, las FIG. 80A, 80B, 80C y 80D muestran ejemplos de patrones de metilación alélica en regiones no impresas. El eje x indica las coordenadas de los sitios CpG. Las coordenadas resaltadas en azul y subrayadas indican islas CpG. Los puntos rellenos de negro indican sitios CpG metilados. Los puntos sin relleno indican sitios CpG no metilados. El alfabeto incorporado entre cada serie horizontal de puntos rellenos y sin relleno (es decir, sitios CpG) indica el alelo en el sitio del polimorfismo de un solo nucleótido (SNP). Los números entre paréntesis a la derecha de cada serie horizontal de puntos indican el tamaño del fragmento, la densidad de metilación a nivel de fragmento y el

número de sitios CpG. Los rectángulos discontinuos indican las regiones seleccionadas al azar para calcular las densidades de metilación informadas entre paréntesis. En contraste con los resultados de las FIG. 79A-79D, no hubo patrones de metilación alélica observables presentes en los genes no impresos. La FIG. 80A no muestra ningún patrón de metilación alélica diferente en una región chr7. La FIG. 80B no muestra ningún patrón de metilación alélica diferente en una región chr12. La FIG. 80C no muestra ningún patrón de metilación alélica diferente en una región chr1. La FIG. 80D no muestra ningún patrón de metilación alélica diferente en otra región chr1.

Con fines ilustrativos, la FIG. 81 muestra una tabla con niveles de metilación de fragmentos específicos de alelo. La primera columna enumera las categorías de "genes impresos" y "regiones seleccionadas aleatoriamente". La segunda columna enumera el gen particular. La tercera columna enumera el primer alelo de un SNP en el gen. La cuarta columna enumera el segundo alelo de un SNP en el gen. La quinta columna muestra el nivel de metilación de los fragmentos ligados al primer alelo. La sexta columna muestra el nivel de metilación de los fragmentos ligados al segundo alelo. Los niveles de metilación de los fragmentos ligados al alelo 2 (media: 88 6 %; rango 84.6 – 91.1 %) son mucho más altos que los de los fragmentos ligados al alelo 1 (media: 12:2 %; rango 7.6 – 15.7 %) para aquellos genes impresos (valor de P = 0.03), lo que indica la presencia de metilación específica alélica. Por el contrario, no hay cambios significativos en los niveles de metilación entre esas regiones seleccionadas al azar (valor de P = 1), lo que sugiere la ausencia de metilación específica alélica.

7. Análisis de ADN libre de células durante el embarazo

En esta ejemplificación, se demuestra que los métodos en el presente documento divulgados son aplicables al análisis de ácidos nucleicos libres de células en plasma o suero obtenidos de mujeres embarazadas con al menos un feto. Durante el embarazo, se encuentran en la circulación materna moléculas de ADN libres de células y ARN libres de células de las células placentarias. Dichas moléculas de ácido nucleico libres de células derivadas de la placenta también se denominan ácidos nucleicos fetales libres de células en el plasma materno o ácidos nucleicos fetales libres de células circulantes. Los ácidos nucleicos fetales libres de células están presentes en el plasma materno entre un fondo de ácidos nucleicos libres de células maternos. Por ejemplo, las moléculas de ADN fetal libre de células circulantes están presentes como una especie menor entre un fondo de ADN materno libre de células en el plasma y suero maternos.

Para distinguir el ADN fetal libre de células del ADN materno libre de células en plasma o suero materno, se sabe que se podrían utilizar medios genéticos o epigenéticos o una combinación. Genéticamente, el genoma fetal puede diferir del genoma materno por alelos SNP específicos del feto heredados por el padre, mutaciones heredadas por el padre o mutaciones de novo. Epigenéticamente, el metiloma placentario generalmente está hipometilado en comparación con el metiloma de las células sanguíneas maternas (Lun et al. Clin Chem. 2013;59:1583-94). Debido a que la placenta es el principal contribuyente de ADN fetal libre de células, mientras que las células sanguíneas maternas son el principal contribuyente de ADN materno libre de células en la circulación materna (plasma o suero), las moléculas de ADN fetal libre de células generalmente están hipometiladas en comparación con ADN materno libre de células en plasma o suero. Hay loci genómicos específicos donde la placenta está hipermetilada en comparación con las células sanguíneas maternas. Por ejemplo, el promotor y la región del exón 1 de RASSF1A están más metilados en la placenta que en las células sanguíneas maternas (Chiu et al. Am J Pathol. 2007;170:941-950). Por lo tanto, el ADN fetal libre de células circulante derivado de este locus RASSF1A estaría hipermetilado en comparación con el ADN materno libre de células circulante del mismo locus.

Describimos cómo el ADN fetal libre de células se puede distinguir de las moléculas de ADN materno libre de células en base al estado de metilación diferencial entre los dos agrupamientos de ácidos nucleicos circulantes. Por ejemplo, se encuentra que los sitios CpG a lo largo de una molécula de ADN libre de células están en su mayoría no metilados; es probable que esta molécula provenga del feto. Si se encuentra que los sitios CpG a lo largo de una molécula de ADN libre de células están en su mayoría metilados, es probable que esta molécula provenga de la madre. Los expertos en la técnica conocen varios métodos para determinar si dichas moléculas provienen realmente del feto o de la madre. Un enfoque es comparar el patrón de metilación de la molécula secuenciada con el perfil de metilación conocido del locus correspondiente en la placenta o en las células sanguíneas maternas.

Con fines ilustrativos, la FIG. 82 muestra un ejemplo para determinar el origen placentario del ADN plasmático durante el embarazo utilizando perfiles de metilación. Las coordenadas resaltadas en azul y subrayadas indican islas CpG. Los puntos rellenos negros indican sitios metilados. Los puntos sin relleno indican sitios no metilados. Los números entre paréntesis cerca de cada línea horizontal con puntos indican el tamaño del fragmento, la densidad de metilación molecular única y el número de sitios CpG.

Como se muestra en la FIG. 82, si la molécula de ADN libre de células plasmáticas maternas se alinea con la región promotora de RASSF1A (una región que se sabe que está específicamente metilada en los tejidos placentarios) y los datos de secuenciación generados utilizando los métodos de esta invención están hipermetilados, esta molécula probablemente derive del feto o la placenta. Por el contrario, las moléculas que muestran hipometilación probablemente derivan del ADN materno (predominantemente de origen hematopoyético).

Con fines ilustrativos, la FIG. 83 ilustra un enfoque para el análisis de metilación específico del feto. El enfoque incluye la utilización de la molécula secuenciada que contiene un alelo SNP específico del feto o una mutación específica del

feto (por ejemplo, heredada por vía paterna o de novo por naturaleza). Cuando se identifican dichas características genéticas específicas del feto, el estado de metilación de las bases presentes en la misma molécula de ADN libre de células refleja el perfil de metilación del ADN fetal libre de células o del metiloma placentario. Las características genéticas específicas del feto se pueden descubrir cuando la secuenciación del ADN libre de células plasmáticas revela alelos o mutaciones que no están presentes en el genoma materno (por ejemplo, al analizar el ADN genómico materno), o al analizar el ADN paterno o que se sabe que se transmite en la familia (por ejemplo, al analizar el ADN de un probando).

La metilación de moléculas de ADN específicas del feto se puede determinar al analizar aquellos fragmentos de ADN que llevan alelos que eran diferentes de los alelos homocigotos en el genoma materno. Se puede esperar que la metilación de las moléculas de ADN fetal sea menor que la de las moléculas de ADN materno.

Como un ejemplo, se secuenciaron el ADN de la capa leucocitaria de una mujer embarazada y su ADN placentario coincidente para obtener una cobertura del genoma haploide de 59x y 58x, respectivamente. Identificamos un total de 822,409 SNP informativos para los cuales la madre era homocigota y el feto heterocigoto. Encontramos 2,652 fragmentos fetales específicos y 24,837 fragmentos compartidos (es decir, los fragmentos que llevan el alelo compartido; predominantemente de origen materno) en el plasma materno (M13160) a través de la secuenciación en tiempo real de única molécula. La fracción de ADN fetal fue del 19.3 %. De acuerdo con la divulgación, se dedujeron los perfiles de metilación de esos fragmentos compartidos y específicos del feto. Como resultado, se encontró que el nivel de metilación de fragmentos específicos del feto era del 57.4 %, mientras que el nivel de metilación de los fragmentos compartidos era del 69.9 %. Este hallazgo fue consistente con el conocimiento actual de que el nivel de metilación del ADN fetal era menor que el del ADN materno en el plasma de una mujer embarazada (Lun et al., Clin Chem. 2013;59:1583-94).

Los patrones de metilación se pueden utilizar con fines de diagnóstico o monitorización. Por ejemplo, el perfil de metilación de una muestra de plasma materno se ha utilizado para determinar la edad gestacional https://www.ncbi.nlm.nih.gov/pubmed/27979959). Una aplicación es como una etapa de control de calidad. Otra aplicación potencial es monitorizar la edad “biológica” versus la “cronológica” de un embarazo. Esta aplicación podrá ser utilizada en la detección o evaluación de riesgos de parto prematuro. Se pueden utilizar otras realizaciones para el análisis de células fetales en sangre materna. En todavía otras realizaciones, dichas células fetales se pueden identificar mediante enfoques basados en anticuerpos o mediante tinción selectiva utilizando marcadores celulares (por ejemplo, sobre la superficie celular o en el citoplasma), o enriquecer mediante citometría de flujo o micromanipulación o microdisección o métodos físicos (por ejemplo, velocidad de flujo diferencial a través de una cámara, superficie o recipiente).

C. Detección de metilación utilizando diferentes reactivos

Esta sección muestra que las técnicas de metilación no se limitan a un sistema reactivo particular.

El análisis de metilación se realizó utilizando diferentes sistemas de reactivos para confirmar que se pueden aplicar técnicas. Como un ejemplo, SMRT-seq se realizó utilizando el sistema Sequel II (Pacific Biosciences) para llevar a cabo la secuenciación en tiempo real de única molécula. Las moléculas de ADN cortadas se sometieron a la construcción de una plantilla de secuenciación en tiempo real de única molécula (SMRT) utilizando un Kit SMRTbell Express Template Prep 2.0 (Pacific Biosciences). Las condiciones de hibridación del cebador de secuenciación y de unión de la polimerasa se calcularon con el software SMRT Link v8.0 (Pacific Biosciences). Brevemente, el cebador de secuenciación v2 se hibridó con la plantilla de secuenciación y luego se unió una polimerasa a las plantillas utilizando un Kit de Control Interno y Unión Sequel II 2.0 (Pacific Biosciences). La secuenciación se realizó en un Sequel II SMRT Cell 8M. Las películas de secuenciación se recopilaron en el sistema Sequel II durante 30 horas con un Kit de Secuenciación Sequel II 2.0 (Pacific Biosciences). En otras realizaciones, se utilizarían otros reactivos químicos y tampones de reacción para SMRT-seq. En una realización, una polimerasa tendría diferentes características cinéticas de incorporación de nucleótidos a lo largo de una hebra molde de ADN dependiendo de su estado de metilación (Huber et al. Nucleic Acids Res. 2016;44:9881-9890). En esta divulgación, los resultados se generan utilizando el cebador de secuenciación v1 a menos que se indique lo contrario.

Para demostrar el uso de la invención en la divulgación descrita en el presente documento con el uso de diferentes reactivos, analizamos los datos de SMRT-seq generados en base a diferentes kits de secuenciación, que incluyen, pero no se limitan a, Kit de Secuenciación Sequel I 3.0, RS II, Kit de Secuenciación Sequel II 1.0 y Kit de Secuenciación Sequel II 2.0. RS II incluye 150,000 ZMW por célula de SMRT. Sequel utiliza 1,000,000 ZMW por célula de SMRT. Sequel II utiliza 8 millones de ZMW por célula de SMRT con dos kits de secuenciación (1.0 y 2.0). Este análisis involucró dos conjuntos de datos. El primer conjunto de datos se preparó basándose en el ADN después de la amplificación del genoma completo, lo que representa el estado no metilado. El segundo tipo de conjunto de datos se preparó basándose en el ADN después del tratamiento con metiltransferasa M.SsssI, lo que representa el estado metilado. Estos datos se generaron utilizando el Kit de Secuenciación Sequel 3.0 en el secuenciador Sequel; y el Kit de Secuenciación Sequel II 1.0 y el Kit de Secuenciación Sequel II 2.0 en el secuenciador Sequel II. Por tanto, obtuvimos tres conjuntos de datos con perfiles cinéticos generados con los diferentes reactivos (por ejemplo, polimerasas). Cada conjunto de datos se dividió en un conjunto de datos de entrenamiento y un conjunto de datos de prueba para evaluar el rendimiento utilizando modelos CNN de acuerdo con esta divulgación de esta invención.

1. Ventanas de medición

Las FIG. 84A, 84B y 84C muestran el rendimiento de diferentes tamaños de ventana de medición en diferentes kits de reactivos para SMRT-seq en conjuntos de datos de entrenamiento que comprenden datos amplificados del genoma completo (sitios CpG no metilados) y datos tratados con M.SsssI (sitios CpG metilados). La tasa positiva verdadera se grafica en el eje y y la tasa positiva falsa se grafica en el eje x. La FIG. 84A muestra datos de SMRT-seq generados en base al Kit de Secuenciación Sequel 3.0. La FIG. 84B muestra datos de SMRT-seq generados en base al Kit de Secuenciación Sequel II 1.0. La FIG. 84C muestra datos de SMRT-seq generados en base al Kit de Secuenciación Sequel II 2.0. En las figuras, "-" indica señales en dirección ascendente de un sitio de citosina CpG que se está analizando. '+' indicó señales en dirección descendente de un sitio de citosina CpG que se estaba analizando. Por ejemplo, "-6 nt" representaba las señales de 6 nt en dirección ascendente de un sitio de citosina CpG que se estaba analizando. '+6 nt' representaba las señales de 6 nt en dirección descendente de un sitio de citosina CpG que se estaba analizando. '±6 nt' indica que incluye señales en dirección ascendente de 6 nt y señales en dirección descendente de 6 nt de un sitio de citosina CpG que se está analizando (es decir, un total de una secuencia de 12 nt que flanquea un sitio de citosina CpG).

Para el conjunto de datos de entrenamiento basado en el Kit de Secuenciación Sequel 3.0, como se muestra en la FIG. 84A, utilizando la ventana de medición que comprende señales en una citosina CpG que se está analizando y señales en dirección ascendente de 6 nt (por ejemplo, IPD, PW, posiciones relativas y composiciones de secuencia) de ese sitio de citosina (indicado por -6 nt), el valor de AUC de 0.50 sugirió sin poder discriminativo para diferenciar las citosinas CpG metiladas de las no metiladas. Sin embargo, para los conjuntos de datos de entrenamiento basados en el Kit de Secuenciación Sequel II 1.0 y 2.0, los valores de AUC correspondientes fueron 0.62 (FIG. 84B) y 0.75 (FIG. 84C). Estos datos demostraron que existían diferentes perfiles cinéticos inherentes a los diferentes reactivos utilizados en SMRT-seq. Estos datos muestran que los métodos divulgados en el presente documento se adaptan fácilmente al uso de diferentes reactivos. Además, la precisión de la detección de modificaciones de bases se puede mejorar potencialmente con nuevos desarrollos en reactivos, por ejemplo, el uso de diferentes polimerasas y otras químicas.

Como otro ejemplo, para el conjunto de datos de entrenamiento basado en el Kit de Secuenciación Sequel 3.0, como se muestra en la FIG. 84A, utilizando una ventana de medición que comprende señales en dirección ascendente de 10 bp de un sitio de citosina CpG (indicado por -10 nt), el valor de AUC de 0.50 sugirió que no hay poder discriminativo para diferenciar las citosinas CpG metiladas de las no metiladas. Sin embargo, para los conjuntos de datos de entrenamiento basados en el Kit de Secuenciación Sequel II 1.0 y 2.0, los valores AUC correspondientes fueron 0.66 (FIG. 84B) y 0.79 (FIG. 84C), lo que mostró ser mejorado en comparación con la ventana de medición que comprende señales en dirección ascendente de 6 nt. Estos datos confirmaron que existían diferentes perfiles cinéticos inherentes a diferentes reactivos que se utilizaron para SMRT-seq. Estos datos muestran que los métodos divulgados en el presente documento se adaptan fácilmente al uso de diferentes reactivos.

A diferencia de la ventana de medición con señales en dirección ascendente, la ventana de medición con señales en dirección descendente podría conducir a una mayor mejora del rendimiento de clasificación. Por ejemplo, para el conjunto de datos de entrenamiento basado en Kit de Secuenciación Sequel 3.0, como se muestra en la FIG. 84A, utilizando una ventana de medición que comprende señales de 6 nt en dirección descendente de un sitio de citosina CpG (+6 nt), el valor de AUC de 0.94 fue mucho mayor que el que utiliza señales de 6 nt en dirección ascendente (AUC: 0.5). Para los conjuntos de datos de entrenamiento basados en el Kit de Secuenciación Sequel II 1.0 y 2.0, los valores AUC correspondientes fueron 0.95 (FIG. 84B) y 0.92 (FIG. 84C), respectivamente, lo que muestra una mejora en comparación con la ventana de medición que comprende 6 nt en dirección ascendente. Estos datos sugirieron que las características cinéticas ligadas al contexto de la secuencia mejorarían el poder de clasificación utilizando, pero no se limitan a, modelos CNN. Estos datos también sugirieron que la divulgación en el presente documento sería aplicable a conjuntos de datos producidos por diferentes reactivos y condiciones de secuenciación (por ejemplo, diferentes polimerasas, otros reactivos químicos, sus concentraciones y parámetros de reacción de secuenciación (por ejemplo, duración)), a través del ajuste de las ventanas de medición. Se sacaría una conclusión similar del análisis utilizando la ventana de medición que incluye señales de 10 nt en dirección descendente de un sitio de citosina CpG (FIGS. 84A, 84B y 84C).

En otra realización, se podría utilizar una ventana de medición que comprenda señales de citosina que se están analizando, y señales tanto en dirección ascendente como descendente de esa citosina. Por ejemplo, como se muestra en las FIG. 84A, 84B y 84C, utilizando una ventana de medición que comprende señales en dirección ascendente de 6 nt y señales en dirección descendente de 6 nt (indicadas por ±6 nt), se encontró que los valores de AUC eran 0.94, 0.95 y 0.92 para el conjunto de datos de entrenamiento basado en el Kit de Secuenciación Sequel. 3.0, Kit de Secuenciación Sequel II 1.0 y 2.0, respectivamente. Utilizando una ventana de medición que comprende señales en dirección ascendente de 10 nt y señales en dirección descendente de 10 nt (indicadas por ±10 nt), se encontró que los valores de AUC eran 0.94, 0.95 y 0.94 para el conjunto de datos de entrenamiento basado en Kit de Secuenciación Sequel 3.0, Kit de Secuenciación Sequel II 1.0 y 2.0, respectivamente. Estos datos sugirieron que la divulgación en el presente documento sería ampliamente aplicable a conjuntos de datos producidos por diferentes reactivos y parámetros de reacción de secuenciación.

Las FIG. 85A, 85B y 85C mostraron que se obtuvieron resultados al probar conjuntos de datos con diferentes ventanas de medición en diferentes kits de secuenciación al aplicar modelos CNN entrenados a partir de los conjuntos de datos de entrenamiento. La tasa positiva verdadera se grafica en el eje y y la tasa positiva falsa se grafica en el eje x. El etiquetado en la leyenda es equivalente al etiquetado utilizado en las FIG. 84A, 84B y 84C. La FIG. 85A muestra datos SMRT-seq generados en base al Kit de Secuenciación Sequel 3.0. La FIG. 85B muestra datos de SMRT-seq generados en base al Kit de Secuenciación Sequel II 1.0. La FIG. 85C muestra SMRT-seq generado en base al Kit de Secuenciación Sequel II 2.0. Todas las conclusiones extraídas de los conjuntos de datos de entrenamiento podrían validarse en estos conjuntos de datos de prueba independientes que no participaron en el proceso de entrenamiento. Adicionalmente, entre tres conjuntos de datos de prueba independientes, los análisis de dos conjuntos de datos (2/3) que involucran el Kit de Secuenciación Sequel II 1.0 y 2.0 mostraron que el uso de la ventana de medición que incluye señales en dirección ascendente y descendente de 10 nt (indicadas por ±10 nt) superó a los demás.

2. Comparación con la secuenciación con bisulfito.

Las FIG. 86A, 86B y 86C muestran la correlación de los niveles de metilación generales cuantificados mediante secuenciación con bisulfito y SMRT-seq (Kit de Secuenciación Sequel II 2.0). La FIG. 86A muestra el nivel de metilación como un porcentaje cuantificado por SMRT-seq en el eje y. La FIG. 86B muestra el nivel de metilación como un porcentaje cuantificado mediante secuenciación con bisulfito en el eje x. La línea negra es una línea de regresión ajustada. La línea discontinua es la línea diagonal en la que las dos medidas son iguales. La FIG. 86B muestra un diagrama de Bland-Altman. El eje x indica el promedio de los niveles de metilación cuantificados por SMRT-seq de acuerdo con la divulgación y la secuenciación con bisulfito. El eje y indica la diferencia en el nivel de metilación entre SMRT-seq de acuerdo con la divulgación y la secuenciación con bisulfito (es decir, metilación de Pacific Biosciences - metilación basada en bisulfito). La línea discontinua corresponde a una línea horizontal que cruza el cero en la que no hay diferencia entre dos medidas. Los puntos de datos desviados de la línea discontinua sugieren que existen desviaciones entre las medidas. La FIG. 86C muestra el cambio porcentual con respecto al valor cuantificado mediante secuenciación con bisulfito. El eje x indica el promedio de los niveles de metilación cuantificados por SMRT-seq de acuerdo con la divulgación y la secuenciación con bisulfito. El eje y indica el porcentaje de la diferencia en los niveles de metilación entre dos medidas en relación con el promedio de los niveles de metilación. La línea discontinua corresponde a una línea horizontal que cruza el cero en la que no hay diferencia entre dos medidas. Los puntos de datos desviados de la línea discontinua sugieren que existen desviaciones entre las medidas.

Para la FIG. 86A, la fórmula de regresión lineal fue Y=aX+b, donde “Y” representa los niveles de metilación determinados por SMRT-seq de acuerdo con la divulgación; “X” representa los niveles de metilación determinados mediante secuenciación con bisulfito; “a” representa la pendiente de la línea de regresión (por ejemplo, a= 1.45); “b” representa la intersección en el eje y (por ejemplo, b= -20.98). En esta situación, los valores de metilación determinados por SMRT-seq se calcularían mediante (Y-b)/a. Este gráfico muestra que los niveles de metilación determinados por SMRT-seq se pueden convertir en niveles de metilación determinados por secuenciación con bisulfito y viceversa para el Kit de Secuenciación Sequel II 2.0 como con el Kit de Secuenciación Sequel II 1.0.

La FIG. 86B es un gráfico de Bland-Altman que muestra el sesgo de la cuantificación de la metilación entre SMRT-seq de acuerdo con la divulgación y la secuenciación con bisulfito, en el que el eje x indica el promedio de los niveles de metilación cuantificados por SMRT-seq de acuerdo con la divulgación y la secuenciación con bisulfito. y el eje y indica la diferencia en los niveles de metilación cuantificados por SMRT-seq de acuerdo con la divulgación y la secuenciación con bisulfito. La diferencia de mediana entre las dos mediciones fue del -6.85 % (rango: -10.1 – 1.7 %). El cambio porcentual de mediana de un nivel de metilación cuantificado mediante la presente divulgación con respecto al valor mediante secuenciación con bisulfito fue -9.96 % (rango: -14.76 – 3.21 %). La diferencia varió dependiendo de los valores promediados. Cuanto mayor sea el promedio de dos medidas, mayor será el sesgo.

La FIG. 86C muestra los mismos datos que la FIG. 86B, pero con la diferencia en los niveles de metilación dividida por el promedio de los dos niveles de metilación. La FIG. 86C también muestra que cuanto mayor sea el promedio de las dos medidas, mayor será el sesgo.

El error puede estar relacionado con la secuenciación con bisulfito y no estar relacionado con los métodos con SMRT-seq. Se informó que la secuenciación convencional con bisulfito del genoma completo (Illumina) introdujo un resultado de secuencia significativamente sesgado y sobrestimó la metilación global, con variaciones sustanciales en la cuantificación de los niveles de metilación entre métodos en regiones genómicas específicas (Olova et al. Genome Biol. 2018;19: 33). Las realizaciones divulgadas en el presente documento tienen una serie de ventajas ejemplares mediante las cuales se pueden realizar sin conversión con bisulfito que degradaría drásticamente el ADN y se pueden realizar sin amplificación por PCR.

3. Origen del tejido

Realizamos el análisis de metilación en varios tipos de cáncer de acuerdo con las realizaciones en esta divulgación utilizando secuenciación en tiempo real de única molécula (SMRT-seq, Pacific Biosciences). Los tipos de cáncer utilizados para SMRT-seq incluyen, pero no se limitan a, cáncer colorrectal (n=3), cáncer de esófago (n=2), cáncer de mama (n=2), carcinoma de células renales (n=2), cáncer de pulmón. (n=2), cáncer de ovario (n=2), cáncer de próstata (n=2), cáncer de estómago (n=2) y cáncer de páncreas (n=1). Sus tejidos no tumorales adyacentes coincidentes

también se incluyeron para SMRT-seq. El conjunto de datos se generó a partir de ADN preparado con el Kit de Secuenciación Sequel II 2.0.

Las FIG. 87A y 87B muestran una comparación del nivel de metilación global entre diversos tejidos tumorales y tejidos no tumorales adyacentes emparejados. El nivel de metilación como porcentaje está en el eje y. En la Fig. 87A, el nivel de metilación se cuantifica mediante SMRT-seq. En la Fig. 87B, los niveles de metilación cuantificados mediante secuenciación con bisulfito. El tipo de tejido (es decir, tejido tumoral o tejido no tumoral adyacente) está en el eje x. Los diferentes símbolos representan diferentes tejidos de origen.

La FIG. 87A muestra que los niveles generales de metilación de los tejidos tumorales, que incluyen el cáncer de mama, cáncer colorrectal, cáncer de esófago, cáncer de hígado, cáncer de pulmón, cáncer de ovario, cáncer de páncreas, carcinoma de células renales y cáncer de estómago, fueron significativamente más bajos que los tejidos no tumorales correspondientes (valor de p = 0.006, muestras pareadas, prueba de rangos con signo de Wilcoxon), que incluyen mama, colon, esófago, hígado, pulmón, ovario, páncreas, próstata, riñón y estómago, respectivamente. La diferencia mediana en el nivel de metilación entre los tejidos tumorales y no tumorales pareados fue del -2,7 % (IQR: -6.4 ~ -0.8 %).

La FIG. 84B confirma niveles más bajos de metilación en tejidos tumorales. Por lo tanto, estos resultados sugirieron que los patrones de metilación en varios tipos y tejidos de cáncer se podrían determinar con precisión mediante SMRT-seq de acuerdo con la divulgación, lo que implica una aplicación amplia de esta divulgación para la detección temprana, el pronóstico, el diagnóstico y el tratamiento del cáncer, en base a la biopsia de tejido. Los diferentes grados de reducción del nivel de metilación en los distintos tipos de tumores probablemente sugirieron que los patrones de metilación estaban asociados con los tipos de cáncer, lo que permitió determinar el tejido de origen de un cáncer.

D. Detección potenciada y otras técnicas

En algunas realizaciones de la invención, el análisis de la metilación se puede realizar utilizando uno o más de los siguientes parámetros: el contexto de secuencia, la IPD y PW. IPD y PW se pueden determinar a partir de la reacción de secuenciación, sin alineación con un genoma de referencia. Algunos aspectos del enfoque de secuenciación en tiempo real de única molécula pueden potenciar aún más la precisión de la determinación del contexto de la secuencia, la IPD y la PW. Un aspecto es el rendimiento de la secuenciación por consenso circular en la que una porción particular de una plantilla de secuenciación se puede medir múltiples veces, permitiendo de esta manera medir el contexto de la secuencia, IPD y PW en base al promedio o la distribución de valores a través de estas múltiples lecturas. En ciertas realizaciones, el análisis de la modificación de la base sin un proceso de alineación puede aumentar la eficiencia computacional, reducir el tiempo de respuesta y puede reducir los costes del análisis. Si bien las realizaciones se pueden realizar sin un proceso de alineación, todavía en otras realizaciones, se puede utilizar un proceso de alineación y puede ser preferible, por ejemplo, si el proceso de alineación se utiliza para determinar las implicaciones clínicas o biológicas de la modificación de base detectada (por ejemplo, si un supresor de tumores está hipermetilado); o si el proceso de alineación se utiliza para seleccionar un subconjunto de los datos de secuenciación que corresponde a ciertas regiones genómicas de interés para su análisis adicional. Para realizaciones en las que se desean datos de regiones genómicas seleccionadas, estas realizaciones pueden implicar dirigirse a dichas regiones utilizando una o más enzimas o metodologías basadas en enzimas que pueden escindir en regiones de interés en el genoma, por ejemplo, una enzima de restricción o un sistema CRISPR-Cas9. El sistema CRISPR-Cas9 puede ser preferible al método basado en PCR, ya que la amplificación por PCR normalmente no conserva información sobre las modificaciones de bases del ADN. Los niveles de metilación de dichas regiones seleccionadas (ya sea bioinformáticamente [por ejemplo, a través de alineación] o mediante métodos como CRISPR-Cas9) se pueden analizar para proporcionar información sobre el origen del tejido, trastornos fetales, trastornos del embarazo y cáncer.

1. Análisis de metilación utilizando sublecturas sin alineación con un genoma de referencia

En realizaciones, el análisis de metilación se podría realizar utilizando las ventanas de medición que comprenden características cinéticas y contexto de secuencia de sublecturas sin alineación con un genoma de referencia. Como se muestra en la FIG. 88, se utilizaron sublecturas provenientes de una guía de ondas de modo cero (ZMW) para construir una secuencia consenso 8802 (también conocida como secuencia consenso circular, CCS). Se calcularon los valores cinéticos promedio en cada posición en un CCS, que incluyen, pero no se limitan a, los valores de PW e IPD. El contexto de secuencia que rodea un sitio CpG se determinó a partir de CCS en base a las secuencias en dirección ascendente y en dirección descendente de ese sitio CpG. Por lo tanto, se construiría una ventana de medición como se define en esta divulgación para el entrenamiento, con la ventana de medición que incluye valores PW, IPD y contexto de secuencia de acuerdo con las sublecturas con características cinéticas relativas a CCS. Este procedimiento evita la alineación de sublecturas con un genoma de referencia.

Para probar el principio mostrado en la FIG. 88, utilizamos 601,942 sitios CpG no metilados que se originaron a partir de ADN amplificado del genoma completo y 163,527 sitios CpG metilados que se originaron a partir de ADN tratado con CpG metiltransferasa (por ejemplo, M.SssI), formando el conjunto de datos de entrenamiento. Utilizamos 546,393 sitios CpG no metilados que se originaron a partir de ADN amplificado del genoma completo y 193,641 sitios CpG metilados que se originaron a partir de ADN tratado con CpG metiltransferasa (por ejemplo, M.SssI), formando el

conjunto de datos de prueba. El conjunto de datos se generó a partir de ADN preparado con el Kit de Secuenciación Sequel II 2.0.

Como se muestra en la FIG. 89, en una realización, utilizando características cinéticas y contexto de secuencia asociados con sublecturas y CCS para entrenar el modelo de red neuronal convolucional (CNN) para determinar la metilación, se podría lograr un valor de AUC de 0.94 y 0.95 para diferenciar sitios CpG metilados de sitios CpG no metilados. en los conjuntos de datos de prueba y entrenamiento, respectivamente. En otras realizaciones, se podrían utilizar otros modelos de redes neuronales, algoritmos de aprendizaje profundo, inteligencia artificial y/o algoritmos de aprendizaje automático.

Si establecemos un valor de corte de 0.2 para la probabilidad de metilación, podríamos obtener una sensibilidad del 82.4 % y una especificidad del 91.7 % en la detección de sitios CpG metilados. Estos resultados ilustraron que se podrían diferenciar los sitios CpG metilados y no metilados utilizando sublecturas con características cinéticas sin la alineación previa con un genoma de referencia.

En otra realización, para determinar el estado de metilación en los sitios CpG, también se podrían utilizar las características cinéticas junto con el contexto de secuencia directamente de sublecturas sin información CCS y alineación previa con un genoma de referencia. Utilizamos características cinéticas que incluyen valores de PW e IPD en posiciones que abarcan 20 nt en dirección ascendente y 20 nt en dirección descendente de un CpG presente en una sublectura para entrenar un modelo CNN para determinar el estado de metilación. Como se muestra en la FIG. 90, de acuerdo con las realizaciones en esta divulgación, un AUC de la curva ROC que utiliza características cinéticas relacionadas con sublecturas fue de 0.70 y 0.69 para detectar sitios CpG metilados en conjuntos de datos de entrenamiento y prueba, respectivamente. Estos datos sugirieron que sería factible utilizar las realizaciones en esta divulgación para inferir los patrones de metilación para una molécula de ADN utilizando características cinéticas asociadas con sublecturas pero sin una alineación y construcción previa de secuencias consenso. Sin embargo, el rendimiento de la determinación de la metilación en esta realización fue inferior al de las realizaciones que utilizan combinatoriamente la información de alineación o secuencias consenso como se describe en esta divulgación. Se podría imaginar que la precisión potenciada en la generación de sublecturas y valores cinéticos mejoraría el rendimiento de la determinación de las modificaciones base utilizando sublecturas y sus características cinéticas asociadas.

2. Análisis de metilación de regiones suprimidas utilizando secuenciación en tiempo real de una única molécula dirigida

Los métodos descritos en el presente documento también se pueden aplicar para analizar una o más regiones genómicas seleccionadas. En una realización, la(s) región(es) de interés se pueden enriquecer primero mediante un método de hibridación que permite la hibridación de moléculas de ADN de las regiones de interés con oligonucleótidos sintéticos con secuencias complementarias. Para el análisis de la metilación utilizando los métodos descritos en el presente documento, las moléculas de ADN diana no se pueden amplificar mediante PCR antes de someterlas a secuenciación porque la información de metilación en la molécula de ADN original no se transferiría a los productos de la PCR. Se han desarrollado varios métodos para enriquecer estas regiones objetivo sin realizar una amplificación por PCR.

En otra realización, la(s) región(es) diana se pueden enriquecer mediante el uso del sistema CRISPR-Cas9 (Stevens et al. PLOS One 2019;14(4):e0215441; Watson et al. Lab Invest 2020;100: 135-146). En una realización, los extremos de las moléculas de ADN en una muestra de ADN se desfosforilan primero, de tal manera que no sean susceptibles a la ligación directa a adaptadores de secuenciación. Luego, la proteína Cas9 dirige la(s) región(es) de interés con ARN guía (ARNcr) para crear cortes de hebra doble. La(s) región(es) de flanqueo interesado por cortes de hebra dobles en ambos lados se ligarían luego a los adaptadores de secuenciación especificados por la plataforma de secuenciación elegida. En otra realización, el ADN se puede tratar con exonucleasa para que las moléculas de ADN no unidas por las proteínas Cas9 se degraden (Stevens et al. PLOS One 2019;14(4):e0215441). Como estos métodos no implican amplificación por PCR, se pueden secuenciar las moléculas de ADN originales con metilación y se determinará la metilación. En una realización, este método se puede utilizar para dirigir un gran número de regiones que comparten secuencias homólogas, por ejemplo, las repeticiones largas de elementos nucleares intercalados (LINE). En un ejemplo, dicho análisis se puede utilizar para el análisis del ADN libre de células circulante en el plasma materno para la detección de aneuploidías fetales (Kinde et al. PLOS One 2012;7(7):e41162.

Como se muestra en la FIG. 91, la secuenciación en tiempo real de la molécula única dirigida se puede implementar mediante el uso del sistema CRISPR (repeticiones palindrómicas cortas agrupadas regularmente interespaciadas)/Cas9 (proteína 9 asociada a CRISPR). Los fragmentos de ADN (por ejemplo, la molécula 9102) que llevaban grupos fosforilo 5' (es decir, 5'-P) y grupos hidroxilo 3' (es decir, 3'-OH) se sometieron a un proceso de bloqueo terminal mediante el cual se eliminó el 5'-P y 3'-OH se ligó con didesoxinucleótidos (es decir, ddNTP). Por lo tanto, las moléculas resultantes (por ejemplo, la molécula 9104) cuyos extremos se modificaron no pudieron ligarse con adaptadores para la posterior preparación de la biblioteca de ADN. Sin embargo, las moléculas bloqueadas en los extremos se sometieron a una escisión específica diana mediada por el sistema CRISPR/Cas9, que introduce terminales 5'-P y 3'-OH en las moléculas de interés. Dichas moléculas de ADN recién escindidas (por ejemplo, la molécula 9106) que llevan terminales 5'-P y 3'-OH adquirieron la capacidad de ligarse con adaptadores de horquilla para formar la molécula circular 9108. Los adaptadores no ligados, el ADN lineal y las moléculas que solo llevan una

escisión se sometieron a digestión con exonucleasa III y VII. Como resultado, las moléculas ligadas con dos adaptadores de horquilla se enriquecieron y se sometieron a secuenciación en tiempo real de única molécula. Estas moléculas diana eran adecuadas para el análisis de metilación de acuerdo con las realizaciones presentes en esta divulgación (es decir, secuenciación en tiempo real de molécula única dirigida).

Como se muestra en la FIG. 92, la proteína Cas9 en el sistema CRISPR/Cas9 interactuó con el ARN guía (es decir, ARNg), que incluye ARN CRISPR (ARNcr, responsable de la orientación del ADN) y ARNcr transactivador (ARNtracr, responsable de formar el complejo con Cas9) (Pickar- Oliver et al. Nat Rev Mol Cell biol. 2019;20:490-507). La forma curva representa la proteína Cas9, que es una enzima que utiliza secuencias CRISPR como guía para reconocer y cortar hebras específicas de ADN que son complementarias a una parte de las secuencias CRISPR. El ARNcr se hibridó con ARNtracr. En una realización, una secuencia de ARN única sintética contenía secuencias de ARNcr y ARNtracr, denominada ARN guía única (ARNgu). Un segmento del ARNcr, denominado secuencia espaciadora, guiaría a la proteína Cas9 para reconocer y cortar hebras específicas de ADN de hebra doble (ADNhd), a través del emparejamiento de bases complementarias con la región dirigida. En una realización, no hubo emparejamientos incorrectos involucradas en la complementariedad entre la secuencia espaciadora y el ADNhd dirigido. En otra realización, el emparejamiento de bases complementarias entre la secuencia espaciadora y el ADNhd dirigido permitiría emparejamientos erróneos. Por ejemplo, el número de emparejamientos erróneos es, pero no se limita a, 1, 2, 3, 4, 5, 6, 7, 8, etc. En una realización, las secuencias CRISPR serían programables, dependiendo de la eficiencia de corte, la especificidad y la sensibilidad. y la capacidad de multiplexación para diferentes diseños complejos CRISPR/Cas.

Como se ilustra en la FIG. 93, diseñamos un par de complejos CRISPR/Cas9 dirigidos a dos cortes que abarcan un elemento Alu en un genoma humano. 'XXX' indica tres nucleótidos que flanquean el sitio de corte de la nucleasa Cas9. 'YYY' indica tres nucleótidos correspondientes complementarios a 'XXX'. 5'-NGG representa la secuencia del motivo adyacente al protoespaciador (PAM). En otros sistemas CRISPR/Cas, la secuencia PAM puede ser diferente y las secuencias que flanquean un sitio de corte de la nucleasa Cas pueden ser diferentes. En esta figura, una región Alu tenía un tamaño de 223 bp. Se presentan 1,175,329 regiones Alu, cada una de las cuales contenía homólogos de dicho elemento Alu en un genoma humano. Una mediana de 5 sitios CpG residían en este elemento Alu (rango: 0-34). Como un ejemplo, este diseño contenía un ARNcr de 36 nt que contenía una secuencia espaciadora de 20 nt. La información detallada de la secuencia de ARNg se muestra a continuación:

Un primer complejo CRISPR/Cas9 para introducir un primer corte: (todas las secuencias desde 5' hasta 3')

ARNcr: GCCUGUAAUCCCAGCACUUUGUUUUAGAGCUAUGCU

ARNtracr:

AGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACC

GAGUCGGUGCUUU

Un segundo complejo CRISPR/Cas9 para introducir un segundo corte:

ARNcr: AGGGUCUCGCUCUGUCGCCCGUUUUAGAGCUAUGCU

ARNtracr:

AGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACC

GAGUCGGUGCUUU

Las moléculas de ARNcr se hibridaron con un ARNtracr (por ejemplo, 67 nt) para formar la estructura principal de ARNg. La nucleasa Cas9 con ARNg diseñado puede escindir ambas hebras de moléculas bloqueadas en los extremos que albergan los sitios de corte específicos, con un cierto nivel de especificidad. Había 116,184 regiones Alu de interés en un genoma humano que se suponía que debían ser cortadas por los complejos CRISPR/Cas9 diseñados. Por lo tanto, aquellas regiones Alu después del corte dirigido por complejos Cas9 se pueden ligar con adaptadores de horquilla. Aquellas moléculas ligadas con adaptadores de horquilla se pueden secuenciar mediante secuenciación en tiempo real de única molécula. Los patrones de metilación para esas regiones Alu se pueden determinar de manera dirigida. En una realización, las secuencias espaciadoras de dos complejos Cas9 pueden tener pares de bases con la misma hebra (por ejemplo, hebra de Watson o hebra de Crick) de un sustrato de ADN de hebra doble. En una realización, las secuencias espaciadoras en ARNg de dos complejos Cas9 se pueden aparear por bases con las diferentes hebras de un sustrato de ADN de hebra doble. Por ejemplo, una secuencia espaciadora en un complejo Cas9 era complementaria a la hebra de Watson de un sustrato de ADN de hebra doble y la otra secuencia espaciadora en un complejo Cas9 era complementaria a la hebra de Crick de un sustrato de ADN de hebra doble, o viceversa.

En una realización, las moléculas de ADN ligadas con adaptadores de horquilla tenían forma circular, lo que sería resistente a la digestión con exonucleasa. Por lo tanto, se puede tratar el producto de ADN ligado al adaptador con exonucleasa (por ejemplo, exonucleasa III y VII) para eliminar el ADN lineal (por ejemplo, moléculas de ADN fuera de

diana). Esta etapa con el uso de exonucleasas puede enriquecer aún más las moléculas dirigidas. Los tamaños de las moléculas dirigidas que se van a secuenciar dependían del tamaño que abarca entre dos sitios de corte introducidos por una o más nucleasas Cas9, por ejemplo, que incluyen, pero no se limitan a, 10 bp, 20 bp, 30 bp, 40 bp, 50 bp, 100 bp, 200 bp, 300 bp, 400 bp, 500 bp, 1000 bp, 2000 bp, 3000 bp, 4000 bp, 5000 bp, 10 kb, 20 kb, 30 kb, 40 kb, 50 kb, 100 kb, 200 kb, 300 kb, 500 kb y 1 Mb.

Como un ejemplo, utilizando Cas9 con ARNg dirigido a regiones Alu, secuenciamos 187,010 moléculas de una muestra de tejido tumoral de carcinoma hepatocelular humano (HCC), utilizando secuenciación en tiempo real de única molécula. Entre ellas, 113,491 moléculas llevaban cortes dirigidos (es decir, la tasa de escisión en diana era de alrededor del 60.7 % de las moléculas). El conjunto de datos se generó a partir de ADN preparado con el Kit de Secuenciación Sequel II 2.0. En otras palabras, la especificidad de los sitios de corte introducidos en las moléculas de interés por los complejos Cas9 en este ejemplo fue del 60.7 %. En otras realizaciones, la especificidad de los sitios de corte introducidos en las moléculas de interés por Cas9 u otros complejos de Cas se variaría, que incluyen, pero no se limitan a, 1 %, 5 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 % y 100 %. Los valores de IPD, PW y el contexto de secuencia derivados de CCS y sublecturas sin alineación con un genoma de referencia se utilizaron para determinar el estado de metilación en sitios CpG en secuencias Alu.

Como se muestra en la FIG. 94, observamos una distribución de metilación similar entre los niveles de metilación determinados mediante secuenciación con bisulfito y secuenciación en tiempo real de única molécula de acuerdo con la divulgación. La FIG. 94 muestra histogramas de densidades de metilación (en porcentaje) para secuenciación con bisulfito y secuenciación en tiempo real de única molécula (Pacific Biosciences). El eje y indica la proporción de moléculas en la muestra con la densidad de metilación particular mostrada en el eje x. Este resultado sugirió que era factible determinar los patrones de metilación utilizando secuenciación en tiempo real de única molécula dirigida mediada por Cas9. Este resultado también sugirió que se podría determinar la metilación utilizando características cinéticas asociadas a sublecturas, que incluyen los valores de PW e IPD sin alineación con un genoma de referencia. Como se muestra en la FIG. 94, observamos una cantidad considerable de regiones Alu que mostraban hipometilación, lo que era consistente con el conocimiento previo de que el genoma del cáncer se desmetilaría en regiones repetidas de Alu (Rodríguez et al. Nucleic Acids Res. 2008; 36:770-784).

La FIG. 95 muestra la distribución de los niveles de metilación determinada por secuenciación en tiempo real de molécula única, de acuerdo con la divulgación en el eje y y la densidad de metilación determinada por secuenciación con bisulfito en el eje x. Como se muestra en la FIG. 95, los niveles de metilación en las regiones de Alu se agruparon en 5 categorías, a saber, 0 - 20 %, 20 - 40 %, 40 - 60 %, 60 - 80 % y 80 - 100 % de acuerdo con los resultados de la secuenciación con bisulfito. Nuestro modelo determinó además los niveles de metilación del mismo conjunto de regiones Alu utilizando las ventanas de medición que incluyen características cinéticas y contexto de secuencia (eje y) para cada categoría de regiones Alu. La distribución de los niveles de metilación determinados por nuestro modelo aumentó gradualmente de acuerdo con los órdenes ascendentes de los niveles de metilación en las categorías agrupadas. Nuevamente, estos resultados sugirieron que es factible determinar los patrones de metilación utilizando secuenciación en tiempo real de única molécula dirigida mediada por Cas9. Se puede determinar la metilación utilizando características cinéticas asociadas a sublecturas, que incluyen los valores de PW e IPD sin alineación con un genoma de referencia.

En todavía otra realización, se pueden utilizar otros tipos de sistemas CRISPR/Cas, por ejemplo, pero no se limitan a, Cas12a, Cas3 y otros ortólogos (por ejemplo, Cas9 de *Staphylococcus aureus*) o proteínas Cas diseñadas por ingeniería (*Acideminococcus* spp Cas12a potenciada) para realizar una secuenciación específica en tiempo real de única molécula.

En una realización, se puede utilizar Cas9 desactivado (dCas9), sin actividad nucleasa, para enriquecer las moléculas dirigidas sin escisión. Por ejemplo, las moléculas de ADN dirigidas estaban unidas por el complejo que comprende dCas9 biotinilado y ARNg específicos de la secuencia diana. Es posible que dCas9 no corte estas moléculas de ADN dirigidas porque dCas9 tenía deficiencia de nucleasa. A través del uso de perlas magnéticas recubiertas de estreptavidina, se pueden enriquecer las moléculas de ADN dirigidas.

En una realización, se pueden utilizar las exonucleasas para digerir la mezcla de ADN después de incubarla con proteínas Cas. Las exonucleasas pueden degradar las moléculas de ADN no unidas a la proteína Cas, mientras que las exonucleasas pueden no degradarse o pueden ser en gran medida menos eficientes en degradar las moléculas de ADN unidas a la proteína Cas. Por lo tanto, la información relativa a las moléculas diana unidas por las proteínas Cas puede enriquecerse aún más en los resultados finales de la secuenciación.

La FIG. 96 muestra una tabla de tejidos y los niveles de metilación de regiones Alu en los tejidos. Muchos tejidos muestran niveles de metilación en el rango del 85 %-92 %, que se incluye en el rango del 88 % al 92 %. El tejido tumoral de HCC y el tejido de placenta mostraron niveles de metilación inferiores al 80 %. Como se ve en la FIG. 96, se mostró que el tumor HCC estaba frecuentemente hipometilado en las regiones Alu que fueron dirigidas por nuestros diseños. Por lo tanto, la determinación de metilación de regiones Alu presentes en esta divulgación se puede utilizar para detectar, estadificar y monitorizar cánceres durante la progresión o el tratamiento del tumor utilizando ADN extraído de biopsias de tumores u otros tejidos o células.

La hipometilación de tejidos placentarios a través de regiones Alu se puede utilizar para realizar pruebas prenatales no invasivas utilizando el ADN plasmático de mujeres embarazadas. Por ejemplo, un mayor grado de hipometilación puede indicar una mayor fracción de ADN fetal en una mujer embarazada. En otro ejemplo, si una mujer está embarazada de un feto con aneuploidía cromosómica, el número de fragmentos de Alu procedentes de un cromosoma afectado detectado mediante este enfoque puede ser cuantitativamente diferente (es decir, aumentado o disminuido) que el de las mujeres embarazadas con fetos euploides. Por lo tanto, si un feto tiene trisomía 21, entonces el número de fragmentos de Alu que se originan en el cromosoma 21 detectados mediante este método puede aumentar en comparación con las mujeres embarazadas con fetos euploides. Por otro lado, si un feto tiene un cromosoma monosómico, entonces el número de fragmentos de Alu que se originan en ese cromosoma detectados mediante este método puede disminuir en comparación con las mujeres embarazadas con fetos euploides. En comparación con los cromosomas no afectados, la determinación de la presentación de hipometilación adicional de un cromosoma afectado (13, 18 o 21) en plasma se puede utilizar como indicador molecular para diferenciar mujeres embarazadas con fetos normales y anormales.

3. Análisis de metilación en las regiones Alu a las que se dirige el complejo Cas9 para diferentes tipos de cáncer

Aunque las repeticiones de Alu a las que nos dirigimos estaban altamente metiladas en diferentes tejidos, planteamos la hipótesis de que diferentes tipos de cáncer albergarían diferentes patrones de desmetilación en esas repeticiones de Alu. En una realización, se puede utilizar la secuenciación en tiempo real de una única molécula dirigida basada en Cas9 para analizar los patrones de metilación para determinar diferentes tipos de cáncer de acuerdo con la presente divulgación en el presente documento.

La FIG. 97 muestra un análisis de agrupamiento de señales de metilación relacionadas con repeticiones de Alu para diferentes tipos de cáncer. Los sujetos con cáncer de la base de datos TCGA (www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga) tenían un estado de metilación en los sitios CpG analizados utilizando tecnología de microarrays (Infinium HumanMethylation450 BeadChip, Illumina Inc). Se analizaron los estados de metilación en 3,024 sitios CpG presentes en el chip de micromatriz y que se superponen con las regiones Alu a las que se dirigen los complejos CRISPR/Cas9. Hay una serie de CpG que se originan en las regiones Alu de interés en un paciente. El nivel de metilación de cada CpG se cuantificó mediante micromatriz (también llamado índice de metilación o valor beta). Realizamos un análisis de agrupamiento jerárquico basado en una serie de niveles de metilación en esos sitios CpG en todos los pacientes. Por lo tanto, los pacientes con un patrón similar de niveles de metilación en esos sitios CpG se agruparían formando un clado. La similitud de los patrones de metilación entre diferentes pacientes estaría indicada por los valores de altura en el dendrograma de agrupamiento. La altura se calculó de acuerdo con distancias euclidianas en este ejemplo. En otras realizaciones, se utilizarían otras métricas de distancia, que incluyen, pero no se limitan a distancias de Minkowski, Chebychev, Mahalanobismo, Manhattan, Coseno, Correlación, Spearman, Hamming, Jaccard, etc. La altura utilizada en el presente documento representa el valor de la métrica de distancia entre grupos, lo que refleja la relación entre los grupos. Por ejemplo, si uno observaba dos agrupaciones fusionados a una altura x, sugería que la distancia entre esas agrupaciones era x (por ejemplo, la distancia promedio entre todos los pacientes entre grupos).

Con el uso de los estados de metilación en los sitios CpG, los pacientes se agruparon en diferentes grupos distintos dependiendo de los tipos de cáncer en los resultados del análisis de agrupación. Los tipos de cáncer incluyeron carcinoma urotelial de vejiga (BLCA), carcinoma invasivo de mama (BRCA), cistadenocarcinoma seroso de ovario (OV), adenocarcinoma de páncreas (PAAD), HCC, adenocarcinoma de pulmón (LUAD), adenocarcinoma de estómago (STAD), melanoma cutáneo de piel (SKCM) y carcinosarcoma uterino (UCS). El número después del tipo de cáncer en la figura indica un paciente. Por lo tanto, la agrupación sugiere que las señales de metilación en las repeticiones de Alu que seleccionamos fueron informativas para clasificar los tipos de cáncer, que incluyen los tipos de cáncer no mostrados en la FIG. 97. En una realización, se pueden diferenciar los tumores primarios y secundarios en base a los patrones de metilación en una biopsia de tejido.

4. Valores de corte de tamaño y profundidad de sublectura

Esta sección muestra que se pueden utilizar valores de corte de profundidad y/o tamaño de sublectura para mejorar la precisión y/o eficiencia de la detección de metilación de acuerdo con la invención. La preparación de la biblioteca puede modificarse para probar ciertas profundidades o tamaños de sublecturas.

Sobre la base del Kit de Secuenciación Sequel II 2.0, analizamos el efecto de la profundidad de lectura en la cuantificación del nivel de metilación general en los conjuntos de datos de prueba que se generaron a partir de muestras después de la amplificación del genoma completo o el tratamiento con M.SsssI. Estudiamos sitios genómicos que estaban cubiertos por sublecturas con al menos un valor de corte determinado, por ejemplo, pero no se limitan a, ≥ 1x, 10x, 20x, 30x, 40x, 50x, 60x, 70x, 80x, 90x, 100x, etc.

La FIG. 98A muestra el efecto de la profundidad de lectura en la cuantificación general del nivel de metilación en los conjuntos de datos de prueba que estuvieron involucrados con la amplificación del genoma completo. La FIG. 98B muestra el efecto de la profundidad de lectura en la cuantificación general del nivel de metilación en los conjuntos de datos de prueba que estuvieron involucrados con el tratamiento con M. SsssI. El eje y muestra el nivel general de

metilación como un porcentaje. El eje x muestra la profundidad de las sublecturas. Las líneas discontinuas indican los valores esperados de los niveles generales de metilación.

Como se muestra en la FIG. 98A, para el conjunto de datos que involucra la amplificación del genoma completo, la metilación general disminuyó en los pocos valores de corte iniciales, tal como, pero no limitados a, 1x, 10x, 20x, 40x, 50x, que varían desde el 5.7 % hasta el 5.2 %. Los niveles de metilación se estabilizaron progresivamente alrededor del 5 % con un calor de corte de 50x o más.

Por otro lado, en la FIG. 98B, para el conjunto de datos generado a partir de muestras después del tratamiento con M.Ssssl, la metilación general aumentó en los pocos valores de corte iniciales, tales como, pero no limitados a, 1x, 10x, 20x, 40x, 50x, que varían desde el 70 % hasta el 83 %. Los niveles de metilación se estabilizaron progresivamente en alrededor del 83 % con el límite de 50x o más.

En una realización, se podrían ajustar los valores de corte de profundidad de la sublectura, haciendo que la realización del análisis de metilación sea factible en diferentes aplicaciones. En otras realizaciones, se podría utilizar el valor de corte de profundidad de sublectura menos estricto para obtener más ZMW (es decir, número de moléculas) que fueran adecuados para el análisis en dirección descendente. En todavía otra realización, se podría calibrar la lectura de los niveles de metilación determinados por SMRT-seq de acuerdo con la divulgación para una segunda medición, por ejemplo, pero no limitada a BS-seq, PCR de gotitas digitales (en muestras convertidas con bisulfito), PCR específica de metilación, o anticuerpos de unión a citosina metilados u otras proteínas. En otra realización, se obtendría una segunda medición al someter las moléculas de ADN después de la amplificación del genoma completo retenido en 5mC a BS-seq, PCR de gotitas digitales (en muestras convertidas con bisulfito), PCR específica de metilación o secuenciación del genoma enriquecida con proteínas (MBD-seq) del dominio de unión a metil-CpG (MBD). Como un ejemplo, la amplificación del genoma completo retenido en 5mC podría estar mediada por la ADN primasa TthPrimPol, la polimerasa phi29 y la DNMT1 (ADN metiltransferasa 1).

Analizamos los niveles de metilación en varios tipos de cáncer y tejidos no tumorales para diferentes profundidades de sublecturas. Los niveles de metilación determinados por SMRT-seq de acuerdo con la divulgación también se compararon con los resultados de la secuenciación BS-seq. Utilizando el Kit de Secuenciación Sequel II 2.0, obtuvimos una mediana de 43 millones de sublecturas (rango intercuartil (IQR): 30 - 52 millones), lo que permitió generar una mediana de 4.6 millones de secuencias de consenso circulares (CCS) que se alinearon con un genoma humano de referencia (IQR: 2.8 – 5.8 millones). Entre esas muestras, 22 muestras también se sometieron a una secuenciación masiva de bisulfito paralela (BS-seq) bien establecida para determinar los patrones de metilación, lo que proporciona una segunda medición para comparar los niveles de metilación.

La FIG. 99 muestra una comparación entre los niveles de metilación generales determinados por SMRT-seq (Kit de Secuenciación Sequel II 2.0) de acuerdo con la divulgación y BS-seq con el uso de diferentes valores de corte de profundidad de sublectura. El nivel de metilación como porcentaje determinado por SMRT-seq se muestra en el eje y. El nivel de metilación como porcentaje determinado mediante secuenciación con bisulfito está en el eje x. Los símbolos indican diferentes profundidades de sublecturas de 1x, 10x y 30x. Las tres líneas diagonales muestran líneas ajustadas para las diferentes profundidades de sublectura.

La FIG. 99 mostraron que los niveles de metilación en los sitios CpG determinados por SMRT-seq de acuerdo con la divulgación estaban bien correlacionados con ($r = 0.8$; valor de $P < 0.0001$) los determinados por BS-seq, al analizar sitios genómicos que estaban cubiertos por sublecturas al menos una vez. (es decir, valor de corte de profundidad de sublectura ≥ 1x). Estos resultados sugirieron que las realizaciones presentes en esta divulgación se podrían utilizar para medir los niveles de metilación para diferentes tipos de tejido, que incluyen, pero no se limitan a, cáncer colorrectal, tejidos colorrectales, cáncer de esófago, tejidos de esófago, cáncer de mama, tejidos de mama no cancerosos. carcinoma de células renales, tejidos de riñón, cáncer de pulmón y tejidos de pulmón. También observamos que la correlación entre estas dos mediciones mejoró a 0.87 (valor de $P < 0.0001$) y 0.95 (valor de $P < 0.0001$) a medida que los valores de corte de profundidad de la sublectura aumentaron a 10x y 30x, respectivamente. En algunas realizaciones, el aumento de la profundidad de las sublecturas o la selección de regiones genómicas con una cobertura de más sublecturas mejoraría el rendimiento de la determinación de metilación basada en SMRT-seq de acuerdo con la divulgación.

La FIG. 100 es una tabla que muestra el efecto de la profundidad de la sublectura en la correlación de los niveles de metilación entre dos mediciones mediante SMRT-seq (Kit de Secuenciación Sequel II 2.0) y BS-seq. La primera columna muestra el valor de corte de profundidad de la sublectura. La segunda columna muestra la r de Pearson, un coeficiente de correlación. La tercera columna muestra el número de sitios CpG asociados con el valor de corte, con el rango del número de sitios entre paréntesis.

Como se muestra en la FIG. 100, la correlación de los niveles de metilación entre dos mediciones mediante SMRT-seq y BS-seq varió de acuerdo con los diferentes valores de corte de profundidad de la sublectura. En una realización, se podría hacer uso de la relación entre los valores de corte de profundidad de la sublectura y los coeficientes de correlación (por ejemplo, el coeficiente de correlación de Pearson) entre dos mediciones para determinar el valor de corte óptimo de la profundidad de la sublectura para diferenciar citosinas metiladas de citosinas no metiladas. La FIG. 100 mostró que con un valor de corte de profundidad de sublectura de 30x (es decir, ≥ 30x), los niveles de metilación

medidos por SMRT-seq de acuerdo con esta divulgación dieron la correlación más alta con los resultados producidos por BS-seq (r de Pearson = 0.952). En otras realizaciones, se pueden utilizar, pero no se limitan a, valores de corte de profundidad de sublectura de 1x, 10x, 30x, 40x, 50x, 60x, 70x, 80x, 900x, 100x, 200x, 300x, 400x, 500x, 600x, 700x, 800x., etc.

El número de sitios CpG utilizados para el análisis de metilación disminuye con un aumento del valor de corte de profundidad de la sublectura, como se muestra en la FIG. 100. Con un valor de corte de profundidad de sublectura de 100x, se observó una correlación más baja (r de Pearson = 0.875) entre dos mediciones de niveles de metilación, en comparación con un valor de corte de profundidad de sublectura de 30x (r de Pearson = 0.952). La correlación más baja para un valor de corte de sublectura más alto se puede atribuir al menor número de sitios CpG que cumplieron con los valores de corte de profundidad de sublectura más estrictos. En una realización, se puede considerar el equilibrio entre el requisito de profundidad de sublectura y el número de moléculas que se pueden utilizar para el análisis de metilación. Por ejemplo, si el objetivo fuera escanear un genoma completo en busca de patrones de metilación, podrían ser deseables más moléculas. Si uno se centra en una región particular con el uso de SMRT-seq dirigido, puede ser deseable una mayor profundidad de sublectura para obtener patrones de metilación para esa región.

La FIG. 101 muestra la distribución de profundidad de sublectura con respecto a los tamaños de fragmentos en los datos generados por el Kit de Secuenciación Sequel II 2.0. La profundidad de las sublecturas se muestra en el eje y y la longitud de la molécula de ADN se muestra en el eje x. Las longitudes de las moléculas de ADN se dedujeron del tamaño de las secuencias circulares consenso (CCS).

Como la profundidad de la sublectura puede afectar el rendimiento de la determinación de metilación utilizando datos de SMRT-seq y la profundidad de la sublectura es una función de la longitud de una molécula de ADN que se secuencia, los tamaños de las moléculas de ADN pueden ser cruciales para obtener una profundidad de sublectura óptima para analizar patrones de metilación en una muestra. Como se muestra en la FIG. 101, cuanto más largo es el ADN, menor es la profundidad de la sublectura. Por ejemplo, para la población de moléculas con un tamaño de 1 kb, la profundidad mediana de la sublectura fue 50x. Para la población de moléculas con un tamaño de 10 kb, la profundidad mediana de la sublectura fue de 15x.

En una realización, como se muestra en la FIG. 100, el valor de corte óptimo de profundidad de sublectura puede ser al menos 30x, lo que da como resultado el coeficiente de correlación más alto. Para mejorar aún más el rendimiento de las moléculas que cumplirían el valor de corte de profundidad de sublectura óptimo de 30x, se puede hacer uso de la relación entre las profundidades de sublectura y las longitudes de las moléculas plantilla de ADN. Por ejemplo, en la FIG. 101, 30x es la profundidad mediana de sublectura para moléculas que tienen una longitud de aproximadamente 4 kb. Por lo tanto, se pueden fraccionar moléculas de ADN de 4 kb antes de la preparación de la biblioteca SMRT-seq y limitar la secuenciación a las moléculas de ADN de 4 kb. En otras realizaciones, se podrían utilizar otros valores de corte de tamaño para el fraccionamiento de moléculas de ADN, que incluyen, pero no se limitan a, 100 bp, 200 bp, 300 bp, 400 bp, 500 bp, 600 bp, 700 bp, 800 bp, 900 bp, 1 kb, 2 kb, 3 kb, 4 kb, 5 kb, 6 kb, 7 kb, 9 kb, 10 kb, 20 kb, 30 kb, 40 kb, 50 kb, 60 kb, 70 kb, 80 kb, 90 kb, 100 kb, 500 kb, 1 Mb, o diferentes combinaciones de valores de corte de tamaño.

5. Secuenciación en tiempo real de molécula única dirigida basada en enzimas de restricción

Esta sección describe el uso de enzimas de restricción para mejorar la practicabilidad y/o el rendimiento y/o la rentabilidad de la detección de modificaciones. Los fragmentos de ADN generados con enzimas de restricción se pueden utilizar para determinar el origen de una muestra.

a) Utilización de enzimas de restricción para digerir moléculas de ADN.

También describimos cómo se pueden utilizar una o más enzimas de restricción para digerir moléculas de ADN antes de la secuenciación en tiempo real de única molécula (por ejemplo, utilizando el sistema Pacific Biosciences). Debido a que la distribución de los sitios de reconocimiento de las enzimas de restricción estaría presente de manera desigual en un genoma humano, el ADN digerido por las enzimas de restricción puede generar una distribución de tamaño sesgada. Las regiones genómicas con más sitios de reconocimiento de enzimas de restricción se pueden digerir en fragmentos más pequeños, mientras que las regiones genómicas con menos sitios de reconocimiento de enzimas de restricción se pueden digerir en fragmentos más largos. De acuerdo con los rangos de tamaño, se pueden obtener selectivamente las moléculas de ADN que se originan desde una o más regiones que tienen patrones de corte similares de una o más enzimas de restricción. Los rangos de tamaño deseados para la selección de tamaño se pueden determinar mediante análisis de corte in silico para una o más enzimas de restricción. Se puede utilizar un programa informático para determinar el número de sitios de reconocimiento de enzimas de restricción de interés en un genoma de referencia (por ejemplo, un genoma de referencia humano). Dicho genoma de referencia se cortó in silico en fragmentos de acuerdo con esos sitios de reconocimiento, lo que proporcionó información sobre el tamaño de las regiones genómicas de interés.

Con fines ilustrativos, la FIG. 126 muestra un método de secuenciación en tiempo real de única molécula dirigida basada en MspI con el uso de reparación de extremos de ADN y cola A. Como se muestra en la FIG. 126, se puede

utilizar MspI, que reconoce sitios 5'C^CGG3', para digerir una muestra de ADN de un organismo, por ejemplo, pero no se limitan a, una muestra de ADN humano. Los fragmentos de ADN digeridos con salientes 5'CG se sometieron a selección de tamaño, enriqueciendo las moléculas de ADN originadas en las islas CpG. Las regiones genómicas que están enriquecidas con residuos G y C (también llamados contenido de GC) pueden generar fragmentos más cortos. Por lo tanto, se puede determinar el rango de tamaños de fragmentos para realizar la selección en función del contenido de GC de las regiones de interés. Un experto en la técnica dispone de una variedad de herramientas de selección del tamaño de fragmentos de ADN que incluyen, pero no se limitan a, electroforesis en gel, electroforesis por exclusión de tamaño, electroforesis capilar, cromatografía, espectrometría de masas, enfoques de filtración, enfoques basados en precipitación, microfluidos y nanofluidos. Las moléculas de ADN fraccionadas por tamaño se sometieron a reparación del extremo del ADN y a una cola en A de tal manera que el producto de ADN deseado se pudiera ligar con adaptadores de horquilla que llevaban un saliente en T 5', formando plantillas de ADN circulares.

Después de la eliminación de los adaptadores no ligados, el ADN lineal y el ADN circular incompleto, por ejemplo, pero no se limita a, utilizando exonucleasas (por ejemplo, exonucleasa III y VII), las moléculas de ADN ligadas con adaptadores en horquilla se pueden utilizar para secuenciación en tiempo real de única molécula para determinar la IPD, PW y el contexto de secuencia para determinar los perfiles de metilación como se divulga en el presente documento. Al analizar las regiones genómicas enriquecidas con CpG, el ADN obtenido de diferentes tejidos o tejidos con diferentes enfermedades y/o afecciones fisiológicas o muestras biológicas se puede distinguir y clasificar por su perfil de metilación determinado por los métodos de análisis de datos de secuenciación de esta divulgación.

Para la etapa que implica la selección de tamaño en la FIG. 126, los rangos de tamaño deseados se pueden determinar mediante el análisis de corte in silico de MspI. Determinamos un total de 2,286,541 sitios de corte de MspI en una referencia humana. Se cortó in silico un genoma de referencia humano en fragmentos de acuerdo con los sitios de corte de MspI. Obtuvimos un total de 2,286,565 fragmentos. El tamaño de cada fragmento individual se determinó por el número total de nucleótidos de ese fragmento.

Con fines ilustrativos, las FIG. 127A y 127B muestran la distribución de tamaño de los fragmentos digeridos con MspI. El eje y de estas cifras es la frecuencia en porcentaje para un tamaño particular de fragmento. La FIG. 127A tiene una escala logarítmica para el eje x que varía desde 50 hasta 500,000 bp. La FIG. 127B tiene una escala lineal para el eje x que varía desde 50 hasta 1,000 bp.

Como se muestra en las FIG. 127A y 127B, las moléculas de ADN digeridas con MspI tienen una distribución de tamaño sesgada. El tamaño de mediana de los fragmentos digeridos con MspI fue de 404 bp (IQR: 98 - 1411 bp). Aproximadamente el 53 % de los fragmentos digeridos con MspI tenían menos de 1 kb. Hubo una serie de picos en el perfil de tamaño que podrían ser causados por elementos repetidos. Ciertos elementos repetidos pueden compartir patrones similares de sitios de corte de MspI, lo que lleva a un conjunto de moléculas derivadas de la digestión de MspI que poseían tamaños de fragmentos similares. Por ejemplo, el pico puntiagudo con la frecuencia más alta (es decir, un total de 49,079) correspondió a un tamaño de 64 bp. Entre ellos, 45,894 (94 %) estaban superpuestos con repeticiones de Alu. Se pueden seleccionar moléculas de ADN con un tamaño de 64 bp para enriquecer las moléculas de ADN que se originan a partir de repeticiones de Alu. Los datos sugieren que se puede utilizar la selección de tamaño para enriquecer las moléculas de ADN deseadas para el análisis de metilación en dirección descendente de acuerdo con la divulgación.

Con fines ilustrativos, la FIG. 128 muestra una tabla con el número de moléculas de ADN para ciertos rangos de tamaño seleccionados. La primera columna muestra rangos de tamaño en pares de bases. La segunda columna muestra el porcentaje de moléculas dentro de un rango de tamaño en relación con el total de fragmentos. La tercera columna muestra el número de moléculas dentro del rango de tamaño que se superponen a las islas CpG. La cuarta columna muestra el porcentaje de moléculas dentro de un rango de tamaño que se superponen a las islas CpG. La quinta columna muestra el número de sitios CpG que se secuencian. La sexta columna muestra el número de sitios CpG que caen dentro de las islas CpG. La séptima columna muestra el porcentaje de sitios CpG dirigidos por selección de tamaño y que caen dentro de islas CpG. Como se muestra en la FIG. 128, la cantidad de moléculas de ADN generadas a partir de un genoma humano sometido a digestión con MspI varió de acuerdo con los diferentes rangos de tamaño en cuestión. El número de moléculas de ADN que se superponían a las islas CpG varió con diferentes rangos de tamaño.

Como el motivo CCGG se produjo preferentemente en islas CpG, la selección de moléculas con un tamaño menor a un cierto valor de corte puede permitir enriquecer las moléculas de ADN que se originan en islas CpG. Por ejemplo, para un rango de tamaño de 50 a 200 bp, el número de moléculas fue 526,543, lo que representó el 23.03 % del total de fragmentos de ADN derivados de un genoma humano sometido a digestión con MspI. Entre 526,543 moléculas de ADN, 104,079 (19.76 %) se superpusieron con islas CpG. Para un rango de tamaño de 600 a 800 bp, el número de moléculas fue 133,927, lo que representó el 5.86 % del total de fragmentos de ADN derivados de un genoma humano sometido a digestión con MspI. Entre 133,927 moléculas, 3,673 (2.74 %) moléculas se superpusieron con islas CpG. Como un ejemplo, se puede seleccionar un tamaño de 50 a 200 bp para enriquecer fragmentos de ADN que se originan en islas CpG.

Para calcular el grado de enriquecimiento de los sitios CpG que se superponen a las islas CpG a través de una secuenciación en tiempo real de única molécula dirigida basada en MspI, realizamos una simulación para ADN cortado

por sonicación, simulamos 526,543 fragmentos generados a partir de ZMW con un tamaño medio de 200 bp y una desviación estándar de 20 bp sobre la base de una distribución normal. Sólo había un 0.88 % de moléculas de ADN superpuestas a las islas CpG. Un total de 71,495 sitios CpG se superpusieron con islas CpG. Como se muestra en la FIG. 128, la selección de fragmentos digeridos con MspI que varían desde 50 hasta 200 bp daría como resultado un 19.8 % de fragmentos que se superponen a las islas CpG. Por lo tanto, estos datos sugirieron que el ADN preparado mediante digestión con MspI puede tener un enriquecimiento de 22.5 veces en fragmentos de ADN que se originan de islas CpG, en comparación con el ADN preparado mediante sonicación. Además, analizamos los sitios CpG que se enriquecen en islas CpG a través de la digestión con MspI. La selección de fragmentos digeridos con MspI que varían desde 50 hasta 200 bp puede dar lugar a 885,041 sitios CpG que se superponen a islas CpG, lo que representa el 37.5 % del total de sitios CpG de fragmentos secuenciados dentro de ese rango de tamaño. Hubo un enriquecimiento de 12.3 veces (es decir, 885,041/71,495) de los sitios CpG que se superponen a las islas CpG, en comparación con el del ADN preparado mediante sonicación. En base a la información mostrada en la FIG. 128, se puede seleccionar un rango de tamaño adecuado para incluir el número deseable de sitios CpG y el enriquecimiento de veces deseable de los sitios CpG dentro de las islas CpG.

Con fines ilustrativos, la FIG. 129 es un gráfico del porcentaje de cobertura de sitios CpG dentro de islas CpG versus el tamaño de los fragmentos de ADN después de la digestión con enzimas de restricción. El eje y muestra el porcentaje de sitios CpG dentro de islas CpG cubiertos por fragmentos que tienen los tamaños dados. El eje x muestra el límite superior del rango de tamaño de los fragmentos de ADN después de la digestión con enzimas de restricción. La FIG. 129 mostró el porcentaje de sitios CpG dentro de islas CpG que se cubrirán al ampliar el rango de selección de tamaño. En la Fig. 129, el rango de tamaño es desde 50 bp hasta el tamaño mostrado en el eje x. En otras realizaciones, el límite inferior del rango de tamaño se puede personalizar, por ejemplo, pero no se limita a, 60 bp, 70 bp, 80 bp, 90 bp, 100 bp, 200 bp, 300 bp, 400 bp y 500 bp. Con la ampliación del rango de tamaño al aumentar el límite de tamaño superior, podemos observar que el porcentaje de cobertura de sitios CpG dentro de las islas CpG aumenta gradualmente y se estabiliza en 65 %. Algunos de los sitios CpG no están cubiertos porque están dentro de fragmentos de ADN de menos de 50 bp o están dentro de fragmentos dentro de moléculas extremadamente largas (por ejemplo, >100,000 bp).

En algunas realizaciones, se puede analizar una muestra de ADN utilizando dos o más enzimas de restricción diferentes (con diferentes sitios de restricción) para aumentar la cobertura de sitios CpG dentro de islas CpG. La digestión de la muestra de ADN por diferentes enzimas se puede llevar a cabo en reacciones individuales de tal manera que solo haya una enzima de restricción en cada reacción. Por ejemplo, AccII, que reconoce sitios CG^CG, se puede utilizar para cortar preferentemente en islas CpG. En otras realizaciones, se pueden utilizar otras enzimas de restricción con dinucleótidos CG como parte del sitio de reconocimiento. Dentro del genoma humano, había 678,669 sitios de corte AccII. Realizamos un corte in silico del genoma humano de referencia mediante restricción AccII y obtuvimos un total de 678,693 fragmentos. Luego realizamos una selección de tamaño in silico de estos fragmentos y calculamos el porcentaje de cobertura de los sitios CpG dentro de las islas CpG de acuerdo con el método descrito anteriormente para la digestión con MspI. Podemos observar un aumento gradual en el porcentaje de cobertura de sitios CpG con la ampliación del rango de selección de tamaño. El porcentaje de cobertura se estabiliza alrededor del 50 %. La cobertura de los sitios CpG aumenta aún más al combinar datos de los dos experimentos de digestión enzimática, a saber, la digestión con MspI y la digestión con AccII. El 80 % de los sitios CpG dentro de las islas CpG se cubren a través de la selección de fragmentos de ADN con un tamaño de 50 bp a 400 bp. Este porcentaje es mayor que los números respectivos de los experimentos de digestión con cualquiera de las dos enzimas solas. La cobertura se puede aumentar aún más mediante el análisis de la muestra de ADN utilizando otras enzimas de restricción. Si una muestra de ADN se divide en dos alícuotas. Una alícuota se digiere con MspI y la otra con AccII. Las dos muestras de ADN digeridas se mezclan en moles iguales y se secuencian utilizando una secuenciación en tiempo real de única molécula con 5 millones de ZMW. En base al análisis in silico, el 83 % de los sitios CpG dentro de las islas CpG (es decir, 1,734,345) se secuenciarían al menos 4 veces en términos de secuencias de consenso circulares.

La FIG. 130 muestra una secuenciación en tiempo real de única molécula dirigida basada en MspI sin el uso de reparación de extremos de ADN y cola A. La ligación entre las moléculas de ADN digeridas y los adaptadores de horquilla se puede realizar sin el proceso de reparación del extremo del ADN y cola A. Se pueden ligar directamente las moléculas de ADN digeridas que llevan salientes 5' CG con adaptadores de horquilla que llevan salientes 5' CG, formando la plantilla de ADN circular para la secuenciación en tiempo real de única molécula. Después de la limpieza de los adaptadores no ligados y los dímeros adaptadores autoligados, y en algunas realizaciones después de la eliminación de los adaptadores no ligados, el ADN lineal y el ADN circular incompleto, las moléculas de ADN ligadas con adaptadores en horquilla pueden ser adecuadas para una secuenciación en tiempo real de única molécula para obtener la IPD, PW y el contexto de secuencia. El perfil de metilación de una única molécula se determinaría utilizando IPD, PW y contexto de secuencia de acuerdo con la divulgación.

La FIG. 131 muestra una secuenciación en tiempo real de única molécula dirigida basada en MspI con una probabilidad reducida de autoligación del adaptador. La base de citosina subyacente indica una base sin grupos fosfato 5'. Para minimizar la posibilidad de formación de dímeros adaptadores autoligados que pueden ocurrir durante el proceso de ligación del adaptador, se pueden utilizar adaptadores de horquilla desfosforilados para realizar la ligación del adaptador con aquellas moléculas de ADN digeridas con MspI. Es posible que esos adaptadores en horquilla desfosforilados no formen dímeros adaptadores autoligados debido a la falta de grupos fosfato 5'. Después de la ligación, el producto se sometió a la etapa de limpieza del adaptador para purificar las moléculas de ADN ligadas con

adaptadores de horquilla. Las moléculas de ADN ligadas con adaptadores de horquilla que pueden llevar las mellas se sometieron además a fosforilación (por ejemplo, polinucleótido quinasa T4) y sellado de mellas mediante ADN ligasa (por ejemplo, ADN ligasa T4). Además, se pueden realizar la eliminación de los adaptadores no ligados, el ADN lineal y el ADN circular incompleto. Las moléculas de ADN ligadas con adaptadores de horquilla eran adecuadas para la secuenciación en tiempo real de única molécula para obtener la IPD, PW y el contexto de secuencia. El perfil de metilación de una única molécula se determinaría utilizando IPD, PW y contexto de secuencia de acuerdo con la divulgación.

Además de MspI, también se pueden utilizar otras enzimas de restricción, tales como SmaI, con un sitio de reconocimiento CCCGGG.

El proceso de selección del tamaño deseado se puede realizar después de la etapa de reparación final del ADN. El proceso de selección del tamaño deseado se puede realizar después de la ligación de los adaptadores de horquilla, cuando se determinó el efecto de los adaptadores de horquilla en el resultado de la selección del tamaño. Los órdenes de las etapas del procedimiento que involucran la secuenciación en tiempo real de única molécula dirigida basada en MspI pueden cambiar dependiendo de las situaciones experimentales.

La selección del tamaño se llevaría a cabo utilizando métodos basados en electroforesis en gel y/o basados en perlas magnéticas. En realizaciones, las enzimas de restricción pueden incluir, pero no se limitan a, BglII, EcoRI, EcoRII, BamHI, HindIII, TaqI, NotI, HinFI, PvuII, Sau3AI, SmaI, HaeIII, HgaI, HpaII, AluI, EcoRV, EcoP15I, KpnI, PstI, SacI, SalI, ScaI, SpeI, SphI, StuI, XbaI y combinaciones de las mismas.

b) Distinción de tipos de muestras biológicas con metilación.

Esta sección describe el uso de perfiles de metilación determinados utilizando fragmentos generados por digestión con enzimas de restricción para facilitar la distinción entre diferentes muestras biológicas.

Evaluamos las diferencias en los perfiles de metilación entre muestras biológicas utilizando perfiles de metilación determinados por secuenciación en tiempo real de única molécula basada en MspI de acuerdo con las realizaciones en esta divulgación. Tomamos como ejemplo muestras de ADN de tejido placentario y de ADN de la capa leucocitaria. Realizamos una simulación por ordenador para generar los datos relacionados con la muestra de ADN de la placenta y la capa leucocitaria sobre la base de una secuenciación en tiempo real de única molécula dirigida basada en MspI. La simulación se basó en los valores cinéticos que incluyen IPD y PW para cada nucleótido generado previamente mediante la secuenciación SMRT del ADN del tejido placentario y el ADN de la capa leucocitaria hasta la cobertura del genoma completo utilizando el Kit de Secuenciación Sequel II 1.0. Luego simulamos la condición mediante la cual las muestras de ADN placentario y de ADN de la capa leucocitaria se sometieron a digestión con MspI, seguida de una selección de tamaño basada en gel utilizando un rango de tamaño de 50 a 200 bp. Las moléculas de ADN seleccionadas se ligaron con adaptadores de horquilla para formar plantillas de ADN circulares. Las plantillas circulares de ADN se sometieron a secuenciación en tiempo real de única molécula para obtener la información con respecto a IPD, PW y contexto de secuencia.

Suponiendo que había 500,000 ZMW que generaban sublecturas de secuenciación SMRT, esas sublecturas siguieron las distribuciones genómicas de los fragmentos digeridos con MspI dentro de un rango de tamaño de 50 a 200 bp como se muestra en la Tabla 1. Se supuso que la profundidad de la sublectura era 30x para ambas Muestras de ADN de placenta y capa leucocitaria. Repetimos la simulación 10 veces para la muestra de ADN de placenta y la muestra de ADN de capa leucocitaria, respectivamente. Por lo tanto, el conjunto de datos generado in silico por una secuenciación en tiempo real de única molécula dirigida digerida con MspI comprendió un total de 10 muestras de ADN de placenta y se obtuvieron 10 muestras de ADN de capa leucocitaria. CNN analizó más a fondo el conjunto de datos y determinó los perfiles de metilación para cada muestra de acuerdo con la divulgación. Obtuvimos una mediana de 9,198 sitios CpG de islas CpG (rango: 5,497 – 13,928), que representaron el 13.6 % del total de sitios CpG secuenciados (rango: 45,304 – 90,762). El estado de metilato para cada sitio CpG en cada molécula se determinó mediante un modelo CNN de acuerdo con la divulgación.

Con fines ilustrativos, la FIG. 132 es un gráfico de los niveles generales de metilación entre muestras de ADN de placenta y leucocitos determinados mediante secuenciación en tiempo real de única molécula dirigida basada en MspI. El eje y es el nivel de metilación como porcentaje. El tipo de muestras enumeradas en el eje x. La FIG. 132 muestra que los niveles generales de metilación (mediana: 57.6 %; rango: 56.9 % - 59.1 %) fueron más bajos en las muestras de placenta en comparación con las muestras de capa leucocítica (mediana: 69.5 %; rango: 68.9 % - 70.4 %) (valor de $P < 0.0001$, prueba U de Mann-Whitney). Estos resultados sugirieron que los perfiles de metilación determinados por la secuenciación en tiempo real de única molécula basada en MspI se pueden utilizar para diferenciar muestras de tejido o muestras biológicas en base a sus diferencias de metilación. Debido a que estos datos muestran que el ADN de la placenta se puede distinguir del ADN de la capa leucocitaria debido a sus diferencias de metilación detectadas mediante secuenciación en tiempo real de única molécula basada en MspI, se puede aplicar este método para medir la fracción de ADN fetal en el plasma materno. La fracción de ADN fetal se puede medir utilizando metilación porque el ADN fetal en el plasma o suero materno proviene de la placenta, mientras que las moléculas de ADN restantes en la muestra se derivan en su mayoría de células de la capa leucocitaria materna. En realizaciones, esta

tecnología sería una herramienta útil para diferenciar diferentes tejidos o tejidos con diferentes enfermedades y/o condiciones fisiológicas o muestras biológicas.

Para realizar análisis de agrupamiento entre muestras de ADN de placenta y muestra de ADN de capa leucocítica utilizando perfiles de metilación de islas CpG, calculamos los niveles de metilación de ADN de una isla CpG utilizando la proporción de sitios CpG clasificados como metilación entre los sitios CpG totales de esa isla CpG. Utilizamos los niveles de metilación de las regiones de isla CpG para realizar el análisis de agrupamiento con fines ilustrativos.

Con fines ilustrativos, la FIG. 133 muestra un análisis de agrupamiento de muestras de placenta y capa leucocitaria utilizando sus perfiles de metilación del ADN determinados mediante secuenciación en tiempo real de única molécula dirigida basada en MspI. La similitud de los patrones de metilación de las islas CpG en diferentes pacientes se indica mediante los valores de altura en el dendrograma de agrupamiento. La altura se calcula de acuerdo con distancias euclidianas en este ejemplo. En una realización, se puede utilizar el valor de corte de altura 100 para cortar el árbol de agrupamiento en dos grupos, lo que permite diferenciar muestras de placenta y capa leucocitaria con 100 % de sensibilidad y especificidad. En otras realizaciones, se pueden utilizar otros valores de corte de altura que incluyen, pero no se limitan a, 50, 60, 70, 80, 90, 120, 130, 140 y 150, etc. La FIG. 133 mostraron que 10 muestras de ADN de placenta y 10 muestras de ADN de capa leucocitaria se agruparon claramente por separado en dos grupos utilizando los perfiles de metilación de las islas CpG determinados por una secuenciación en tiempo real de única molécula basada en MspI de acuerdo con la divulgación.

V. Métodos de entrenamiento y detección

Esta sección muestra métodos de ejemplo para entrenar un modelo de aprendizaje automático para la detección de una metilación de acuerdo con la invención y utilizar el modelo de aprendizaje automático para detectar una metilación de acuerdo con la invención.

A. Entrenamiento modelo

La FIG. 102 muestra un método de ejemplo 1020 para detectar una metilación de un nucleótido en una molécula de ácido nucleico de acuerdo con la presente invención. El método de ejemplo 1020 puede ser un método para entrenar un modelo para detectar la metilación. La metilación puede incluir cualquier metilación descrita en el presente documento. La metilación puede tener estados discretos, como metilado y no metilado, y potencialmente especificar un tipo de metilación. Por tanto, puede haber más de dos estados (clasificaciones) de un nucleótido.

En el bloque 1022, se recibe una pluralidad de primeras estructuras de datos. En el presente documento se describen varios ejemplos de estructuras de datos, por ejemplo, en las FIG. 4-16. Cada primera estructura de datos de la primera pluralidad de primeras estructuras de datos corresponde a una ventana respectiva de nucleótidos secuenciados en una molécula de ácido nucleico respectiva de una pluralidad de primeras moléculas de ácido nucleico. Cada ventana asociada con la primera pluralidad de estructuras de datos puede incluir 4 o más nucleótidos consecutivos, que incluyen 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21 o más nucleótidos consecutivos. Cada ventana puede tener el mismo número de nucleótidos consecutivos. Es posible que las ventanas se superpongan. Cada ventana puede incluir nucleótidos en una primera hebra de la primera molécula de ácido nucleico y nucleótidos en una segunda hebra de la primera molécula de ácido nucleico. La primera estructura de datos también puede incluir para cada nucleótido dentro de la ventana un valor de una propiedad de la hebra. La propiedad de la hebra puede indicar que el nucleótido está presente o ya seas la primera o la segunda hebra. La ventana puede incluir nucleótidos en la segunda hebra que no son complementarios a un nucleótido en una posición correspondiente en la primera hebra. En algunas realizaciones, todos los nucleótidos de la segunda hebra son complementarios a los nucleótidos de la primera hebra. En algunas realizaciones, cada ventana puede incluir nucleótidos en solo una hebra de la primera molécula de ácido nucleico.

La primera molécula de ácido nucleico puede ser una molécula de ADN circular. La molécula de ADN circular se puede formar al cortar una molécula de ADN de hebra doble utilizando un complejo Cas9 para formar una molécula de ADN de hebra doble cortada. Se puede ligar un adaptador de horquilla sobre un extremo de la molécula de ADN de hebra doble cortada. En realizaciones, se pueden cortar y ligar ambos extremos de una molécula de ADN de hebra doble. Por ejemplo, el corte, ligación y análisis posterior se pueden realizar como se describe en la FIG. 91.

La primera pluralidad de primeras estructuras de datos puede incluir de 5,000 a 10,000, 10,000 a 50,000, 50,000 a 100,000, 100,000 a 200,000, 200,000 a 500,000, 500,000 a 1,000,000, o 1,000,000 o más primeras estructuras de datos. La pluralidad de primeras moléculas de ácido nucleico puede incluir al menos 1,000, 10,000, 50,000, 100,000, 500,000, 1,000,000, 5,000,000, o más moléculas de ácido nucleico. Como un ejemplo adicional, se pueden generar al menos 10,000 o 50,000 o 100,000 o 500,000 o 1,000,000 o 5,000,000 lecturas de secuencia.

Cada una de las primeras moléculas de ácido nucleico se secuencia al medir pulsos en una señal correspondiente a los nucleótidos. La señal puede ser una señal de fluorescencia u otro tipo de señal óptica (por ejemplo, quimioluminiscencia, fotométrica). La señal puede resultar de los nucleótidos o etiquetas asociadas con los nucleótidos.

La metilación tiene un primer estado conocido en el nucleótido en una posición diana en cada ventana de cada primera molécula de ácido nucleico. El primer estado puede ser que la metilación esté ausente en el nucleótido o puede ser que la modificación esté presente en el nucleótido. Se puede saber que la metilación está ausente en las primeras moléculas de ácido nucleico, o las primeras moléculas de ácido nucleico se pueden someter a un tratamiento tal que la metilación esté ausente. Se puede saber que la metilación se presenta en las primeras moléculas de ácido nucleico, o las primeras moléculas de ácido nucleico se pueden someter a un tratamiento de manera que la metilación esté presente. Si el primer estado es que la metilación está ausente, la metilación puede estar ausente en cada ventana de cada primera molécula de ácido nucleico y no estar ausente sólo en la posición diana. Los primeros estados conocidos pueden incluir un estado metilado para una primera porción de las primeras estructuras de datos y un estado no metilado para una segunda porción de las primeras estructuras de datos.

La posición diana puede ser el centro de la ventana respectiva. Para una ventana que abarca un número par de nucleótidos, la posición diana puede ser la posición inmediatamente en dirección ascendente o inmediatamente en dirección descendente del centro de la ventana. En algunas realizaciones, la posición diana puede estar en cualquier otra posición de la ventana respectiva, que incluye la primera posición o la última posición. Por ejemplo, si la ventana abarca n nucleótidos de una hebra, desde la 1ª posición hasta la n^{ésima} posición (ya sea en dirección ascendente o en dirección descendente), la posición diana puede estar en cualquier posición desde la 1ª hasta la n^{ésima} posición.

Cada primera estructura de datos incluye valores para propiedades dentro de la ventana. Las propiedades son para cada nucleótido dentro de la ventana. Las propiedades incluyen una identidad del nucleótido. La identidad puede incluir la base (por ejemplo, A, T, C o G). Las propiedades también incluyen una posición del nucleótido con respecto a la posición diana dentro de la ventana respectiva. Por ejemplo, la posición puede ser una distancia de nucleótidos relativa a la posición diana. La posición puede ser +1 cuando el nucleótido está a un nucleótido de la posición diana en una dirección, y la posición puede ser -1 cuando el nucleótido está a un nucleótido de la posición diana en la dirección opuesta.

Las propiedades incluyen una anchura del pulso correspondiente al nucleótido. La anchura del pulso puede ser la anchura del pulso a la mitad del valor máximo del pulso. Las propiedades incluyen además una duración de interpulso (IPD) que representa un tiempo entre el pulso correspondiente al nucleótido y un pulso correspondiente a un nucleótido vecino. La duración de interpulso puede ser el tiempo entre el valor máximo del pulso asociado con el nucleótido y el valor máximo del pulso asociado con el nucleótido vecino. El nucleótido vecino puede ser el nucleótido adyacente. Las propiedades también pueden incluir una altura del pulso correspondiente a cada nucleótido dentro de la ventana. Las propiedades pueden incluir además un valor de una propiedad de la hebra, que indica si el nucleótido está presente en la primera o en la segunda hebra de la primera molécula de ácido nucleico. La indicación de la hebra puede ser similar a la matriz mostrada en la FIG. 6.

Cada estructura de datos de la pluralidad de primeras estructuras de datos puede excluir primeras moléculas de ácido nucleico con un IPD o ancho por debajo de un valor de corte. Por ejemplo, sólo se pueden utilizar primeras moléculas de ácido nucleico con un valor de IPD mayor que un percentil 10 (o un percentil 1, 5, 15, 20, 30, 40, 50, 60, 70, 80, 90 o 95). El percentil se puede basar en datos de todas las moléculas de ácido nucleico en una muestra de referencia o muestras de referencia. El valor de corte de la anchura también puede corresponder a un percentil.

En el bloque 1024, se almacena una pluralidad de primeras muestras de entrenamiento. Cada primera muestra de entrenamiento incluye una de la primera pluralidad de primeras estructuras de datos y una primera etiqueta que indica el primer estado para la metilación del nucleótido en la posición diana.

En el bloque 1026, se recibe una segunda pluralidad de segundas estructuras de datos. El bloque 1026 puede ser opcional. Cada segunda estructura de datos de la segunda pluralidad de segundas estructuras de datos corresponde a una ventana respectiva de nucleótidos secuenciados en una molécula de ácido nucleico respectiva de una pluralidad de segundas moléculas de ácido nucleico. La segunda pluralidad de moléculas de ácido nucleico puede ser igual o diferente que la pluralidad de primeras moléculas de ácido nucleico. La metilación tiene un segundo estado conocido en un nucleótido en una posición diana dentro de cada ventana de cada segunda molécula de ácido nucleico. El segundo estado es un estado diferente al primero. Por ejemplo, si el primer estado es que la metilación está presente, entonces el segundo estado es que la metilación está ausente, y viceversa. Cada segunda estructura de datos incluye valores para las mismas propiedades que la primera pluralidad de primeras estructuras de datos.

La pluralidad de primeras muestras de entrenamiento se puede generar utilizando amplificación de desplazamiento múltiple (MDA). En algunas realizaciones, la pluralidad de primeras muestras de entrenamiento se puede generar al amplificar una primera pluralidad de moléculas de ácido nucleico utilizando un conjunto de nucleótidos. El conjunto de nucleótidos puede incluir un primer tipo de metilación (por ejemplo, 6mA o cualquier otra metilación [por ejemplo, CpG]) en una relación especificada. La relación especificada puede incluir 1:10, 1:100, 1:1000, 1:10000, 1:100000 o 1:1000000 en relación con los nucleótidos no metilados. La pluralidad de segundas moléculas de ácido nucleico se puede generar utilizando amplificación por desplazamiento múltiple con nucleótidos no metilados del primer tipo.

En el bloque 1028, se almacena una pluralidad de segundas muestras de entrenamiento. El bloque 1028 puede ser opcional. Cada segunda muestra de entrenamiento incluye una de la segunda pluralidad de segundas estructuras de datos y una segunda etiqueta que indica el segundo estado para la metilación del nucleótido en la posición diana.

En el bloque 1029, se entrena un modelo utilizando la pluralidad de primeras muestras de entrenamiento y opcionalmente la pluralidad de segundas muestras de entrenamiento. El entrenamiento se realiza al optimizar los parámetros del modelo en base a las salidas del modelo que coinciden o no con las etiquetas correspondientes de las primeras etiquetas y opcionalmente de las segundas etiquetas cuando se introducen la primera pluralidad de primeras estructuras de datos y opcionalmente la segunda pluralidad de segundas estructuras de datos al modelo. Una salida del modelo especifica si el nucleótido en la posición diana en la ventana respectiva tiene la metilación. El método puede incluir sólo la pluralidad de primeras muestras de entrenamiento porque el modelo puede identificar un valor atípico como de un estado diferente al primer estado. El modelo puede ser un modelo estadístico, también denominado modelo de aprendizaje automático.

En algunas realizaciones, la salida del modelo puede incluir una probabilidad de estar en cada uno de una pluralidad de estados. El estado con mayor probabilidad se puede tomar como estado.

El modelo puede incluir una red neuronal convolucional (CNN). La CNN puede incluir un conjunto de filtros convolucionales configurados para filtrar la primera pluralidad de estructuras de datos y, opcionalmente, la segunda pluralidad de estructuras de datos. El filtro puede ser cualquier filtro descrito en el presente documento. El número de filtros para cada capa puede ser desde 10 a 20, 20 a 30, 30 a 40, 40 a 50, 50 a 60, 60 a 70, 70 a 80, 80 a 90, 90 a 100, 100 a 150, 150 a 200, o más. El tamaño del núcleo para los filtros puede ser 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, desde 15 hasta 20, desde 20 hasta 30, desde 30 hasta 40, o más. La CNN puede incluir una capa de entrada configurada para recibir la primera pluralidad de estructuras de datos filtradas y, opcionalmente, la segunda pluralidad de estructuras de datos filtradas. La CNN también puede incluir una pluralidad de capas ocultas que incluyen una pluralidad de nodos. La primera capa de la pluralidad de capas ocultas acopladas a la capa de entrada. La CNN puede incluir además una capa de salida acoplada a una última capa de la pluralidad de capas ocultas y configurada para emitir una estructura de datos de salida. La estructura de datos de salida puede incluir las propiedades.

El modelo puede incluir un modelo de aprendizaje supervisado. Los modelos de aprendizaje supervisado pueden incluir diferentes enfoques y algoritmos, que incluyen el aprendizaje analítico, redes neuronales artificiales, retropropagación, reforzamiento (metaalgoritmo), estadística bayesiana, razonamiento basado en casos, aprendizaje de árboles de decisión, programación lógica inductiva, regresión de procesos gaussianos, programación genética, método de grupo de manejo de datos, estimadores de núcleo, autómatas de aprendizaje, sistemas clasificadores de aprendizaje, longitud mínima de mensaje (árboles de decisión, gráficos de decisión, etc.), aprendizaje subespacial multilineal, clasificador simple de Bayes, clasificador de máxima entropía, campo aleatorio condicional, Algoritmo del Vecino más Cercano, aprendizaje probablemente aproximadamente correcto (PAC), reglas de propagación, una metodología de adquisición de conocimientos, algoritmos de aprendizaje automático simbólico, algoritmos de aprendizaje automático subsimbólico, máquinas de vectores de soporte, Máquinas de Complejidad Mínima (MCM), bosques aleatorios, grupos de clasificadores, clasificación ordinal, preprocesamiento de datos, manejo de conjuntos de datos desequilibrados, aprendizaje relacional estadístico o Proaftn, un algoritmo de clasificación multicriterio. El modelo puede ser regresión lineal, regresión logística, red neuronal recurrente profunda (por ejemplo, memoria a largo plazo, LSTM), clasificador de Bayes, modelo oculto de Markov (HMM), análisis discriminante lineal (LDA), agrupamiento de k-medias, agrupamiento espacial de aplicaciones con ruido basado en densidad (DBSCAN), algoritmo de bosque aleatorio, máquina de vectores de soporte (SVM) o cualquier modelo descrito en el presente documento.

Como parte del entrenamiento de un modelo de aprendizaje automático, los parámetros del modelo de aprendizaje automático (tales como pesos, umbrales, por ejemplo, como se pueden utilizar para funciones de activación en redes neuronales, etc.) se pueden optimizar en base a las muestras de entrenamiento (conjunto de entrenamiento) para proporcionar una precisión optimizada en la clasificación de la modificación del nucleótido en la posición diana. Se pueden realizar varias formas de optimización, por ejemplo, retropropagación, minimización del riesgo empírico y minimización del riesgo estructural. Se puede utilizar un conjunto de validación de muestras (estructura de datos y etiqueta) para validar la precisión del modelo. La validación cruzada se puede realizar utilizando varias porciones del conjunto de entrenamiento para entrenamiento y validación. El modelo puede comprender una pluralidad de submodelos, proporcionando de esta manera un modelo de grupo. Los submodelos pueden ser modelos más débiles que, una vez combinados, proporcionan un modelo final más preciso.

En algunas realizaciones, se pueden utilizar moléculas de ácido nucleico quiméricas o híbridas para validar el modelo. Al menos algunas de la pluralidad de primeras moléculas de ácido nucleico incluyen cada una una primera porción correspondiente a una primera secuencia de referencia y una segunda porción correspondiente a una segunda secuencia de referencia. La primera secuencia de referencia puede ser de un cromosoma, tejido (por ejemplo, tumoral o no tumoral), organismo o especie diferente que la segunda secuencia de referencia. La primera secuencia de referencia puede ser humana y la segunda secuencia de referencia puede ser de un animal diferente. Cada molécula de ácido nucleico quimérico puede incluir la primera porción correspondiente a la primera secuencia de referencia y la segunda porción correspondiente a la segunda secuencia de referencia. La primera porción puede tener un primer patrón de metilación y la segunda porción puede tener un segundo patrón de metilación. La primera porción se puede tratar con una metilasa. La segunda porción no se puede tratar con la metilasa y puede corresponder a una porción no metilada de la segunda secuencia de referencia.

B. Detección de modificaciones

La FIG. 103 muestra un método 1030 para detectar una metilación de un nucleótido en una molécula de ácido nucleico de acuerdo con la presente invención. La metilación puede ser cualquier metilación descrita con el método 1020 de la FIG. 102.

En el bloque 1032, se recibe una estructura de datos de entrada. La estructura de datos de entrada corresponde a una ventana de nucleótidos secuenciados en una molécula de ácido nucleico de muestra. La molécula de ácido nucleico de muestra se secuencia al medir los pulsos en una señal óptica correspondiente a los nucleótidos. La ventana puede ser cualquier ventana descrita con el bloque 1022 en la FIG. 102, y la secuenciación puede ser cualquier secuenciación descrita con el bloque 1022 en la FIG. 102. La estructura de datos de entrada puede incluir valores para las mismas propiedades descritas con el bloque 1022 en la FIG. 102. El método 1030 puede incluir secuenciar la molécula de ácido nucleico de muestra.

Los nucleótidos dentro de la ventana pueden estar alineados o no con un genoma de referencia. Los nucleótidos dentro de la ventana se pueden determinar utilizando una secuencia consenso circular (CCS) sin alineación de los nucleótidos secuenciados con un genoma de referencia. El CCS puede identificar los nucleótidos en cada ventana en lugar de alinearlos con un genoma de referencia. En algunas realizaciones, la ventana se puede determinar sin una CCS y sin alineamiento de los nucleótidos secuenciados con un genoma de referencia.

Los nucleótidos dentro de la ventana se pueden enriquecer o filtrar. El enriquecimiento se puede realizar mediante un enfoque que involucre Cas9. El enfoque de Cas9 puede incluir cortar una molécula de ADN de hebra doble utilizando un complejo de Cas9 para formar una molécula de ADN de hebra doble cortada y ligar un adaptador de horquilla en un extremo de la molécula de ADN de hebra doble cortada, similar a la FIG. 91. La filtración se puede realizar al seleccionar moléculas de ADN de hebra doble que tienen un tamaño dentro de un rango de tamaños. Los nucleótidos pueden provenir de estas moléculas de ADN de hebra doble. Se pueden utilizar otros métodos que preserven el estado de metilación de las moléculas (por ejemplo, proteínas de unión a metilo).

En el bloque 1034, la estructura de datos de entrada se introduce en un modelo. El modelo se puede entrenar mediante el método 1020 en la FIG. 102.

En algunas realizaciones, se pueden utilizar moléculas de ácido nucleico quiméricas para validar el modelo. Al menos algunas de la pluralidad de primeras moléculas de ácido nucleico incluyen cada una una primera porción correspondiente a una primera secuencia de referencia y una segunda porción correspondiente a una segunda secuencia de referencia que está separada de la primera secuencia de referencia. La primera secuencia de referencia puede ser de un cromosoma, tejido (por ejemplo, tumoral o no tumoral), orgánulos (por ejemplo, mitocondrias, núcleos, cloroplastos), organismo (mamíferos, virus, bacterias, etc.) o especies diferentes a la segunda secuencia de referencia. La primera secuencia de referencia puede ser humana y la segunda secuencia de referencia puede ser de un animal diferente. Cada molécula de ácido nucleico quimérico puede incluir la primera porción correspondiente a la primera secuencia de referencia y la segunda porción correspondiente a la segunda secuencia de referencia. La primera porción puede tener un primer patrón de metilación y la segunda porción puede tener un segundo patrón de metilación. La primera porción puede tratarse con una metilasa. La segunda porción puede no estar tratada con la metilasa y puede corresponder a una porción no metilada de la segunda secuencia de referencia.

En el bloque 1036, se determina utilizando el modelo si la metilación está presente en un nucleótido en la posición diana dentro de la ventana en la estructura de datos de entrada.

La estructura de datos de entrada puede ser una estructura de datos de entrada de una pluralidad de estructuras de datos de entrada. Cada estructura de datos de entrada corresponde a una ventana respectiva de nucleótidos secuenciados en una molécula de ácido nucleico de muestra respectiva de la pluralidad de moléculas de ácido nucleico de muestra. La pluralidad de moléculas de ácido nucleico de muestra se puede obtener a partir de una muestra biológica de un sujeto. La muestra biológica puede ser cualquier muestra biológica descrita en el presente documento. El método 1030 se puede repetir para cada estructura de datos de entrada. El método puede incluir recibir la pluralidad de estructuras de datos de entrada. La pluralidad de estructuras de datos de entrada se puede introducir en el modelo. Utilizando el modelo se puede determinar si hay una metilación presente en un nucleótido en la ubicación objetivo en la ventana respectiva de cada estructura de datos de entrada.

Cada molécula de ácido nucleico de muestra de la pluralidad de moléculas de ácido nucleico de muestra puede tener un tamaño mayor que un tamaño de valor de corte. Por ejemplo, el tamaño de valor de corte puede ser 100 bp, 200 bp, 300 bp, 400 bp, 500 bp, 600 bp, 700 bp, 800 bp, 900 bp, 1 kb, 2 kb, 3 kb, 4 kb, 5 kb, 6 kb, 7 kb, 9 kb, 10 kb, 20 kb, 30 kb, 40 kb, 50 kb, 60 kb, 70 kb, 80 kb, 90 kb, 100 kb, 500 kb o 1 Mb. Tener un valor de corte de tamaño puede dar como resultado una profundidad de sublectura mayor, cualquiera de las cuales puede aumentar la precisión de la detección de metilación. En algunas realizaciones, el método puede incluir fraccionar las moléculas de ADN para ciertos tamaños antes de secuenciar las moléculas de ADN.

La pluralidad de moléculas de ácido nucleico de muestra se pueden alinear con una pluralidad de regiones genómicas. Para cada región genómica de la pluralidad de regiones genómicas, se pueden alinear una serie de moléculas de ácido nucleico de muestra con la región genómica. El número de moléculas de ácido nucleico de muestra puede ser mayor que un número de corte. El número de corte puede ser un valor de corte de profundidad de sublectura. El

número de valor de corte de profundidad de sublectura puede ser 1x, 10x, 30x, 40x, 50x, 60x, 70x, 80x, 900x, 100x, 200x, 300x, 400x, 500x, 600x, 700x u 800x. El número de valor de corte de profundidad de sublectura se puede determinar para mejorar u optimizar la precisión. El número de valor de corte de profundidad de la sublectura puede estar relacionado con el número de la pluralidad de regiones genómicas. Por ejemplo, un número de valor de corte de profundidad de sublectura mayor, un número menor de la pluralidad de regiones genómicas.

Se puede determinar que la metilación está presente en uno o más nucleótidos. Se puede determinar una clasificación de un trastorno utilizando la presencia de metilación en uno o más nucleótidos. La clasificación del trastorno puede incluir utilizar el número de modificaciones. El número de metilaciones se puede comparar con un umbral. Alternativa o adicionalmente, la clasificación puede incluir la ubicación de una o más metilaciones. La ubicación de una o más metilaciones se puede determinar al alinear lecturas de secuencia de una molécula de ácido nucleico con un genoma de referencia. El trastorno se puede determinar si se demuestra que ciertas ubicaciones que se sabe que están correlacionadas con el trastorno tienen metilación. Por ejemplo, se puede comparar un patrón de sitios metilados con un patrón de referencia para un trastorno, y la determinación del trastorno se puede basar en la comparación. Una coincidencia con el patrón de referencia o una coincidencia sustancial (por ejemplo, 80 %, 90 % o 95 % o más) con el patrón de referencia puede indicar el trastorno o una alta probabilidad del trastorno. El trastorno puede ser cáncer o cualquier trastorno (por ejemplo, trastorno asociado al embarazo, enfermedad autoinmunitaria) descrito en el presente documento.

Se puede analizar un número estadísticamente significativo de moléculas de ácido nucleico para proporcionar una determinación precisa de un trastorno, origen de tejido o fracción de ADN clínicamente relevante. En algunas realizaciones, se analizan al menos 1,000 moléculas de ácido nucleico. En otras realizaciones, se pueden analizar al menos 10,000 o 50,000 o 100,000 o 500,000 o 1,000,000 o 5,000,000 moléculas de ácido nucleico, o más. Como ejemplo adicional, se pueden generar al menos 10,000 o 50,000 o 100,000 o 500,000 o 1,000,000 o 5,000,000 de lecturas de secuencia.

El método puede incluir determinar que la clasificación del trastorno es que el sujeto tiene el trastorno. La clasificación puede incluir un nivel del trastorno, utilizando el número de metilaciones y/o los sitios de las metilaciones.

Se puede determinar una fracción de ADN clínicamente relevante, un perfil de metilación fetal, un perfil de metilación materna, la presencia de una región genética de impronta o un tejido de origen (por ejemplo, de una muestra que contiene una mezcla de diferentes tipos de células) utilizando la presencia de la modificación en uno o más nucleótidos. La fracción de ADN clínicamente relevante incluye, pero no se limita a, fracción de ADN fetal, fracción de ADN tumoral (por ejemplo, de una muestra que contiene una mezcla de células tumorales y células no tumorales) y fracción de ADN de trasplante (por ejemplo, de una muestra que contiene una mezcla de células del donante y células del receptor).

De acuerdo con la divulgación y no de acuerdo con la invención, el método puede incluir además tratar el trastorno. El tratamiento puede proporcionarse de acuerdo con un nivel determinado del trastorno, las modificaciones identificadas y/o el tejido de origen (por ejemplo, de células tumorales aisladas de la circulación de un paciente con cáncer). Por ejemplo, una modificación identificada se puede dirigir con un fármaco o quimioterapia en particular. El tejido de origen se puede utilizar para guiar una cirugía o cualquier otra forma de tratamiento. Y el nivel de trastorno se puede utilizar para determinar qué tan agresivo debe ser con cualquier tipo de tratamiento.

El tratamiento puede incluir tratar el trastorno en el paciente después de determinar el nivel del trastorno en el paciente. El tratamiento puede incluir cualquier terapia, fármaco, quimioterapia, radiación o cirugía adecuada, que incluye cualquier tratamiento descrito en una referencia mencionada en el presente documento.

VI. Análisis de haplotipos

Se encontraron diferencias en los perfiles de metilación entre dos haplotipos en muestras de tejido tumoral. Por lo tanto, los desequilibrios de metilación entre haplotipos se pueden utilizar para determinar una clasificación de un nivel de cáncer u otro trastorno. Los desequilibrios en los haplotipos también se pueden utilizar para identificar la herencia de un haplotipo por parte de un feto. Los trastornos fetales también se pueden identificar mediante el análisis de los desequilibrios de metilación entre haplotipos. El ADN celular se puede utilizar para analizar los niveles de metilación de haplotipos. El análisis de haplotipos no está de acuerdo con la invención y está presente solo con fines ilustrativos.

A. Análisis de metilación asociado a haplotipos

La tecnología de secuenciación en tiempo real de única molécula permite la identificación de SNP individuales. Las lecturas largas producidas a partir de pocillos de secuenciación en tiempo real de única molécula (por ejemplo, hasta varias kilobases) permitirían eliminar variantes en los genomas al aprovechar la información de haplotipos presente en cada lectura de consenso (Edge et al. Genome Res. 2017;27: 801-812; Wenger et al. Nat Biotechnol. 2019;37:1155-1162). El perfil de metilación del haplotipo se podría analizar a partir de los niveles de metilación de los sitios CpG ligados por el CCS a los alelos en los haplotipos respectivos, como se ilustra en la FIG. 77. Este análisis de haplotipos de metilación por fases se podría utilizar para resolver la cuestión de si dos copias de cromosomas homólogos comparten patrones de metilación similares o diferentes en diferentes afecciones clínicamente relevantes, tales como el cáncer. También describimos cómo la metilación del haplotipo serían los niveles de metilación agregados aportados por una serie de fragmentos de ADN asignados a ese haplotipo. El haplotipo podría ser bloques de diferentes tamaños,

que incluyen, pero no se limitan a, 50 nt, 100 nt, 200 nt, 300 nt, 400 nt, 500 nt, 1 knt, 2 knt, 3 knt, 4 knt, 5 knt, 10 knt, 20 knt, 30 knt, 40 knt, 50 knt, 100 knt, 200 knt, 300 knt, 400 knt, 500 knt, 1 Mnt, 2 Mnt y 3 Mnt.

B. Análisis relativo del desequilibrio de metilación basado en haplotipos

Con fines ilustrativos, la FIG. 104 ilustra el análisis de desequilibrio de metilación relativo basado en haplotipos. Los haplotipos (es decir, Hap I y Hap II) se determinaron al analizar los resultados de la secuenciación en tiempo real de única molécula. Los patrones de metilación ligados a cada haplotipo se podrían determinar utilizando aquellos fragmentos asociados a haplotipo cuyos perfiles de metilación se determinaron de acuerdo con el enfoque descrito en la FIG. 77. De este modo, se podrían comparar los patrones de metilación entre Hap I y Hap II.

Para cuantificar la diferencia en la metilación entre Hap I y Hap II, se calculó la diferencia de niveles de metilación (ΔF) entre Hap I y Hap II. La diferencia ΔF se calcula como:

$$\Delta F = M_{HapI} - M_{HapII}$$

donde ΔF representa la diferencia en el nivel de metilación entre Hap I y Hap II, y M_{HapI} y M_{HapII} representan los niveles de metilación de Hap I y Hap II, respectivamente. Un valor positivo de ΔF sugirió un mayor nivel de metilación del ADN para Hap I en comparación con Hap II.

C. Análisis de desequilibrio de metilación relativo basado en haplotipos para el ADN del tumor HCC

El análisis de metilación de haplotipos puede ser útil para detectar aberraciones de metilación en genomas de cáncer. Por ejemplo, se analizaría el cambio de metilación entre dos haplotipos dentro de una región genómica. Un haplotipo dentro de una región genómica se define como un bloque de haplotipo. Un bloque de haplotipo podría considerarse como un conjunto de alelos en un cromosoma que han estado en fase. Un bloque de haplotipo se extendería tanto como fuera posible de acuerdo con un conjunto de información de secuencia que respalda dos alelos ligados físicamente en un cromosoma. Para el caso 3033, obtuvimos 97,475 bloques de haplotipos a partir de los resultados de la secuenciación del ADN del tejido normal adyacente. El tamaño de la mediana de los bloques de haplotipos fue de 2.8 kb. El 25 % de los bloques de haplotipos tenían un tamaño mayor al 8.2 kb. El tamaño máximo de los bloques de haplotipos fue de 282.2 kb. El conjunto de datos se generó a partir de ADN preparado con el Kit de Secuenciación Sequel II 1.0.

Con fines ilustrativos, utilizamos una serie de criterios para identificar los bloques de haplotipos potenciales que exhibían la metilación diferencial entre Hap I y Hap II en el ADN tumoral en comparación con el ADN del tejido no tumoral adyacente. Los criterios fueron: (1) el bloque de haplotipo que se estaba analizando contenía al menos 3 tres secuencias CCS que se produjeron a partir de tres pocillos de secuenciación, respectivamente; (2) la diferencia absoluta en el nivel de metilación entre Hap I y Hap II en el ADN del tejido no tumoral adyacente fue menor del 5 %; (3) la diferencia absoluta en el nivel de metilación entre Hap I y Hap II en el ADN del tejido tumoral fue mayor al 30 %. Identificamos 73 bloques de haplotipos que cumplían los criterios anteriores.

Con fines ilustrativos, la FIG. 105A y 105B son tablas de los 73 bloques de haplotipos que muestran niveles de metilación diferenciales entre Hap I y Hap II en el ADN del tumor HCC en comparación con el ADN del tejido no tumoral adyacente para el caso TBR3033. La primera columna muestra el cromosoma asociado con el bloque de haplotipo. La segunda columna muestra la coordenada inicial del bloque de haplotipo dentro del cromosoma. La tercera columna muestra la coordenada final del bloque de haplotipo. La cuarta columna muestra la longitud del bloque de haplotipo. La cuarta columna enumera la identificación del bloque de haplotipo. La quinta columna muestra el nivel de metilación de Hap I en tejido no tumoral adyacente al tejido tumoral. La sexta columna muestra el nivel de metilación de Hap II en el tejido no tumoral. La séptima columna muestra el nivel de metilación de Hap I en tejido tumoral. La octava columna muestra el nivel de metilación de Hap II en tejido tumoral.

En contraste con los 73 bloques de haplotipo que muestran una diferencia mayor al 30 % en el nivel de metilación entre haplotipos para el ADN del tejido tumoral, sólo un bloque de haplotipo mostró una diferencia mayor al 30 % para el ADN del tejido no tumoral pero menor del 5 % de diferencia en el ADN del tejido tumoral. Se podría utilizar otro conjunto de criterios para identificar los bloques de haplotipos que exhiben metilaciones diferenciales. Se pueden utilizar otras diferencias de umbral máximo y mínimo. Por ejemplo, las diferencias de umbral mínimo pueden ser del 10 %, 15 %, 20 %, 25 %, 30 %, 35 %, 40 %, 45 %, 50 % o más. Las diferencias de umbral máximas pueden ser del 1 %, 5 %, 10 %, 15 %, 20 % o 30 %, como ejemplos. Estos resultados sugirieron que la variación de la diferencia de metilación entre haplotipos puede servir como un nuevo biomarcador para el diagnóstico, detección, monitorización, pronóstico y orientación del tratamiento del cáncer.

Un bloque de haplotipo largo se sometería a partición, in silico, en bloques más pequeños al estudiar los patrones de metilación.

Para el caso 3032, obtuvimos 61,958 bloques de haplotipos a partir de los resultados de la secuenciación del ADN del tejido no tumoral adyacente. El tamaño de mediana de los bloques de haplotipos fue de 9.3 kb. El 25 % de los bloques de haplotipos tenían un tamaño mayor a 27.6 kb. El tamaño máximo de los bloques de haplotipos fue de 717.8 kb. A

modo de ilustración, utilizamos los mismos tres criterios descritos anteriormente para identificar los posibles bloques de haplotipos que exhibían la metilación diferencial entre Hap I y Hap II en el ADN del tumor en comparación con el ADN del tejido normal adyacente. Identificamos 20 bloques de haplotipos que cumplían los criterios anteriores. El conjunto de datos se generó a partir de ADN preparado con el Kit de Secuenciación Sequel II 1.0.

Con fines ilustrativos, la FIG. 106 es una tabla de los 20 bloques de haplotipos que muestran niveles de metilación diferenciales entre Hap I y Hap II en el ADN del tumor en comparación con el ADN del tejido normal adyacente para el caso TBR3032. La primera columna muestra el cromosoma asociado con el bloque de haplotipo. La segunda columna muestra la coordenada inicial del bloque de haplotipo dentro del cromosoma. La tercera columna muestra la coordenada final del bloque de haplotipo. La cuarta columna muestra la longitud del bloque de haplotipo. La cuarta columna enumera la identificación del bloque de haplotipo. La quinta columna muestra el nivel de metilación de Hap I en tejido no tumoral adyacente al tejido tumoral. La sexta columna muestra el nivel de metilación de Hap II en el tejido no tumoral. La séptima columna muestra el nivel de metilación de Hap I en tejido tumoral. La octava columna muestra el nivel de metilación de Hap II en tejido tumoral.

En contraste con los 20 bloques de haplotipos que muestran la diferencia en el tejido tumoral de HCC en la FIG. 106, sólo un bloque de haplotipo mostró una diferencia mayor al 30 % en tejido no tumoral pero menor del 5 % en tejido tumoral. Estos resultados sugieren además que la variación de la diferencia de metilación entre haplotipos serviría como un nuevo biomarcador para el diagnóstico, detección, monitorización, pronóstico y orientación del tratamiento del cáncer. Se podrían utilizar otros criterios para identificar los bloques de haplotipos que exhiben metilaciones diferenciales.

D. Análisis del desequilibrio de metilación relativo basado en haplotipos para el ADN de otros tipos de tumores

Como se indicó anteriormente, el análisis de los niveles de metilación entre haplotipos reveló que los tejidos tumorales de HCC albergaban más bloques de haplotipos que exhibían un desequilibrio de metilación en comparación con tejidos no tumorales adyacentes emparejados. Como un ejemplo, los criterios para un bloque de haplotipo que muestra un desequilibrio de metilación en un tejido tumoral fueron: (1) el bloque de haplotipo que se estaba analizando contenía al menos tres secuencias CCS que se produjeron a partir de tres pocillos de secuenciación; (2) la diferencia absoluta en el nivel de metilación entre Hap I y Hap II en el ADN del tejido no tumoral adyacente o el ADN del tejido normal en base a datos históricos fue menor del 5 %; (3) la diferencia absoluta en el nivel de metilación entre Hap I y Hap II en el ADN del tejido tumoral fue mayor al 30 %. El criterio (2) se incluyó porque los tejidos no tumorales/normales que muestran un desequilibrio de haplotipos en los niveles de metilación pueden indicar regiones impresas en lugar de regiones tumorales. Los criterios para un bloque de haplotipo que muestra un desequilibrio de metilación en un tejido no tumoral fueron: (1) el bloque de haplotipo que se estaba analizando contenía al menos tres secuencias CCS que se produjeron a partir de tres pocillos de secuenciación; (2) la diferencia absoluta en el nivel de metilación entre Hap I y Hap II en el ADN del tejido no tumoral adyacente o el ADN del tejido normal en base a los datos históricos fue mayor al 30 %; (3) la diferencia absoluta en el nivel de metilación entre Hap I y Hap II en el ADN del tejido tumoral fue menor del 5 %.

En otras realizaciones, se pueden utilizar otros criterios. Por ejemplo, para identificar el desequilibrio del genoma del cáncer del haplotipo I, la diferencia en el nivel de metilación entre Hap I y Hap II puede ser menor al 1 %, 5 %, 10 %, 20 %, 40 %, 50 % o 60 %, etc., en tejidos no tumorales, mientras que la diferencia en el nivel de metilación entre Hap I y Hap II puede ser mayor al 1 %, 5 %, 10 %, 20 %, 40 %, 50 % o 60 %, etc., en tejidos tumorales. Para identificar el desequilibrio del genoma no canceroso del haplotipo I, la diferencia en el nivel de metilación entre Hap I y Hap II puede ser mayor al 1 %, 5 %, 10 %, 20 %, 40 %, 50 % o 60 %, etc., en tejidos no tumorales, mientras que la diferencia en el nivel de metilación entre Hap I y Hap II puede ser menor del 1 %, 5 %, 10 %, 20 %, 40 %, 50 % o 60 %, etc., en tejidos tumorales.

Con fines ilustrativos, la FIG. 107A es una tabla que resume el número de bloques de haplotipos que muestran un desequilibrio de metilación entre dos haplotipos entre tejidos tumorales y no tumorales adyacentes en base a los datos generados por el Kit de Secuenciación Sequel II 2.0. La primera columna enumera el tipo de tejido. La segunda columna enumera el número de bloques de haplotipos que muestran un desequilibrio de metilación entre dos haplotipos en tejidos tumorales. La tercera columna enumera el número de bloques de haplotipos que muestran un desequilibrio de metilación entre dos haplotipos en tejidos no tumorales adyacentes emparejados. Las filas muestran tejido tumoral con más bloques de haplotipos que muestran un desequilibrio de metilación entre dos haplotipos que el tejido no tumoral adyacente emparejado.

La longitud mediana de los bloques de haplotipos implicados en este análisis fue de 15.7 kb (IQR: 10.3 – 26.1 kb). Incluyendo los resultados de HCC para el hígado, los datos muestran 7 tipos de tejido para los cuales el tejido tumoral albergaba más bloques de haplotipos con desequilibrio de metilación. Además del hígado, los otros tejidos incluyen el colon, mama, riñón, pulmón, próstata y estómago. Por tanto, en algunas realizaciones, se podría utilizar el número de bloques de haplotipos que albergan un desequilibrio de metilación para detectar si un paciente tenía un tumor o cáncer.

Con fines ilustrativos, la FIG. 107B es una tabla que resume el número de bloques de haplotipos que muestran un desequilibrio de metilación entre dos haplotipos en tejidos tumorales para diferentes estadios tumorales basándose en los datos generados por el Kit de Secuenciación Sequel II 2.0. La primera columna muestra el tipo de tejido con

tumor. La segunda columna muestra el número de bloques de haplotipos con un desequilibrio de metilación entre dos haplotipos en tejidos tumorales. La tercera columna enumera la información sobre la estadificación del tumor utilizando la clasificación TNM de tumores malignos. T3 y T3a son tumores de mayor tamaño que T2.

La tabla muestra más bloques de haplotipos que muestran un desequilibrio de metilación para tumores más grandes tanto de mama como de riñón. Por ejemplo, para el tejido mamario, el tejido categorizado como tumor de grado T3 (estadificación TNM), ER positivo y que exhibe amplificación de ERBB2 tenía más bloques de haplotipos (57) que mostraban un desequilibrio de metilación que los bloques de haplotipos (18) para el tejido categorizado como tumor de grado T2 (Estadificación TNM), PR (receptor de progesterona)/ER (receptor de estrógeno) positivo y sin amplificación de ERBB2. Para el tejido renal, el tejido categorizado como tumor de grado T3a tenía más bloques de haplotipos (68) que mostraban un desequilibrio de metilación que los bloques de haplotipos (0) para el tejido categorizado como tumor de grado T2.

Se pueden hacer uso de bloques de haplotipos que muestran un desequilibrio de metilación para la clasificación de tumores y para correlacionarlos con su comportamiento clínico (por ejemplo, progresión, pronóstico o respuesta al tratamiento). Estos datos sugirieron que el grado de desequilibrio de metilación basado en haplotipos puede servir como clasificador de tumores y se puede incorporar en estudios o ensayos clínicos o eventuales servicios clínicos. La clasificación de los tumores puede incluir tamaño y gravedad.

E. Análisis de metilación basado en haplotipos del ADN libre de células plasmáticas maternas

Se pueden determinar los haplotipos de ambos progenitores o de cualquiera de los progenitores. Los métodos de haplotipado pueden incluir secuenciación de única molécula de lectura larga, secuenciación ligada de lectura corta (por ejemplo, genómica 10x), PCR de una única molécula de largo alcance o inferencia poblacional. Si se conocen los haplotipos paternos, el metiloma del ADN fetal libre de células se puede ensamblar al ligar los perfiles de metilación de múltiples moléculas de ADN libre de células, cada una de las cuales contiene al menos un alelo SNP paterno específico que está presente a lo largo del haplotipo paterno. En otras palabras, el haplotipo paterno se utiliza como andamio para ligar las secuencias de lectura específicas del feto.

Con fines ilustrativos, la FIG. 108 ilustra el análisis de haplotipos para el desequilibrio relativo de metilación. Si se conocen los haplotipos maternos, el desequilibrio de metilación entre los dos haplotipos (es decir, Hap I y Hap II) se puede utilizar para determinar el haplotipo materno heredado fetalmente. Como se muestra en la FIG. 108, las moléculas de ADN plasmático de una mujer embarazada se secuencian utilizando tecnología de secuenciación en tiempo real de única molécula. La metilación y la información alélica se pueden determinar de acuerdo con la divulgación en el presente documento. Los SNP ligados a un gen que causa la enfermedad se asignan como Hap I. Si el feto ha heredado Hap I, en el plasma materno estarían presentes más fragmentos que llevan los alelos de Hap I en comparación con los que llevan los alelos de Hap II. La hipometilación de fragmentos de ADN derivados del feto reduciría el nivel de metilación de Hap I en comparación con el de Hap II. Como resultado, si la metilación de Hap I muestra un nivel de metilación más bajo que el de Hap II, es más probable que el feto herede el Hap I materno. De lo contrario, es más probable que el feto herede el Hap II materno. En la práctica clínica, el análisis del desequilibrio de metilación basado en haplotipos se puede utilizar para determinar si un feto no nacido ha heredado un haplotipo materno asociado con trastornos genéticos, por ejemplo, pero no limitados a, trastornos de un solo gen, que incluyen el síndrome de X frágil, distrofia muscular, enfermedad de Huntington o beta-talasemia.

F. Método de clasificación de trastornos de ejemplo

Con fines ilustrativos, la FIG. 109 muestra un método de ejemplo 1090 para clasificar un trastorno en un organismo que tiene un primer haplotipo y un segundo haplotipo. El método 1090 implica comparar los niveles relativos de metilación entre dos haplotipos.

En el bloque 1091, se analizan moléculas de ADN de la muestra biológica para identificar sus ubicaciones en un genoma de referencia correspondiente al organismo. Las moléculas de ADN pueden ser moléculas de ADN celular. Por ejemplo, las moléculas de ADN se pueden secuenciar para obtener lecturas de secuencia, y las lecturas de secuencia se pueden mapear (alinear) con el genoma de referencia. Si el organismo fuera un humano, entonces el genoma de referencia sería un genoma humano de referencia, potencialmente de una subpoblación particular. Como otro ejemplo, las moléculas de ADN se pueden analizar con diferentes sondas (por ejemplo, después de PCR u otros métodos de amplificación), donde cada sonda corresponde a una ubicación genómica, que puede cubrir un heterocigoto y uno o más sitios CpG, como se describe a continuación.

Además, las moléculas de ADN se pueden analizar para determinar un alelo respectivo de la molécula de ADN. Por ejemplo, un alelo de una molécula de ADN se puede determinar a partir de una secuencia leída obtenida a partir de secuenciación o de una sonda particular que se hibrida con la molécula de ADN, donde ambas técnicas pueden proporcionar una secuencia leída (por ejemplo, la sonda se puede tratar como la secuencia leída cuando hay hibridación). Se puede determinar un estado de metilación en cada uno de uno o más sitios (por ejemplo, sitios CpG) para las moléculas de ADN.

En el bloque 1092, se identifican uno o más loci heterocigotos de una primera porción de la primera región cromosómica. Cada locus heterocigoto puede incluir un primer alelo correspondiente en el primer haplotipo y un

segundo alelo correspondiente en el segundo haplotipo. El uno o más loci heterocigotos puede ser una primera pluralidad de loci heterocigotos, donde una segunda pluralidad de loci heterocigotos puede corresponder a una región cromosómica diferente.

En el bloque 1093, se identifica un primer conjunto de la pluralidad de moléculas de ADN. Cada una de la pluralidad de moléculas de ADN está ubicada en cualquiera de los loci heterocigotos del bloque 1096 e incluye un primer alelo correspondiente, de tal manera que la molécula de ADN se puede identificar como correspondiente al primer haplotipo. Es posible que una molécula de ADN esté ubicada en más de uno de los loci heterocigotos, pero normalmente una lectura solo incluiría un locus heterocigoto. Cada uno del primer conjunto de moléculas de ADN también incluye al menos uno de los N sitios genómicos, donde los sitios genómicos se utilizan para medir los niveles de metilación. N es un número entero, por ejemplo, mayor o igual a 1, 2, 3, 4, 5, 10, 20, 50, 100, 200, 500, 1,000, 2,000 o 5,000. Por tanto, una lectura de una molécula de ADN puede indicar la cobertura de 1 sitio, 2 sitios, etc. El 1 sitio genómico puede incluir un sitio en el que está presente un nucleótido CpG.

En el bloque 1094, se determina un primer nivel de metilación de la primera porción del primer haplotipo utilizando el primer conjunto de la pluralidad de moléculas de ADN. El primer nivel de metilación se puede determinar mediante cualquier método descrito en el presente documento. La primera porción puede corresponder a un solo sitio o incluir muchos sitios. La primera porción del primer haplotipo puede tener una longitud mayor o igual a 1 kb. Por ejemplo, la primera porción del primer haplotipo puede tener una longitud mayor o igual a 1 kb, 5 kb, 10 kb, 15 kb o 20 kb. Los datos de metilación pueden ser datos del ADN celular.

Se pueden determinar una pluralidad de primeros niveles de metilación para una pluralidad de porciones del primer haplotipo. Cada porción puede tener una longitud mayor o igual a 5 kb o cualquier tamaño descrito en el presente documento para la primera porción del primer haplotipo.

En el bloque 1095, se identifica un segundo conjunto de la pluralidad de moléculas de ADN. Cada una de la pluralidad de moléculas de ADN está ubicada en cualquiera de los loci heterocigotos del bloque 1096 e incluye un segundo alelo correspondiente, de tal manera que la molécula de ADN se puede identificar como correspondiente al segundo haplotipo. Cada uno del segundo conjunto de moléculas de ADN también incluye al menos uno de los N sitios genómicos, donde los sitios genómicos se utilizan para medir los niveles de metilación.

En el bloque 1096, se determina un segundo nivel de metilación de la primera porción de un segundo haplotipo utilizando el segundo conjunto de la pluralidad de moléculas de ADN. El segundo nivel de metilación se puede determinar mediante cualquier método descrito en el presente documento. La primera porción del segundo haplotipo puede tener una longitud mayor o igual a 1 kb o cualquier tamaño para la primera porción del primer haplotipo. La primera porción del primer haplotipo puede ser complementaria a la primera porción del segundo haplotipo. La primera porción del primer haplotipo y la primera porción del segundo haplotipo pueden formar una molécula de ADN circular. El primer nivel de metilación de la primera porción del primer haplotipo se puede determinar utilizando datos de la molécula de ADN circular. Por ejemplo, el análisis del ADN circular puede incluir el análisis descrito en la FIG. 1, FIG. 2, FIG. 4, FIG. 5, FIG. 6, FIG. 7, FIG. 8, FIG. 50, o FIG. 61.

La molécula de ADN circular se puede formar cortando una molécula de ADN de hebra doble utilizando un complejo Cas9 para formar una molécula de ADN de hebra doble cortada. Se puede ligar un adaptador de horquilla sobre un extremo de la molécula de ADN de hebra doble cortada. En las realizaciones, se pueden cortar y ligar ambos extremos de una molécula de ADN de hebra doble. Por ejemplo, el corte, la ligación y el análisis posterior se pueden realizar como se describe en la FIG. 91.

Se pueden determinar una pluralidad de segundos niveles de metilación para una pluralidad de porciones del segundo haplotipo. Cada porción de la pluralidad de porciones del segundo haplotipo puede ser complementaria a una porción de la pluralidad de porciones del primer haplotipo.

En el bloque 1097, se calcula un valor de un parámetro utilizando el primer nivel de metilación y el segundo nivel de metilación. El parámetro puede ser un valor de separación. El valor de separación puede ser una diferencia entre los dos niveles de metilación o una relación de los dos niveles de metilación.

Si se utiliza una pluralidad de porciones del segundo haplotipo, entonces para cada porción de la pluralidad de porciones del segundo haplotipo, se puede calcular un valor de separación utilizando el segundo nivel de metilación de la porción del segundo haplotipo y el primer nivel de metilación utilizando la porción complementaria del primer haplotipo. El valor de separación se puede comparar con un valor de corte.

El valor de corte se puede determinar a partir de tejidos que no tienen el trastorno. El parámetro puede ser el número de porciones del segundo haplotipo donde el valor de separación excede el valor de corte. Por ejemplo, el número de porciones del segundo haplotipo donde el valor de separación excede el valor de corte puede ser similar al número de regiones que se muestran con una diferencia superior al 30 % en la FIG. 105A, FIG. 105B, y la FIG. 106. Con la figura. 105A, FIG. 105B, y la FIG. 106, el valor de separación es una relación y el valor de corte es 30 %. En algunas realizaciones, el valor límite se puede determinar a partir de tejidos que tienen el trastorno.

En otro ejemplo, el valor de separación para cada porción se puede agregar, por ejemplo, sumar, lo que se puede hacer mediante una suma ponderada o una suma de funciones de los valores de separación respectivos. Dicha agregación puede proporcionar el valor del parámetro.

En el bloque 1098, el valor del parámetro se compara con un valor de referencia. El valor de referencia se puede determinar utilizando un tejido de referencia sin el trastorno. El valor de referencia puede ser un valor de separación. Por ejemplo, el valor de referencia puede representar que no debería haber una diferencia significativa entre los niveles de metilación de los dos haplotipos. Por ejemplo, el valor de referencia puede ser una diferencia estadística de 0 o una relación de aproximadamente 1. Cuando se utiliza una pluralidad de porciones, el valor de referencia puede ser un número de porciones en un organismo sano donde los dos haplotipos muestran un valor de separación que excede el valor de corte. En algunas realizaciones, el valor de referencia se puede determinar utilizando un tejido de referencia con el trastorno.

En el bloque 1099, la clasificación del trastorno en el organismo se determina utilizando la comparación del valor del parámetro con el valor de referencia. Se puede determinar que el trastorno está presente o es más probable si el valor del parámetro excede el valor de referencia. El trastorno puede incluir cáncer. El cáncer puede ser cualquier cáncer descrito en el presente documento. La clasificación del trastorno puede ser una probabilidad del trastorno. La clasificación del trastorno puede incluir la gravedad del trastorno. Por ejemplo, un valor de parámetro mayor que indique una mayor cantidad de porciones con un desequilibrio de haplotipo puede indicar una forma más grave de cáncer.

Si bien el método descrito con la FIG. 109 implica una clasificación de un trastorno, se pueden utilizar métodos similares para determinar cualquier condición o característica que pueda resultar de un desequilibrio en los niveles de metilación entre haplotipos. Por ejemplo, el nivel de metilación de un haplotipo del ADN fetal puede ser menor que el de un haplotipo del ADN materno. Los niveles de metilación se pueden utilizar para clasificar los ácidos nucleicos como maternos o fetales.

Cuando el trastorno es cáncer, diferentes regiones cromosómicas de un tumor pueden exhibir tales diferencias en la metilación. Dependiendo de las regiones afectadas, se puede proporcionar un tratamiento diferente. Además, los sujetos que tienen diferentes regiones que exhiben dichas diferencias en la metilación pueden tener pronósticos diferentes.

Las regiones (porciones) cromosómicas que tienen una separación suficiente (por ejemplo, mayor que un valor de corte) se pueden identificar como aberrantes (o que tienen una separación aberrante). Un patrón de región aberrante (que potencialmente explica qué haplotipo es mayor que el otro) se puede comparar con un patrón de referencia (por ejemplo, determinado a partir de un sujeto que tiene cáncer, potencialmente un tipo particular de cáncer o un sujeto sano). Si los dos patrones son iguales dentro de un umbral (por ejemplo, menor de un número específico de regiones/porciones que difieren) que un patrón de referencia que tiene una clasificación particular, se puede identificar que el sujeto tiene esa clasificación para el trastorno. Dicha clasificación puede incluir un trastorno de impronta, por ejemplo, como se describe en el presente documento.

VII. Análisis de metilación de moléculas únicas para moléculas híbridas

Para evaluar más a fondo el rendimiento y la utilidad de las realizaciones divulgadas en el presente documento con respecto a la determinación de modificaciones de bases de ácidos nucleicos, creamos artificialmente fragmentos de ADN híbridos humanos y de ratón para los cuales la parte humana estaba metilada y la parte de ratón no estaba metilada, o viceversa. La determinación de uniones de moléculas de ADN híbridas o quiméricas puede permitir detectar fusiones de genes para diversos trastornos o enfermedades, que incluyen el cáncer. Los métodos y análisis en esta sección no están de acuerdo con la invención y están presentes solo con fines ilustrativos.

A. Métodos para crear fragmentos de ADN híbridos humanos y de ratón.

Esta sección describe la creación de fragmentos de ADN híbridos y luego un procedimiento para determinar los perfiles de metilación de los fragmentos.

El ADN humano se amplificó a través de la amplificación del genoma completo de manera que la firma de metilación original en el genoma humano se eliminaría porque la amplificación del genoma completo no preservaría los estados de metilación. La amplificación del genoma completo se podría realizar utilizando hexámeros degenerados modificados con tiofosfato resistentes a exonucleasas como cebadores que podrían unirse aleatoriamente a un genoma, permitiendo que la polimerasa (por ejemplo, la ADN polimerasa Phi29) amplifique el ADN sin ciclos térmicos. El producto de ADN amplificado no estaría metilado. Las moléculas de ADN humano amplificadas se trataron adicionalmente con M.SssI, una metiltransferasa CpG, que en teoría metilaría completamente todas las citosinas en el contexto CpG en ADN de hebra doble, no metilado o hemimetilado. Por tanto, dicho ADN humano amplificado tratado con M.SssI se convertiría en moléculas de ADN metiladas.

Por el contrario, el ADN del ratón se sometió a amplificación del genoma completo de tal manera que se produjeran fragmentos de ADN de ratón no metilados.

Con fines ilustrativos, la FIG. 110 ilustra la creación de fragmentos de ADN híbridos humano-ratón para los cuales la parte humana está metilada mientras que la parte del ratón no está metilada. Las paletas rellenas representan sitios CpG metilados. Las paletas sin rellenar representan sitios CpG no metilados. La barra gruesa 11010 con rayas diagonales representa la parte humana metilada. La barra gruesa 11020 con rayas verticales representa la parte del ratón sin metilar.

Para la generación de moléculas de ADN híbridas de humano-ratón, en una realización, las moléculas de ADN amplificadas con el genoma completo y tratadas con M.SssI se digirieron adicionalmente con HindIII y NcoI para generar extremos pegajosos para facilitar la ligación en dirección descendente. Los fragmentos de ADN humano metilados se mezclaron además con los fragmentos de ADN de ratón no metilados en una relación equimolar. Dicha mezcla de ADN humano-ratón se sometió a un proceso de ligación, que en una realización estuvo mediado por ADN ligasa a 20 °C durante 15 minutos. Como se muestra en la FIG. 110, esta reacción de ligación produciría 3 tipos de moléculas resultantes, que incluyen moléculas de ADN híbridas humano-ratón (a: fragmentos híbridos humano-ratón); moléculas de ADN solo humanas (b: ligación humano-humano y c: ADN humano sin ligación); y moléculas de ADN solo de ratón (d: ligación ratón-ratón y e: ADN de ratón sin ligación). El producto de ADN después de la ligación se sometió a secuenciación en tiempo real de única molécula. Los resultados de la secuenciación se analizaron de acuerdo con la divulgación proporcionada en el presente documento para determinar los estados de metilación.

Con fines ilustrativos, la FIG. 111 ilustra la creación de fragmentos de ADN híbridos humano-ratón para los cuales la parte humana no está metilada mientras que la parte del ratón está metilada. Las paletas rellenas representan sitios CpG metilados. Las paletas sin rellenar representan sitios CpG no metilados. La barra gruesa 11110 con rayas diagonales representa la parte metilada del ratón. La barra gruesa 11120 con rayas verticales representa la parte humana sin metilar.

Para el ejemplo de la FIG. 111, las moléculas de ADN del ratón se amplificaron a través de la amplificación del genoma completo de tal manera que se eliminara la metilación original en el genoma del ratón. El producto de ADN amplificado no estaría metilado. El ADN de ratón amplificado se trataría adicionalmente con M.SssI. Por tanto, dicho ADN de ratón amplificado tratado con M.SssI se convertiría en moléculas de ADN metiladas. Por el contrario, los fragmentos de ADN humano se sometieron a una amplificación del genoma completo para obtener fragmentos humanos no metilados. En una realización, los fragmentos humanos metilados se mezclaron además con los fragmentos no metilados en una relación equimolar. Dicha mezcla de ADN humano-ratón se sometió a un proceso de ligación mediado por ADN ligasa. Como se muestra en la FIG. 111, esta reacción de ligación produciría 3 tipos de moléculas resultantes, que incluyen moléculas de ADN híbridas humano-ratón (a: fragmentos híbridos humano-ratón); moléculas de ADN solo humanas (b: ligación humano-humano y c: ADN humano sin ligación); y moléculas de ADN solo de ratón (d: ligación ratón-ratón y e: ADN de ratón sin ligación). El producto de ADN después de la ligación se sometió a secuenciación en tiempo real de única molécula. Los resultados de la secuenciación se analizaron de acuerdo con la divulgación proporcionada en el presente documento para determinar los estados de metilación.

De acuerdo con el ejemplo mostrado en la FIG. 110, preparamos una mezcla de ADN artificial (llamada muestra MIX01) que comprende moléculas de ADN híbridas humano-ratón, ADN solo humano y ADN solo de ratón para las cuales las moléculas de ADN asociadas a humanos estaban metiladas mientras que las moléculas de ADN de ratón no estaban metiladas. Para la muestra MIX01, obtuvimos 166 millones de sublecturas que se podrían alinear con un genoma de referencia humano o de ratón, o parcialmente con un genoma humano y parcialmente con un genoma de ratón. Estas sublecturas se generaron a partir de aproximadamente 5 millones de pocillos de Secuenciación en Tiempo Real de Único Molécula (SMRT) de Pacific Biosciences. Cada molécula en un pocillo de secuenciación en tiempo real de única molécula se secuenció en promedio 32 veces (rango: 1 - 881 veces).

Para determinar la parte de ADN humano y ADN de ratón en un fragmento híbrido, primero construimos secuencias de consenso al combinar la información de nucleótidos de todas las sublecturas relevantes en un pocillo. En total, obtuvimos 3,435,657 secuencias consenso para la muestra MIX01. El conjunto de datos se generó a partir de ADN preparado con el Kit de Secuenciación Sequel II 1.0.

Las secuencias consenso se alinearon con los genomas de referencia que comprenden tanto las referencias humanas como las de ratón. Obtuvimos 3.2 millones de secuencias consenso alineadas. Entre ellas, el 39.6 % se clasificaron como tipo de ADN solo humano; el 26.5 % de ellos se clasificaron como tipo de ADN solo de ratón y el 30.2 % de ellos como ADN híbrido humano-ratón.

Con fines ilustrativos, la FIG. 112 muestra la distribución de longitud de las moléculas de ADN en la mezcla de ADN después de la ligación (muestra MIX01). El eje x muestra la longitud de una molécula de ADN. El eje y muestra la frecuencia asociada con la longitud de la molécula de ADN. Como se muestra en la FIG. 112, las moléculas de ADN híbridas humano-ratón tenían una distribución de longitud más larga, lo que era consistente con el hecho de que eran una combinación de al menos dos tipos de moléculas.

Con fines ilustrativos, la FIG. 113 ilustra una región de unión mediante la cual se unen un primer ADN (A) y un segundo ADN (B). El ADN (A) y el ADN (B) se pueden digerir con una enzima de restricción. En una realización, para mejorar la eficacia de la ligación utilizando los extremos escalonados, utilizamos las enzimas de restricción HindIII y NcoI, que reconocen los sitios A^AGCTT y C^CATGG respectivamente, para digerir el ADN humano y de ratón antes de la etapa

de ligación. A continuación se pueden ligar el ADN (A) y el ADN (B). Entre 698,492 moléculas de ADN híbrido humano-ratón que albergan regiones de unión, encontramos que el 88 % de las moléculas de ADN híbrido humano-ratón que llevaban el sitio de reconocimiento enzimático de A^AGCTT y C^CATGG, lo que sugiere además que se había producido la ligación entre fragmentos de ADN humano y de ratón. Dicha región de unión se define como una región o sitio mediante el cual se unieron físicamente un primer fragmento de ADN y un segundo fragmento de ADN. Debido a que la unión incluye secuencias comunes tanto al ADN (A) como al ADN (B), no se puede determinar que la porción de una hebra correspondiente a la unión sea parte del ADN (A) o del ADN (B) solo por la secuencia. Se puede utilizar el análisis del patrón de metilación o la densidad de la porción de una hebra correspondiente a la unión para determinar si la porción es de ADN (A) o ADN (B). Como un ejemplo, el ADN (A) puede ser ADN viral y el ADN (B) puede ser ADN humano. La determinación de la unión exacta puede informar si dicho ADN integrado altera las estructuras de las proteínas y cómo.

Con fines ilustrativos, la FIG. 114 ilustra el análisis de metilación para la mezcla de ADN. La barra 11410 con rayas diagonales indica una región de unión observada en el análisis de alineación que se introduciría mediante un tratamiento con enzimas de restricción antes de la ligación. “Sitio RE” indica sitio de reconocimiento de enzima de restricción (RE).

Como se muestra en la FIG. 114, en una realización, las secuencias consenso alineadas se agruparon en tres categorías de la siguiente manera:

(1) Un ADN secuenciado solo se alineó con un genoma de referencia humano pero no con un genoma de referencia de ratón, en referencia a uno o más criterios de alineación. En un ejemplo, un criterio de alineación se podría definir como, pero no limitado a, 100 %, 95 %, 90 %, 80 %, 70 %, 60 %, 50 %, 40 %, 30 % o 20 % de nucleótidos contiguos de un ADN secuenciado se podría alinear con una referencia humana. En un ejemplo, un criterio de alineación sería que la parte restante del fragmento secuenciado que no se alineó con la referencia humana no pueda alinearse con un genoma de referencia de ratón. En un ejemplo, un criterio de alineación fue que el ADN secuenciado pudiera alinearse con una única región en un genoma humano de referencia. En un ejemplo, la alineación podría ser perfecta. Todavía en otra realización, el alineamiento se podría adaptar a discrepancias de nucleótidos, que incluyen inserciones, emparejamientos erróneas y supresiones, siempre que dichas discrepancias fueran inferiores a ciertos umbrales, tales como, pero no limitados a, 1 %, 2 %, 3 %, 4 %, 5 %, 10 %, 20 % o 30 % de la longitud de las secuencias alineadas. En otro ejemplo, el alineado podría estar en más de una ubicación en un genoma de referencia. Aún en otros ejemplos, el alineamiento con uno o más sitios en un genoma de referencia podría indicarse de manera probabilística (por ejemplo, indicando la posibilidad de un alineamiento erróneo), y la medición de probabilidades se podría utilizar en el procesamiento posterior.

(2) Un ADN secuenciado solo se alineó con un genoma de referencia de ratón pero no con un genoma de referencia humano, en referencia a uno o más criterios de alineación. En un ejemplo, un criterio de alineación se podría definir como, pero no limitado a, 100 %, 95 %, 90 %, 80 %, 70 %, 60 %, 50 %, 40 %, 30 % o 20 % de nucleótidos contiguos de un ADN secuenciado se podría alinear con una referencia de ratón. En un ejemplo, un criterio de alineación sería que la parte restante no pudiera alinearse con un genoma humano de referencia. En un ejemplo, un criterio de alineación fue que el ADN secuenciado se pudiera alinear con una única región en un genoma de ratón de referencia. En un ejemplo, la alineación podría ser perfecta. Todavía en otras realizaciones, el alineamiento se podría adaptar a discrepancias de nucleótidos, que incluyen inserciones, emparejamientos erróneos y supresiones, siempre que dichas discrepancias fueran menores a ciertos umbrales, tales como, pero no se limitan a, 1 %, 2 %, 3 %, 4 %, 5 %, 10 %, 20 % o 30 % de la longitud de las secuencias alineadas. En otro ejemplo, el alineado podría estar en más de una ubicación en un genoma de referencia. Todavía en otros ejemplos, el alineamiento con uno o más sitios en un genoma de referencia podría indicarse de manera probabilística (por ejemplo, indicando la posibilidad de un alineamiento erróneo), y la medición de probabilidades se podría utilizar en el procesamiento posterior.

(3) Una parte de un ADN secuenciado se alineó de manera única con un genoma de referencia humano, mientras que otra parte se alineó de manera única con un genoma de referencia de ratón. En un ejemplo, si se utilizara una enzima de restricción antes de la ligación, se observaría una región de unión en el análisis de alineación, correspondiente al sitio de corte de la enzima de restricción. En algunos ejemplos, las regiones de unión entre partes de ADN humano y de ratón solo se pudieron determinar aproximadamente dentro de una determinada región debido a errores de secuenciación y alineación. En algunos ejemplos, los sitios de reconocimiento de enzimas de restricción no serían observables en las regiones de unión de fragmentos de ADN híbridos humano-ratón si la ligación involucrara moléculas sin el corte de enzimas de restricción (por ejemplo, si hubiera una ligación de extremos romos).

Las duraciones interpulsos (IPD), las anchuras de pulso (PW) y el contexto de secuencia que rodea los sitios CpG se obtuvieron a partir de aquellas sublecturas correspondientes a las secuencias de consenso. De este modo, la metilación de cada molécula de ADN, que incluyen el ADN solo humano, solo de ratón e híbrido humano-ratón, se podría determinar de acuerdo con las realizaciones presentes en esta divulgación.

B. Resultados de la metilación

Esta sección describe los resultados de la metilación para fragmentos de ADN híbridos. Las densidades de metilación se pueden utilizar para identificar los orígenes de diferentes partes de fragmentos de ADN híbrido.

Con fines ilustrativos, la FIG. 115 muestra un diagrama de caja de las probabilidades de ser metilado para sitios CpG en la muestra MIX01. El eje x muestra las tres moléculas diferentes presentes en la muestra MIX01: ADN solo humano, ADN solo de ratón y ADN híbrido humano-ratón (incluye una parte humana y una parte de ratón). El eje y muestra la probabilidad de que se metile un sitio CpG de una molécula de ADN particular. Este ensayo se realizó de manera que el ADN humano estuviera más metilado mientras que el ADN del ratón estuviera menos metilado.

Como se muestra en la FIG. 115, la probabilidad de ser metilado para un sitio CpG en el ADN solo humano (mediana: 0.66; rango: 0-1) fue significativamente mayor que la del ADN solo de ratón (mediana: 0.06; rango: 0-1) (valor de $p < 0.0001$). Estos resultados estaban en línea con el diseño del ensayo para el cual el ADN humano estaba más metilado debido al tratamiento con la CpG metiltransferasa M.SssI, mientras que el ADN del ratón estaba menos metilado porque la metilación no podía conservarse durante la amplificación del genoma completo. Más aún, los sitios CpG dentro de la parte del ADN humano en una molécula de ADN híbrida humano-ratón mostraron una mayor probabilidad de estar metilados (mediana: 0.69; rango: 0-1) en comparación con aquellos dentro de la parte del ADN del ratón (mediana: 0.06; rango : 0-1) (valor de $P < 0.0001$). Estos datos indican que el método divulgado podría determinar con precisión el estado de metilación de las moléculas de ADN, así como los segmentos dentro de una molécula de ADN.

La probabilidad de metilación se refiere a la probabilidad estimada de un sitio CpG particular dentro de una única molécula en base al modelo estadístico utilizado. Una probabilidad de 1 indica que, de acuerdo con el modelo estadístico, el 100 % de los sitios CpG que utilizan los parámetros medidos (incluidos IPD, PW y contexto de secuencia) estarían metilados. Una probabilidad de 0 indica que, en base al modelo estadístico, el 0 % de los sitios CpG que utilizan los parámetros medidos (que incluyen IPD, PW y contexto de secuencia) estarían metilados. En otras palabras, todos los sitios CpG que utilicen los parámetros medidos no estarían metilados. La FIG. 115 muestra una distribución de las probabilidades de metilación, con una distribución más amplia para el ADN solo humano y la parte humana que para las contrapartes de ratón. La secuenciación con bisulfito se utiliza para medir la metilación de muestras similares para confirmar que la metilación no fue completa y los resultados se muestran a continuación. La FIG. 115 muestra una diferencia significativa entre la metilación en ADN humano versus ADN de ratón.

De acuerdo con la realización mostrada en la FIG. 111, preparamos una mezcla de ADN artificial (denominada muestra MIX02) que comprende moléculas de ADN híbridas humano-ratón, ADN solo humano y ADN solo de ratón en el que la parte humana no está metilada y la parte del ratón está metilada. Para la muestra MIX02, obtuvimos 140 millones de sublecturas que se podrían alinear con un genoma de referencia humano o de ratón, o parcialmente con un genoma humano y parcialmente con un genoma de ratón. Estas sublecturas se generaron a partir de aproximadamente 5 millones de pocillos de Secuenciación en Tiempo Real de Única Molécula (SMRT) de Pacific Biosciencees. Cada molécula en un pocillo de secuenciación en tiempo real de única molécula se secuenció en promedio 27 veces (rango: 1 - 1028 veces).

También construimos secuencias de consenso al combinar la información de nucleótidos de todas las sublecturas relevantes en un pocillo. En total, obtuvimos 3,265,487 secuencias consenso para la muestra MIX02. Las secuencias consenso se alinearon con los genomas de referencia que comprenden tanto las referencias humanas como las de ratón utilizando BWA (Li H et al., Bioinformatics. 2010;26(5):589-595). Obtuvimos 3.0 millones de secuencias consenso alineadas. Entre ellas, el 30.5 % se clasificaron como de tipo ADN solo humano; el 32.2 % se clasificó como tipo de ADN solo de ratón y el 33.8 % se clasificó como ADN híbrido humano-ratón. El conjunto de datos se generó a partir de ADN preparado con el Kit de Secuenciación Sequel II 1.0.

Con fines ilustrativos, la FIG. 116 muestra la distribución de longitud de las moléculas de ADN en la mezcla de ADN después de la ligación cruzada de la muestra MIX02. El eje x muestra la longitud de una molécula de ADN. El eje y muestra la frecuencia asociada con la longitud de la molécula de ADN. Como se muestra en la FIG. 116, las moléculas de ADN híbridas humano-ratón tenían una distribución de longitud más larga, lo que concuerda con el hecho de que se produjeron mediante la ligación de más de una molécula.

Con fines ilustrativos, la FIG. 117 muestra un diagrama de caja de las probabilidades de ser metilado para sitios CpG en la muestra MIX02. El estado de metilación se determinó de acuerdo con los métodos descritos en el presente documento. El eje x muestra las tres moléculas diferentes presentes en la muestra MIX01: ADN solo humano, ADN solo de ratón y ADN híbrido humano-ratón (incluye una parte humana y una parte de ratón). El eje y muestra la probabilidad de que un sitio CpG esté metilado. Este ensayo se realizó de manera que el ADN humano no estuviera metilado mientras que el ADN del ratón estaba metilado.

Como se muestra en la FIG. 117, la probabilidad de ser metilado para los sitios CpG en el ADN solo humano (mediana: 0.06; rango: 0-1) fue significativamente menor que la del ADN solo de ratón (mediana: 0.93; rango: 0-1) (valor de $p < 0.0001$). Estos resultados estaban en línea con el diseño del ensayo para el cual el ADN humano estaba menos metilado porque la metilación no podía conservarse durante la amplificación del genoma completo, mientras que el ADN del ratón estaba más metilado debido al tratamiento con la CpG metiltransferasa M.SssI. Más aún, los sitios CpG dentro de la parte del ADN humano en una molécula de ADN híbrida humano-ratón mostraron menores probabilidades de ser metilados (mediana: 0.07; rango: 0-1) en comparación con aquellos dentro de la parte del ADN del ratón (mediana: 0.93; rango: 0-1) (valor de $p < 0.0001$). Estos datos indican que el método divulgado podría determinar con precisión el estado de metilación de las moléculas de ADN, así como los segmentos dentro de una molécula de ADN.

Se utilizó secuenciación con bisulfito para medir la metilación de fragmentos híbridos humano-ratón cuyos patrones de metilación se determinaron mediante secuenciación en tiempo real de única molécula de acuerdo con realizaciones en esta divulgación. La muestra MIX01 (el ADN humano estaba metilado y el ADN de ratón no estaba metilado) y MIX02 (el ADN humano no estaba metilado y el ADN de ratón estaba metilado) se cortaron dando como resultado una mezcla con un tamaño de fragmento de ADN de mediana de 196 bp (rango intercuartil: 161 - 268). Mdiante sonicación. Luego se realizó la secuenciación con bisulfito de extremos emparejados (BS-Seq) en la plataforma MiSeq (Illumina) con una longitud de lectura de 300 bp x2. Obtuvimos 3.7 millones y 2.9 millones de fragmentos secuenciados para MIX01 y MIX02, respectivamente, que se alinearon con el genoma de referencia humano o de ratón, o parcialmente con un genoma humano y parcialmente con un genoma de ratón. Para MIX01, el 41.6 % de los fragmentos alineados se clasificaron como ADN solo humano, el 56.6 % como ADN solo de ratón y el 1.8 % como ADN híbrido humano-ratón. Para MIX02, el 61.8 % de los fragmentos alineados se clasificaron como ADN solo humano, el 36.3 % como ADN solo de ratón y el 1.9 % como ADN híbrido humano-ratón. El porcentaje de fragmentos secuenciados que se determinó que eran ADN híbrido humano-ratón en BS-Seq (<2 %) fue mucho menor que el observado en los resultados de secuenciación de Pacific Biosciences (>30 %). En particular, los fragmentos largos (una mediana de ~2 kb) fueron secuenciados mediante secuenciación de Pacific Biosciences, mientras que los fragmentos largos se dividieron en fragmentos cortos (una mediana de ~196 bp) que eran adecuados para MiSeq. Dicho proceso de corte diluiría en gran medida los fragmentos híbridos humano-ratón.

Con fines ilustrativos, la FIG. 118 muestra una tabla que compara la metilación determinada mediante secuenciación con bisulfito y secuenciación de Pacific Biosciences para MIX01. La sección más a la izquierda de la tabla muestra el tipo de ADN: 1) solo humano; 2) solo ratón; y 3) híbrido humano-ratón, dividido en la parte humana y la parte de ratón. La sección central de la tabla muestra detalles de la secuenciación de bisulfito, que incluye el número de sitios CG y la densidad de metilación. La sección más a la derecha de la tabla muestra detalles de la secuenciación de Pacific Biosciences, que incluye el número de sitios de CG y la densidad de metilación.

Como se muestra en la FIG. 118, el ADN solo humano exhibió consistentemente una mayor densidad de metilación que el ADN solo de ratón para MIX01 tanto en la secuenciación con bisulfito como en los resultados de la secuenciación de Pacific Biosciences. Para los fragmentos híbridos humano-ratón, se determinó que los niveles de metilación de la parte humana y la parte del ratón eran 46.8 % y 2.3 %, respectivamente, en los resultados de la secuenciación con bisulfito. Estos resultados confirmaron las mayores densidades de metilación para la parte humana en comparación con la parte del ratón según lo determinado por la secuenciación de Pacific Biosciences de acuerdo con la divulgación. Con la secuenciación de Pacific Biosciences se observó una densidad de metilación del 57.4 % en la parte humana y una menor densidad de metilación del 12.1 % en la parte del ratón. Estos resultados sugieren que la metilación determinada mediante la secuenciación de Pacific Biosciences de acuerdo con esta divulgación podría ser factible. En particular, la secuenciación de Pacific Biosciences se puede utilizar para determinar diferentes densidades de metilación, que incluyen en ADN que tiene una sección con una densidad de metilación mayor que otra sección. Observamos que la densidad de metilación determinada por la secuenciación de Pacific Biosciences de acuerdo con la divulgación fue mayor en relación con la secuenciación con bisulfito. Dicha estimación se puede ajustar utilizando la diferencia entre los resultados determinados por estas dos tecnologías para comparar los resultados entre las tecnologías.

Con fines ilustrativos, la FIG. 119 muestra una tabla que compara la metilación determinada mediante secuenciación con bisulfito y secuenciación de Pacific Biosciences para MIX02. La sección más a la izquierda de la tabla muestra el tipo de ADN: 1) solo humano; 2) solo ratón; y 3) híbrido humano-ratón, dividido en la parte humana y la parte de ratón. La sección central de la tabla muestra detalles de la secuenciación de bisulfito, que incluye el número de sitios CG y la densidad de metilación. La sección más a la derecha de la tabla muestra detalles de la secuenciación de Pacific Biosciences, que incluye el número de sitios de CG y la densidad de metilación.

Como se muestra en la FIG. 119, el ADN solo humano exhibió consistentemente una densidad de metilación más baja que el ADN solo de ratón para MIX02 tanto en la secuenciación con bisulfito como en los resultados de la secuenciación de Pacific Biosciences. Para los fragmentos híbridos humano-ratón, se determinó que los niveles de metilación de la parte humana y la parte del ratón eran 1.8 % y 67.4 %, respectivamente, en los resultados de la secuenciación con bisulfito. Estos resultados confirmaron aún más las densidades de metilación más bajas para la parte humana en comparación con la parte del ratón según lo determinado por la secuenciación de Pacific Biosciences de acuerdo con la divulgación. Con la secuenciación de Pacific Biosciences, se observó una densidad de metilación del 13.1 % en la parte humana y una mayor densidad de metilación del 72.2 % en la parte del ratón según lo determinado mediante la secuenciación de Pacific Biosciences de acuerdo con esta divulgación. También sugirió que era factible determinar la metilación mediante la secuenciación de Pacific Biosciences de acuerdo con esta divulgación. En particular, la secuenciación de Pacific Biosciences se puede utilizar para determinar diferentes densidades de metilación, que incluye en ADN que tiene una sección con una densidad de metilación menor que otra sección. También observamos que la densidad de metilación determinada por la secuenciación de Pacific Biosciences de acuerdo con la divulgación fue mayor en relación con la secuenciación con bisulfito. Dicha estimación se puede ajustar utilizando la diferencia entre los resultados determinados por estas dos tecnologías para comparar los resultados entre las tecnologías.

Con fines ilustrativos, la FIG. 120A muestra los niveles de metilación en intervalos de 5 Mb para ADN solo de humanos y de ratón para MIX01. La FIG. 120B muestra los niveles de metilación en intervalos de 5 Mb para ADN solo de humanos y de ratón para MIX02. En ambas figuras, el nivel de metilación en porcentaje se muestra en el eje y. En el

eje x se muestra la secuenciación con bisulfito y la secuenciación de Pacific Biosciences para cada uno de los ADN de solo humano y ADN de solo ratón.

Los resultados en las FIG. 120A y FIG. 120B. determinados por la secuenciación de Pacific Biosciences de acuerdo con la divulgación se encontraron que era sistémicamente más altos en todos los intervalos tanto en la muestra MIX01 como en la MIX02.

Con fines ilustrativos, la FIG. 121A muestra los niveles de metilación en intervalos de 5 Mb para la parte humana y la parte de ratón de fragmentos de ADN híbridos humano-ratón para MIX01. La FIG. 121B muestra los niveles de metilación en intervalos de 5 Mb para la parte humana y la parte de ratón de fragmentos de ADN híbridos humano-ratón para MIX02. En ambas figuras, el nivel de metilación en porcentaje se muestra en el eje y. En el eje x se muestra la secuenciación con bisulfito y la secuenciación de Pacific Biosciences para cada parte del ADN humano y del ADN de la parte del ratón.

La FIG. 121A y FIG. 121B mostraron un aumento en el nivel de metilación cuando se utiliza la secuenciación de Pacific Biosciences en comparación con la secuenciación con bisulfito. Este aumento es similar al aumento en los niveles de metilación mediante la secuenciación de Pacific Biosciences observado con ADN solo de humanos y ADN solo de ratón en la FIG. 120A y FIG. 120B. La mayor variabilidad en los niveles de metilación en los intervalos de 5 Mb presentes en los resultados de la secuenciación con bisulfito para fragmentos híbridos probablemente se debió al menor número de sitios CpG utilizados para el análisis.

Las FIG. 122A y 122B son gráficos representativos que muestran estados de metilación en una única molécula híbrida humano-ratón. La FIG. 122A muestra un fragmento híbrido humano-ratón en la muestra MIX01. La FIG. 122B muestra un fragmento híbrido humano-ratón en la muestra MIX02. Un círculo relleno indica un sitio metilado y un círculo sin relleno indica un sitio no metilado. Los estados de metilación en estos fragmentos se determinaron de acuerdo con las realizaciones descritas en el presente documento.

Como se muestra en la FIG. 122A, se determinó que la parte humana de una molécula híbrida de la muestra MIX01 estaba más metilada. Por el contrario, se determinó que la parte del ADN del ratón estaba más hipometilada. Por el contrario, la FIG. 122B muestra que se determinó que la parte humana de una molécula híbrida de la muestra MIX02 estaba más hipometilada, mientras que se determinó que la parte de ADN de ratón estaba más metilada.

Estos resultados demostraron que las realizaciones presentes en esta divulgación permitieron determinar los cambios de metilación en una única molécula de ADN con diferentes patrones de metilación en diferentes partes de la molécula. En una realización, se puede medir el estado de metilación de un gen u otras regiones genómicas en las que diferentes partes del gen o regiones genómicas exhibirían un estado de metilación diferente (por ejemplo, el promotor versus el cuerpo del gen). En otra realización, los métodos presentados en el presente documento pueden detectar los fragmentos híbridos humano-ratón, proporcionando un enfoque genérico para detectar moléculas de ADN que contienen fragmentos no contiguos (es decir, moléculas quiméricas) con respecto a un genoma de referencia y analizar sus estados de metilación. Por ejemplo, podríamos utilizar este enfoque para analizar, pero no se limitan a, fusiones de genes, reordenamientos genómicos, traducciones, inversiones, duplicaciones, variaciones estructurales, integraciones de ADN viral, recombinaciones meióticas, etc.

En algunas realizaciones, estos fragmentos híbridos podrían enriquecerse antes de la secuenciación utilizando métodos de hibridación basados en sondas o sistemas CRISPR-Cas o sus enfoques variantes para el enriquecimiento del ADN diana. Recientemente, se informó que una transposasa asociada a CRISPR de la cianobacteria *Scytonema hofmanni* fue capaz de insertar segmentos de ADN en una región cercana al sitio dirigido de interés (Strecker et al. Science. 2019;365:48-53). La transposasa asociada a CRISPR podría funcionar como la transposición mediada por Tn7. En una realización, podríamos adaptar esta transposasa asociada a CRISPR para insertar secuencias comentadas etiquetadas, por ejemplo, con biotina en una o más regiones genómicas de interés, guiadas por ARNg. Podríamos utilizar perlas magnéticas recubiertas con, por ejemplo, estreptavidina para capturar las secuencias comentadas, al extraer simultáneamente secuencias de ADN dirigidas para secuenciación y análisis de metilación de acuerdo con las realizaciones en esta divulgación.

En algunas realizaciones, los fragmentos se pueden enriquecer al utilizar enzimas de restricción, que pueden incluir cualquier enzima de restricción divulgada en el presente documento.

C. Método de detección de moléculas quiméricas de ejemplo

La FIG. 123 muestra un método 1230 para detectar moléculas quiméricas en una muestra biológica. La realización mostrada en la FIG. 123 no caen dentro del alcance de las reivindicaciones adjuntas y se debe considerar simplemente como un ejemplo adecuado para comprender la invención. Las moléculas quiméricas pueden incluir secuencias de dos diferentes genes, cromosomas, orgánulos (por ejemplo, mitocondrias, núcleos, cloroplastos), organismos (mamíferos, bacterias, virus, etc.) y/o especies. El método 1230 se puede aplicar a cada una de una pluralidad de moléculas de ADN de la muestra biológica. En algunos ejemplos, la pluralidad de moléculas de ADN puede ser ADN celular. En otros ejemplos, la pluralidad de moléculas de ADN pueden ser moléculas de ADN libres de células del plasma de una mujer embarazada.

En el bloque 1232, se puede realizar la secuenciación de única molécula de una molécula de ADN para obtener una lectura de secuencia que proporciona un estado de metilación en cada uno de los sitios N. N puede ser 5 o más, que incluye 5 a 10, 10 a 15, 15 a 20 o más de 20. Los estados de metilación de la secuencia leída pueden formar un patrón de metilación. La molécula de ADN puede ser una molécula de ADN de la pluralidad de moléculas de ADN, y el método 1230 se puede realizar en la pluralidad de moléculas de ADN. El patrón de metilación puede adoptar varias formas. Por ejemplo, el patrón podría ser N (por ejemplo, 2, 3, 4, etc.) sitios metilados seguidos de N sitios no metilados, o viceversa. Dicho cambio en la metilación puede indicar una unión. El número de sitios contiguos que están metilados puede ser diferente del número de sitios contiguos que no están metilados.

En el bloque 1234, el patrón de metilación se puede deslizar sobre uno o más patrones de referencia que corresponden a moléculas quiméricas que tienen dos porciones de dos partes de un genoma humano de referencia. Un patrón de referencia puede actuar como filtro para identificar un patrón coincidente que sea indicativo de una unión. Se puede realizar un seguimiento del número de sitios que coinciden con el patrón de referencia de tal manera que una posición coincidente corresponda a un número máximo de sitios coincidentes (es decir, el número en el que el estado de metilación coincide con el patrón de referencia). Las dos partes del genoma humano de referencia pueden ser partes discontinuas del genoma humano de referencia. Las dos partes del genoma humano de referencia pueden estar separadas por más de 1 kb, 5 kb, 10 kb, 100 kb, 1 Mb, 5 Mb o 10 Mb. Las dos partes pueden ser de dos brazos cromosómicos o cromosomas diferentes. El uno o más patrones de referencia pueden incluir un cambio entre estados metilados y estados no metilados.

En el bloque 1236, se puede identificar una posición coincidente entre el patrón de metilación y un primer patrón de referencia de uno o más patrones de referencia. La posición coincidente puede identificar una unión entre las dos partes del genoma humano de referencia en la secuencia leída. La posición de coincidencia puede corresponder a un máximo en una función de superposición entre un patrón de referencia y el patrón de metilación. La función de superposición puede utilizar múltiples patrones de referencia, siendo la salida posiblemente un máximo sobre una función agregada (es decir, cada patrón de referencia contribuye a un valor de salida) o un máximo único que se identifica en todos los patrones de referencia.

En el bloque 1238, la unión se puede generar como una ubicación de una fusión genética en una molécula quimérica. La ubicación de la fusión genética se puede comparar con ubicaciones de referencia de fusiones genéticas para diversos trastornos o enfermedades, que incluyen el cáncer.

La posición coincidente se puede emitir a una función de alineación. Se puede refinar la ubicación de la fusión genética. Refinar la ubicación de la fusión genética puede incluir alinear una primera porción de la secuencia leída con una primera parte del genoma humano de referencia. La primera parte puede estar antes de la unión. Refinar la ubicación de la fusión genética puede incluir alinear una segunda porción de la secuencia leída con una segunda parte del genoma humano de referencia. La segunda parte puede estar después de la unión. La primera parte del genoma humano de referencia puede estar separada al menos 1 kb de la segunda parte del genoma humano de referencia. Por ejemplo, la primera parte del genoma humano de referencia y la segunda parte del genoma humano de referencia pueden tener de 1.0 a 1.5 kb, de 1.5 a 2.0 kb, de 2.0 a 2.5 kb, de 2.5 a 3.0 kb, de 3 a 5 kb, o más de 5 kb de diferencia.

Las uniones de múltiples moléculas quiméricas se pueden comparar entre sí para confirmar la ubicación de una fusión genética.

VIII. Conclusión

Hemos desarrollado un enfoque eficaz para predecir los niveles de metilación de ácidos nucleicos con resolución de base única. Este nuevo enfoque implementa un nuevo esquema para capturar simultáneamente la cinética de la polimerasa que rodea la base que se interroga, el contexto de la secuencia y la información de la hebra. Dicha nueva transformación de la cinética permitió identificar y modelar la sutil interrupción que se produce en los pulsos cinéticos. En comparación con los métodos anteriores utilizados solo por IPD, el nuevo enfoque presente en esta solicitud de patente ha mejorado mucho la resolución y la precisión en el análisis de metilación. Este nuevo esquema se podría ampliar fácilmente para otros fines, por ejemplo, detectar 5hmC (5-hidroximetilcitosina), 5fC (5-formilcitosina), 5caC (5-carboxilcitosina), 4mC (4-metilcitosina), 6mA (N6-metiladenina), 8oxoG (7,8-dihidro-8-oxoguanina), 8oxoA (7,8-dihidro-8-oxoadenina) y otras formas de metilación. En otra realización, este nuevo esquema (por ejemplo, transformación cinética análoga a la matriz digital 2-D presente en esta solicitud) se podría utilizar para el análisis de emetilación con el uso de un sistema de secuenciación de nanoporos.

Esta implementación de detección de metilación se podría utilizar para muestras de ácido nucleico de diferentes fuentes, por ejemplo, ácidos nucleicos celulares, ácidos nucleicos de muestreo ambiental (por ejemplo, contaminantes celulares), ácidos nucleicos de patógenos (por ejemplo, bacterias y hongos), y ADNlc en el plasma de mujeres embarazadas. Abriría muchas posibilidades nuevas para la investigación genómica y el diagnóstico molecular, tales como las pruebas prenatales no invasivas, la detección del cáncer y el monitorización de trasplantes. Para el diagnóstico prenatal no invasivo basado en ADNlc, este nuevo método ha hecho posible el uso simultáneo de aberraciones en el número de copias, tamaños, mutaciones, extremos de fragmentos y modificación de bases para cada molécula en diagnósticos sin PCR y conversión experimental antes de la secuenciación, mejorando de esta manera la sensibilidad. Los desequilibrios en los niveles de metilación entre haplotipos se pueden detectar utilizando

los métodos descritos en el presente documento. Dichos desequilibrios pueden indicar el origen de una molécula de ADN (por ejemplo, extraída de o un trastorno, como una célula cancerosa aislada de la sangre de un paciente con cáncer) o un trastorno.

IX. Sistemas de ejemplo

La FIG. 124 ilustra un sistema de medición 12400 de acuerdo con una realización de la presente invención. El sistema como se muestra incluye una muestra 12405, tal como moléculas de ADN dentro de un soporte de muestra 12410, donde la muestra 12405 se puede poner en contacto con un ensayo 12408 para proporcionar una señal de una característica física 12415. Un ejemplo de un soporte de muestra puede ser una celda de flujo que incluye sondas y/o cebadores de un ensayo o un tubo a través del cual se mueve una gotita (la gotita que incluye el ensayo). La característica física 12415 (por ejemplo, una intensidad de fluorescencia, un voltaje o una corriente) de la muestra es detectada por el detector 12420. El detector 12402 puede realizar una medición a intervalos (por ejemplo, intervalos periódicos) para obtener puntos de datos que constituyen una señal de datos. En una realización, un convertidor analógico a digital convierte una señal analógica del detector en forma digital una pluralidad veces. El soporte de muestra 12401 y el detector 12402 pueden formar un dispositivo de ensayo, por ejemplo, un dispositivo de secuenciación que realiza la secuenciación de acuerdo con las realizaciones descritas en el presente documento. Se envía una señal de datos 12425 desde el detector 12402 al sistema lógico 12403. La señal de datos 12425 se puede almacenar en una memoria local 12435, una memoria externa 12404 o un dispositivo de almacenamiento 12445.

El sistema lógico 12403 puede ser, o puede incluir, un sistema informático, ASIC, microprocesador, etc. También puede incluir o estar acoplado con una pantalla (por ejemplo, monitor, pantalla LED, etc.) y un dispositivo de entrada de usuario (por ejemplo, ratón, teclado, botones, etc.). El sistema lógico 12403 y los otros componentes pueden ser parte de un sistema informático independiente o conectado en red, o pueden estar directamente adheridos o incorporados en un dispositivo (por ejemplo, un dispositivo de secuenciación) que incluye el detector 12402 y/o el soporte de muestra 12401. El sistema lógico 12403 también puede incluir software que se ejecuta en un procesador 12405. El sistema lógico 12403 puede incluir un medio legible por ordenador que almacena instrucciones para controlar el sistema 12400 para realizar cualquiera de los métodos descritos en el presente documento. Por ejemplo, el sistema lógico 12403 puede proporcionar comandos a un sistema que incluye un soporte de muestra 12401 de manera que se realicen secuenciaciones u otras operaciones físicas. Dichas operaciones físicas se pueden realizar en un orden particular, por ejemplo, al agregar o eliminar reactivos en un orden particular. Dichas operaciones físicas se pueden realizar mediante un sistema robótico, por ejemplo, que incluye un brazo robótico, como se puede utilizar para obtener una muestra y realizar un ensayo.

Cualquiera de los sistemas informáticos mencionados en el presente documento puede utilizar cualquier número adecuado de subsistemas. Ejemplos de dichos subsistemas se muestran en la FIG. 125 en el sistema informático 10. En algunas realizaciones, un sistema informático incluye un único aparato informático, donde los subsistemas pueden ser los componentes del aparato informático. En otras realizaciones, un sistema informático puede incluir múltiples aparatos informáticos, siendo cada uno un subsistema, con componentes internos. Un sistema informático puede incluir ordenadores de escritorio y portátiles, ordenadores tipo tableta, teléfonos móviles, otros dispositivos móviles y sistemas basados en la nube.

Los subsistemas mostrados en la FIG. 125 están interconectados a través de un bus de sistema 75. Subsistemas adicionales tales como una impresora 74, un teclado 78, un dispositivo(s) de almacenamiento 79, un monitor 76 (por ejemplo, una pantalla de visualización, tal como un LED), que está acoplado al adaptador de visualización 82, y se muestran otros. Los periféricos y los dispositivos de entrada/salida (E/S), que se acoplan al controlador de E/S 71, se pueden conectar al sistema informático mediante cualquier número de medios conocidos en la técnica, tales como el puerto de entrada/salida (E/S) 77 (por ejemplo, USB, FireWire®). Por ejemplo, el puerto de E/S 77 o la interfaz externa 81 (por ejemplo, Ethernet, Wi-Fi, etc.) se pueden utilizar para conectar el sistema informático 10 a una red de área amplia tal como Internet, un dispositivo de entrada de ratón o un escáner. La interconexión a través del bus del sistema 75 permite que el procesador central 73 se comunique con cada subsistema y controle la ejecución de una pluralidad de instrucciones desde la memoria del sistema 72 o los dispositivos de almacenamiento 79 (por ejemplo, un disco fijo, tal como un disco duro), o disco óptico), así como el intercambio de información entre subsistemas. La memoria del sistema 72 y/o el(los) dispositivo(s) de almacenamiento 79 pueden incorporar un medio legible por ordenador. Otro subsistema es un dispositivo de recopilación de datos 85, tal como una cámara, micrófono, acelerómetro y similares. Cualquiera de los datos mencionados en el presente documento se puede emitir de un componente a otro componente y se puede emitir al usuario.

Un sistema informático puede incluir una pluralidad de los mismos componentes o subsistemas, por ejemplo, conectados entre sí mediante una interfaz externa 81, mediante una interfaz interna o mediante dispositivos de almacenamiento extraíbles que se pueden conectar y quitar de un componente a otro componente. En algunas realizaciones, los sistemas, subsistemas o aparatos informáticos pueden comunicarse a través de una red. En tales casos, un ordenador se puede considerar un cliente y otro ordenador un servidor, donde cada uno puede ser parte de un mismo sistema informático. Un cliente y un servidor pueden incluir cada uno múltiples sistemas, subsistemas o componentes.

Algunos aspectos de las realizaciones se pueden implementar en forma de lógica de control utilizando circuitos de hardware (por ejemplo, un circuito integrado de aplicación específica o una matriz de puertas programables en campo) y/o utilizando software informático con un procesador generalmente programable de manera modular o integrada. Como se utiliza en el presente documento, un procesador puede incluir un procesador de un solo núcleo, un procesador de múltiples núcleos en un mismo chip integrado o múltiples unidades de procesamiento en una sola placa de circuito o en red, así como hardware dedicado. Con base en la divulgación y las enseñanzas proporcionadas en el presente documento, una persona con conocimientos habituales en la técnica conocerá y apreciará otras formas y/o métodos para implementar realizaciones de la presente invención utilizando hardware y una combinación de hardware y software.

Cualquiera de los componentes o funciones de software descritos en esta solicitud se puede implementar como código de software para ser ejecutado por un procesador utilizando cualquier lenguaje informático adecuado tal como, por ejemplo, Java, C, C++, C#, Objective-C, Swift, o lenguaje de secuencias de comandos tal como Perl o Python utilizando, por ejemplo, técnicas convencionales u orientadas a objetos. El código de software se puede almacenar como una serie de instrucciones o comandos en un medio legible por ordenador para almacenamiento y/o transmisión. Un medio legible por ordenador no transitorio adecuado puede incluir memoria de acceso aleatorio (RAM), una memoria de sólo lectura (ROM), un medio magnético tal como un disco duro o un disquete, o un medio óptico tal como un disco compacto (CD) o DVD (disco versátil digital) o disco Blu-ray, memoria flash y similares. El medio legible por ordenador puede ser cualquier combinación de dichos dispositivos de almacenamiento o transmisión.

Dichos programas también se pueden codificar y transmitir utilizando señales portadoras adaptadas para la transmisión a través de redes cableadas, ópticas y/o inalámbricas que se ajusten a una variedad de protocolos, que incluyen Internet. Como tal, se puede crear un medio legible por ordenador utilizando una señal de datos codificada con dichos programas. Los medios legibles por ordenador codificados con el código del programa se pueden empacar con un dispositivo compatible o proporcionarse por separado de otros dispositivos (por ejemplo, mediante descarga de Internet). Cualquier medio legible por ordenador puede residir en o dentro de un solo producto informático (por ejemplo, un disco duro, un CD o un sistema informático completo) y puede estar presente en o dentro de diferentes productos informáticos dentro de un sistema o red. Un sistema informático puede incluir un monitor, una impresora u otra pantalla adecuada para proporcionar cualquiera de los resultados mencionados en el presente documento a un usuario.

Cualquiera de los métodos descritos en el presente documento se puede realizar total o parcialmente con un sistema informático que incluye uno o más procesadores, que se pueden configurar para realizar las etapas. Por lo tanto, las realizaciones se pueden dirigir a sistemas informáticos configurados para realizar las etapas de cualquiera de los métodos descritos en el presente documento, potencialmente con diferentes componentes que realizan una etapa respectiva o un grupo de etapas respectivo. Aunque se presentan como etapas numeradas, las etapas de los métodos del presente documento se pueden realizar al mismo tiempo o en momentos diferentes o en un orden diferente. Adicionalmente, se pueden utilizar porciones de estas etapas con porciones de otras etapas de otros métodos. También, todo o parte de una etapa puede ser opcional. Además, cualquiera de las etapas de cualquiera de los métodos se puede realizar con módulos, unidades, circuitos u otros medios de un sistema para realizar estas etapas.

Los detalles específicos de realizaciones particulares se pueden combinar de cualquier manera adecuada. Sin embargo, otras realizaciones de la invención pueden estar dirigidas a realizaciones específicas relacionadas con cada aspecto individual, o combinaciones específicas de estos aspectos individuales.

La descripción anterior de realizaciones de ejemplo de la presente divulgación se ha presentado con fines de ilustración y descripción. No pretende ser exhaustivo ni limitar la divulgación a la forma precisa descrita, y son posibles muchas modificaciones y variaciones a la luz de la enseñanza anterior.

Una recitación de “un”, “una” o “el” pretende significar “uno o más” a menos que se indique específicamente lo contrario. El uso de “o” pretende significar “o inclusivo” y no “o exclusivo” a menos que se indique específicamente lo contrario. La referencia a un “primer” componente no requiere necesariamente que se proporcione un segundo componente. Más aún, la referencia a un “primer” o un “segundo” componente no limita el componente al que se hace referencia a una ubicación particular a menos que se indique expresamente. El término “basado en” pretende significar “basado al menos en parte en”.

Ninguna patente, solicitudes de patente, publicaciones y descripción mencionadas en el presente documento se admiten como técnica anterior.

Referencias

Albert, T.J. et al. (2007) Direct selection of human genomic loci by microarray hybridization. Nat. Methods, 4, 903-905.
Beckmann et al. (2014) Detecting epigenetic motifs in low coverage and metagenomics settings. BMC Bioinformatics, 15(Suppl 9): S16.

Beaulaurier, J. et al. (2019) Deciphering bacterial epigenomes using modern sequencing technologies. Nature Reviews Genetics, 20:157-172.

Blow, M.J. et al. (2016) The Epigenomic Landscape of Prokaryotes. PLOS Genet., 12, e1005854.

Breiman, L. (2001) Random Forests. Mach. Learn., 45, 5-32.

Chan, K.C.A. et al. (2013) Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. Proc. Natl. Acad. Sci. U. S. A., 110, 18761-8.

Clark, T.A. et al. (2013) Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. BMC Biol., 11, 4.

Clark, T.A. et al. (2012) Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. Nucleic Acids Res., 40: e29.

Eid, J. et al. (2009) Real-Time DNA Sequencing from Single Polymerase Molecules. Science 323, 133-138.

Feinberg, A.P. and Irizarry, RA. (2010) Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proc. Natl. Acad. Sci., 107, 1757-1764.

Feng, Z. et al. (2013) Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic. PLoS Comput Biol, 9:e1002935.

Flusberg, B.A. et al. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat. Methods, 7, 461-465.

Frommer, M. et al. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc. Natl. Acad. Sci., 89, 1827-1831.

Gai, W. et al. (2018) Liver- and colon-specific DNA methylation markers in plasma for investigation of colorectal cancers with or without liver metastases. Clin. Chem., 64, 1239-1249.

Gouil, Q. et al. (2019) Latest techniques to study DNA methylation. Essays Biochem. 63(6):639-648.

Grunau, C. (2001) Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. Nucleic Acids Res., 29, 65e - 65.

Herman, J.G. et al. (1996) Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. Proc. Natl. Acad. Sci. U. S. A., 93, 9821-9826.

Jiang, P. et al. (2014) Methy-Pipe: An Integrated Bioinformatics Pipeline for Whole Genome Bisulfite Sequencing Data Analysis. PLoS One, 9, e100360.

LeCun, Y. et al. (1989) Backpropagation Applied to Handwritten Zip Code Recognition. Neural Comput., 1, 541-551.

Lee, E.-J. et al. (2011) Targeted bisulfite sequencing by solution hybrid selection and massively parallel sequencing. Nucleic Acids Res., 39, e127-e127.

Lehmann-Werman, R. et al. (2016) Identification of tissue-specific cell death using methylation patterns of circulating DNA. Proc. Natl. Acad. Sci., 113, E1826-E1834.

Lister, R. et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature, 462, 315-322.

Liu, Q. et al. (2019) Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. Nature Commun., 10, 2449.

Liu, Y. et al. (2019) Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. Nat. Biotechnol., 37, 424-429.

Lun, F.M.F. et al. (2013) Noninvasive prenatal methylomic analysis by genomewide bisulfite sequencing of maternal plasma DNA. Clin. Chem., 59, 1583-1594.

Nattestad, M. et al. (2018) Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. Genome Res., 28, 1126-1135.

Ng, A.Y. (2004) Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In, Twenty-first International Conference on Machine Learning - ICML '04. ACM Press, New York, New York, USA, p. 78.

Ni, P. et al. (2019) DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep- learning. Bioinformatics,35, 4586-4595

Okou, D.T. et al. (2007) Microarray-based genomic selection for high-throughput resequencing. Nat. Methods, 4, 907-909.

Olova, N. et al. (2018) Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biol., 19, 33.

Robertson, K.D. (2005) DNA methylation and human disease. Nat. Rev. Genet., 6, 597-610.

Smith, Z.D. and Meissner, A. (2013) DNA methylation: roles in mammalian development. Nat. Rev. Genet., 14, 204-20.

Schadt, E.E. et al. (2013) Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. Genome Res., 23(1):129-41.

Sun, K. et al. (2015) Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proc. Natl. Acad. Sci., 112, E5503-E5512.

Suzuki, Y. et al. (2016) AgIn: measuring the landscape of CpG methylation of individual repetitive elements. Bioinformatics, 32, 2911-2919.

Watson, C.M. et al. (2019) Cas9-based enrichment and single-molecule sequencing for precise characterization of genomic duplications. Lab. Investig, 100, 135-146.

Zhang, W. et al. (2015) Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. Genome Biol., 16, 14.

REIVINDICACIONES

1. Un método implementado por ordenador para detectar una metilación de un nucleótido en una molécula de ácido nucleico, el método comprende:

recibir una estructura de datos de entrada, la estructura de datos de entrada corresponde a una ventana de nucleótidos secuenciados en una molécula de ácido nucleico de muestra, en la que la molécula de ácido nucleico de muestra se secuencia al medir los pulsos en una señal óptica que corresponde a los nucleótidos, la estructura de datos de entrada comprende valores para las siguientes propiedades:

para cada nucleótido dentro de la ventana:

una identidad del nucleótido,

una posición del nucleótido con respecto a una posición diana dentro de la respectiva ventana,

una anchura del pulso que corresponde al nucleótido, y

una duración interpulso que representa un tiempo entre el pulso que corresponde al nucleótido y un pulso que corresponde a un nucleótido vecino;

ingresar la estructura de datos de entrada en un modelo, el modelo entrenado para:

recibir una primera pluralidad de primeras estructuras de datos, cada primera estructura de datos de la primera pluralidad de primeras estructuras de datos que corresponde a una ventana respectiva de nucleótidos secuenciados en una molécula de ácido nucleico respectiva de una pluralidad de primeras moléculas de ácido nucleico, en la que cada una de las primeras moléculas de ácido nucleico se secuencia al medir los pulsos en una señal que corresponde a los nucleótidos, en la que la metilación tiene un primer estado conocido en un nucleótido en una posición diana en cada ventana de cada primer molécula de ácido nucleico, cada primera estructura de datos comprende valores para las mismas propiedades como la estructura de datos de entrada,

almacenar una pluralidad de primeras muestras de entrenamiento, cada una incluye una de la primera pluralidad de primeras estructuras de datos y una primera etiqueta que indica el primer estado del nucleótido en la posición diana, y

optimizar, utilizando la pluralidad de primeras muestras de entrenamiento, los parámetros del modelo basado en salidas del modelo que coinciden o no coinciden con las etiquetas correspondientes de las primeras etiquetas cuando la primera pluralidad de primeras estructuras de datos es la entrada al modelo, en el que una salida del modelo especifica si el nucleótido en la posición diana en la respectiva ventana tiene la metilación,

determinar, utilizando el modelo, si la metilación está presente en un nucleótido en la posición diana dentro de la ventana en la estructura de datos de entrada.

2. El método de la reivindicación 1, en el que:

la estructura de datos de entrada es una estructura de datos de entrada de una pluralidad de estructuras de datos de entrada, la molécula de ácido nucleico de muestra es una molécula de ácido nucleico de muestra de una pluralidad de moléculas de ácido nucleico de muestra,

la pluralidad de moléculas de ácido nucleico de muestra se obtiene a partir de una muestra biológica de un sujeto, y cada estructura de datos de entrada corresponde a una ventana respectiva de nucleótidos secuenciados en una molécula de ácido nucleico respectiva de muestra de la pluralidad de moléculas de ácido nucleico de muestra, y

el método comprende además:

recibir la pluralidad de estructuras de datos de entrada,

ingresar la pluralidad de estructuras de datos de entrada en el modelo, y

determinar, utilizando el modelo, si está presente una metilación en un nucleótido en una ubicación diana en la respectiva ventana de cada estructura de datos de entrada.

3. El método de la reivindicación 2, que comprende además:

determinar si la metilación está presente en uno o más nucleótidos, y

determinar una clasificación de un trastorno utilizando la presencia de la metilación en uno o más nucleótidos.

4. El método de la reivindicación 3, en el que el trastorno comprende cáncer.

5. El método de la reivindicación 3, en el que determinar la clasificación del trastorno utiliza el número de metilaciones o los sitios de las metilaciones.

6. El método de la reivindicación 2, en el que cada molécula de ácido nucleico de muestra de la pluralidad de moléculas de ácido nucleico de muestra tiene un tamaño mayor que un tamaño de valor corte.

7. El método de la reivindicación 2, comprende además:

alinear la pluralidad de moléculas de ácido nucleico de muestra a una pluralidad de regiones genómicas, para cada región genómica de la pluralidad de regiones genómicas:

determinar que un número de moléculas de ácido nucleico de muestra alineado con la región genómica es mayor que un número de valor de corte.

8. El método de la reivindicación 1, en el que:

la ventana de nucleótidos que corresponde a la estructura de datos de entrada comprende nucleótidos sobre una primera hebra de la molécula de ácido nucleico de muestra y nucleótidos sobre una segunda hebra de la molécula de ácido nucleico de muestra, y la estructura de datos de entrada comprende además para cada nucleótido dentro de la ventana un valor de una propiedad de hebra, la propiedad de hebra indica que el nucleótido está presente sobre ya sea la primera hebra o la segunda hebra.

9. El método de la reivindicación 8, en el que la molécula de ácido nucleico de muestra es una molécula de ADN circular formada al:

cortar una molécula de ADN de doble hebra utilizando un complejo Cas9 para formar una molécula de ADN de doble hebra cortada, y

ligar un adaptador de horquilla en un extremo de la molécula de ADN de doble hebra cortada.

10. El método de la reivindicación 1, en el que los nucleótidos dentro de la ventana se determinan utilizando una secuencia de consenso circular y se determinan sin alineación de los nucleótidos secuenciados con un genoma de referencia.

11. El método de la reivindicación 1, el método comprende además:

recibir la primera pluralidad de primeras estructuras de datos;

almacenar la pluralidad de primeras muestras de entrenamiento; y

entrenar el modelo utilizando la pluralidad de primeras muestras de entrenamiento al optimizar parámetros del modelo basado en las salidas del modelo que coincide o no coincide con las etiquetas correspondientes de las primeras etiquetas cuando la primera pluralidad de primeras estructuras de datos es la entrada al modelo.

12. El método de la reivindicación 11, que comprende además:

recibir una segunda pluralidad de segundas estructuras de datos, cada segunda estructura de datos de la segunda pluralidad de segundas estructuras de datos que corresponden a una ventana respectiva de nucleótidos secuenciados en una molécula de ácido nucleico respectiva de una pluralidad de segundas moléculas de ácido nucleico, en la que la metilación tiene un segundo estado conocido en un nucleótido en una posición diana dentro de cada ventana de cada segunda molécula de ácido nucleico, cada segunda estructura de datos comprende valores para las mismas propiedades como la primera pluralidad de primeras estructuras de datos;

almacenar una pluralidad de segundas muestras de entrenamiento, cada una incluye una de la segunda pluralidad de segundas estructuras de datos y una segunda etiqueta que indica el segundo estado del nucleótido en la posición diana;

en el que:

el primer estado o el segundo estado es aquel en el que está presente la metilación y el otro estado es aquel en el que está ausente la metilación, y

entrenar el modelo además comprende utilizar la pluralidad de segundas muestras de entrenamiento el optimizar los parámetros del modelo basado en las salidas del modelo que coinciden o no coinciden con las etiquetas correspondientes de las segundas etiquetas cuando la segunda pluralidad de segundas estructuras de datos son entradas al modelo.

13. El método de la reivindicación 1 o 11, en el que cada ventana asociada con la primera pluralidad de primeras estructuras de datos comprende al menos 4 nucleótidos consecutivos sobre una primera hebra de cada primer molécula de ácido nucleico.

14. El método de la reivindicación 11, en el que el modelo comprende una red neuronal convolucional que comprende:

un conjunto de filtros convolucionales configurados para filtrar la primera pluralidad de primeras estructuras de datos,

una capa de entrada configurada para recibir la primera pluralidad filtrada de primeras estructuras de datos, una pluralidad de capas ocultas que incluyen una pluralidad de nodos, una primera capa de la pluralidad de capas ocultas acopladas a la capa de entrada; y

una capa de salida acoplada a una última capa de la pluralidad de capas ocultas y configurada para emitir una estructura de datos de salida, la estructura de datos de salida que comprende las propiedades.

15. El método de la reivindicación 1 o 11, en el que los primeros estados incluyen un estado metilado para una primera porción de las primeras estructuras de datos y un estado no metilado para una segunda porción de las primeras estructuras de datos, y en la que la metilación comprende 4mC (N4-metilcitosina), 5mC (5-metilcitosina), 5hmC (5-hidroximetilcitosina), 5fC (5-formilcitosina), 5caC (5-carboxilcitosina), 1mA (N1-metiladenina), 3mA (N3-metiladenina), 6mA (N6-metiladenina), 7mA (N7-metiladenina), 3mC (N3-metilcitosina), 2mG (N2-metilguanina), 6mG (O6-metilguanina), 7mG (N7-metilguanina), 3mT (N3-metiltimina), o 4mT (O4-metiltimina).

16. El método de la reivindicación 1 o 11, en el que la metilación comprende 5mC (5-metilcitosina).

17. El método de la reivindicación 1 o 11, en el que la metilación comprende 6mA (N6-metiladenina).

18. El método de la reivindicación 1 o 11, en el que cada estructura de datos comprende además un valor para una altura del pulso que corresponde a cada nucleótido dentro de la ventana.

19. El método de la reivindicación 11, en el que:

cada primera estructura de datos de la primera pluralidad de primeras estructuras de datos excluye primeras moléculas de ácido nucleico con una duración interpulso o una anchura de un pulso por debajo de un valor de corte.

20. El método de la reivindicación 11, en el que:

la pluralidad de primeras muestras de entrenamiento se genera al:

amplificar una pluralidad de moléculas de ácido nucleico utilizando un conjunto de nucleótidos, en el que el conjunto de nucleótidos incluye 6mA a una relación especificada.

21. Un producto informático que comprende un medio no transitorio legible por ordenador que almacena una pluralidad de instrucciones que cuando se ejecutan controlan un sistema informático para realizar el método de una cualquiera de las reivindicaciones anteriores.

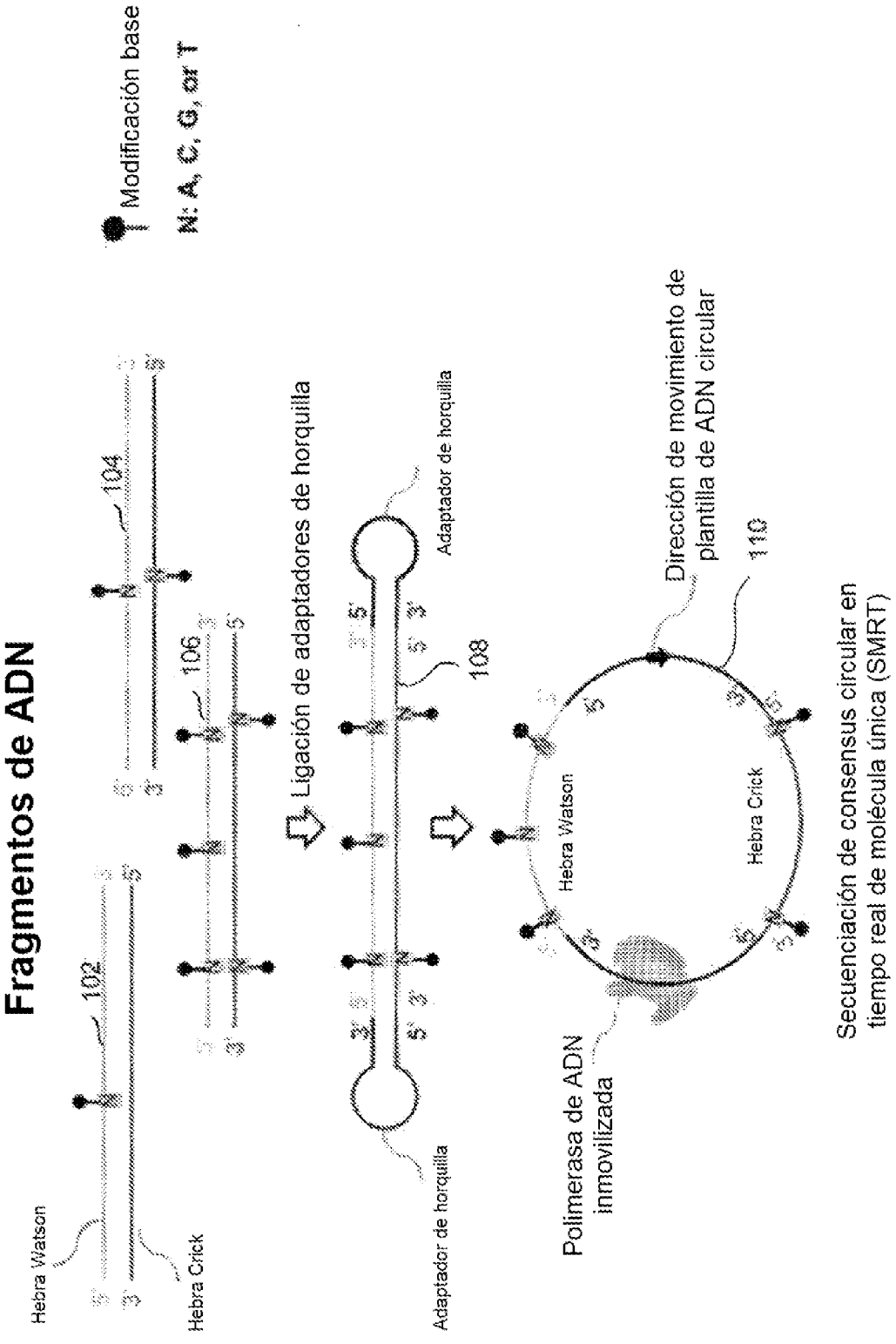

Fragmentos de ADN
Hebra Watson
102
Hebra Crick
104
106
Modificación base
N: A, C, G, or T
Ligación de adaptadores de horquilla
Adaptador de horquilla
Adaptador de horquilla
108
Polimerasa de ADN inmovilizada
Hebra Watson
Hebra Crick
Dirección de movimiento de plantilla de ADN circular
110
Secuenciación de consensus circular en tiempo real de molécula única (SMRT)

FIG. 1

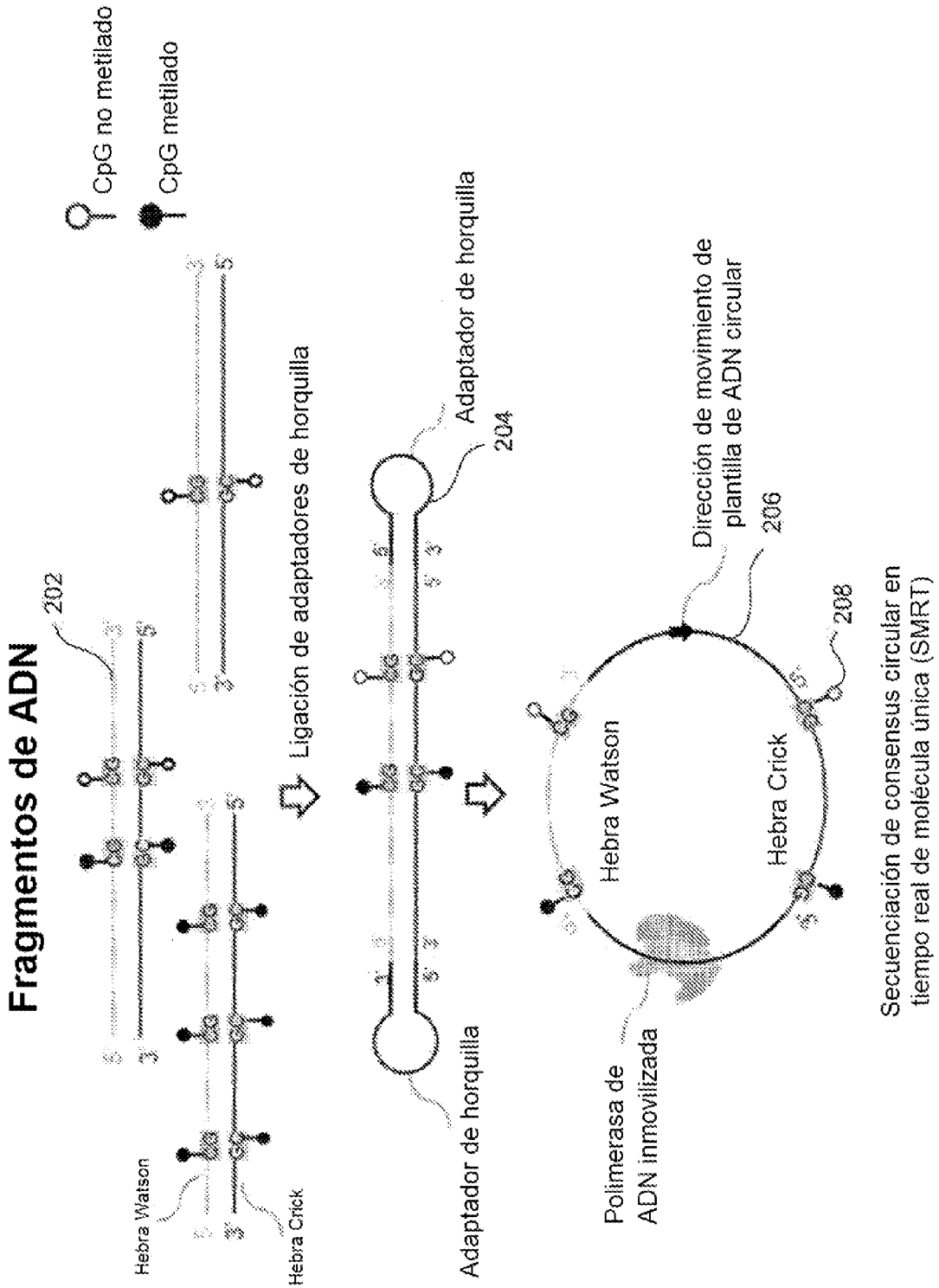

Fragmentos de ADN
202
CpG no metilado
CpG metilado
Hebra Watson
Hebra Crick
Ligación de adaptadores de horquilla
Adaptador de horquilla
204
Adaptador de horquilla
Polimerasa de ADN inmovilizada
Hebra Watson
Hebra Crick
Dirección de movimiento de plantilla de ADN circular
206
208
Secuenciación de consensus circular en tiempo real de molécula única (SMRT)

FIG. 2

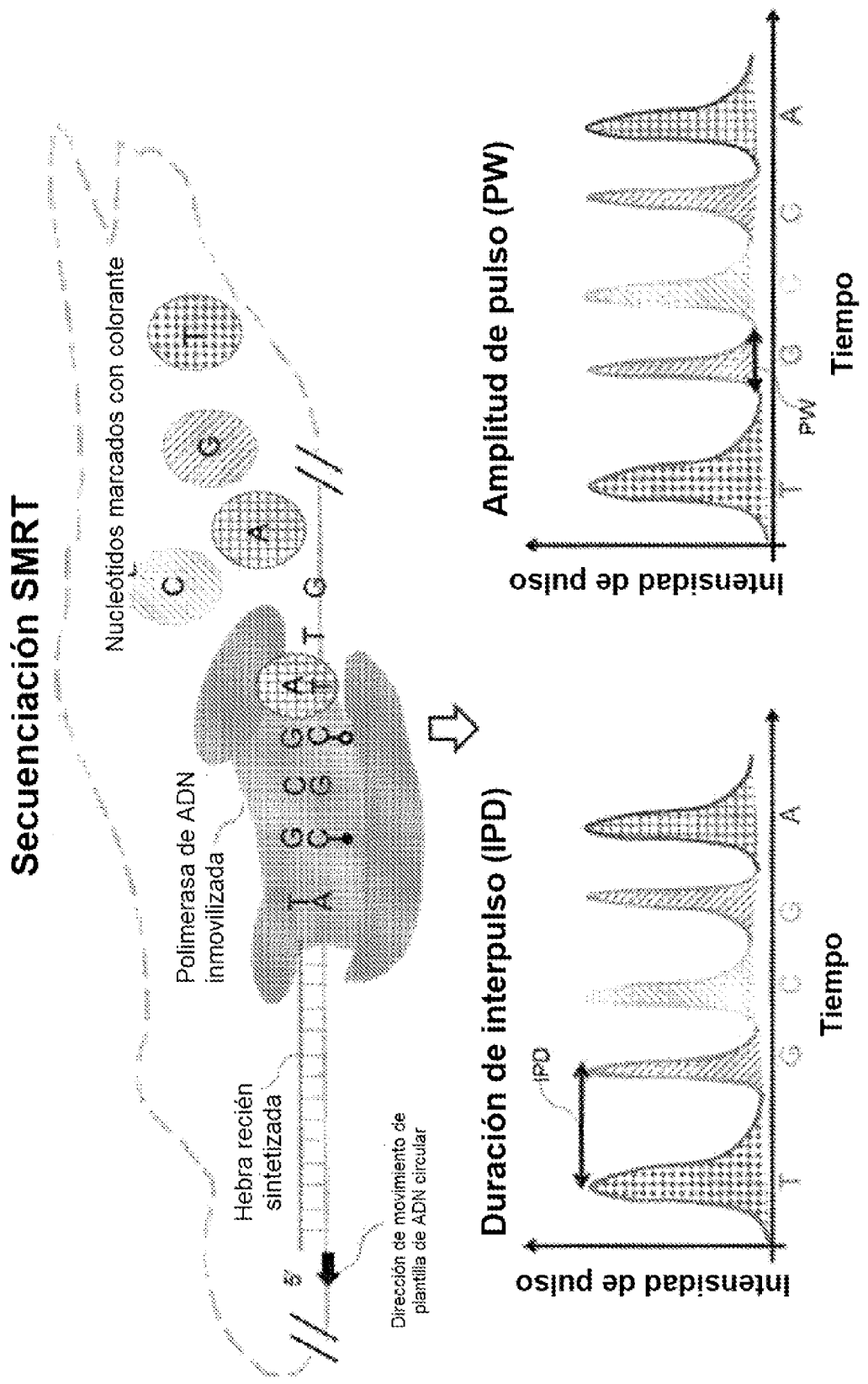

Secuenciación SMRT
Nucleótidos marcados con colorante
Polimerasa de ADN inmovilizada
Hebra recién sintetizada
Dirección de movimiento de plantilla de ADN circular
Duración de interpulso (IPD)
Intensidad de pulso
Tiempo
IPD
Amplitud de pulso (PW)
Intensidad de pulso
Tiempo
PW

FIG. 3

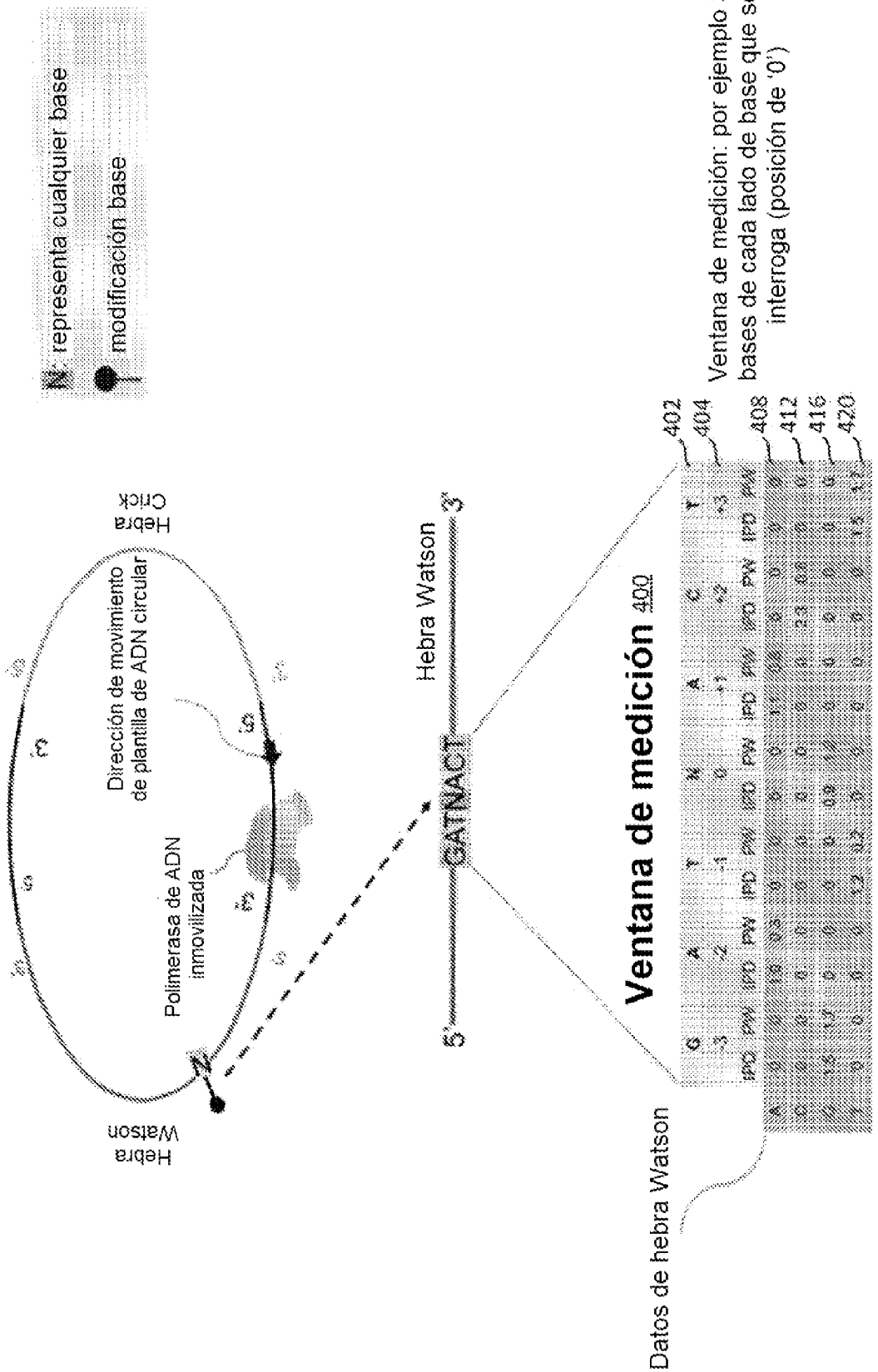

N representa cualquier base
modificación base
Hebra Watson
Hebra Crick
Dirección de movimiento de plantilla de ADN circular
Polimerasa de ADN inmovilizada
Hebra Watson
5'
3'
GATNACT
Ventana de medición 400
Ventana de medición: por ejemplo 3 bases de cada lado de base que se interroga (posición de '0')
Datos de hebra Watson
402
404
408
412
416
420

FIG. 4

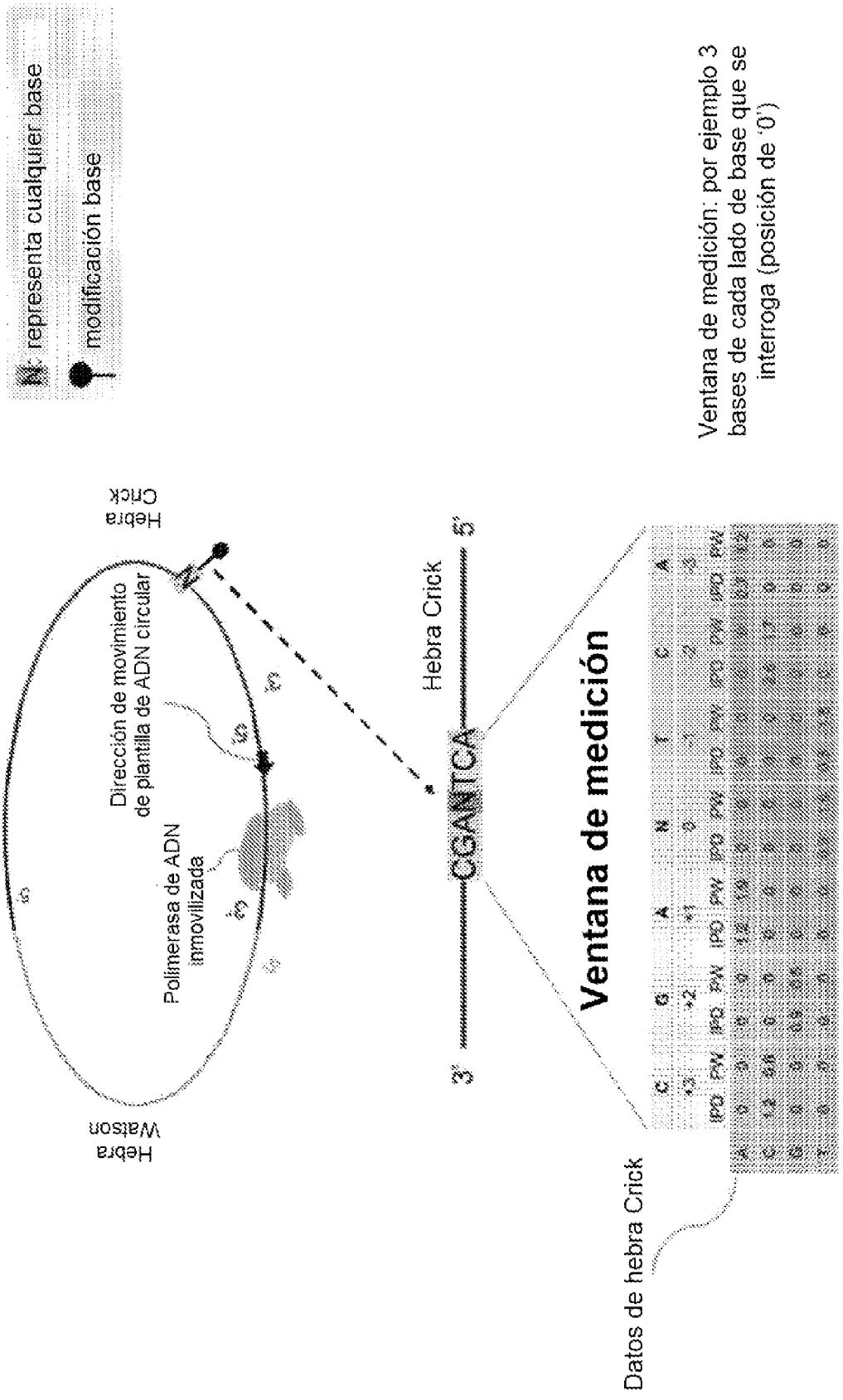

N representa cualquier base
modificación base
Ventana de medición: por ejemplo 3 bases de cada lado de base que se interroga (posición de '0')
Hebra Crick
Dirección de movimiento de plantilla de ADN circular
Polimerasa de ADN inmovilizada
Hebra Watson
Hebra Crick
5'
3'
CGANTCA
Ventana de medición
Datos de hebra Crick

FIG. 5

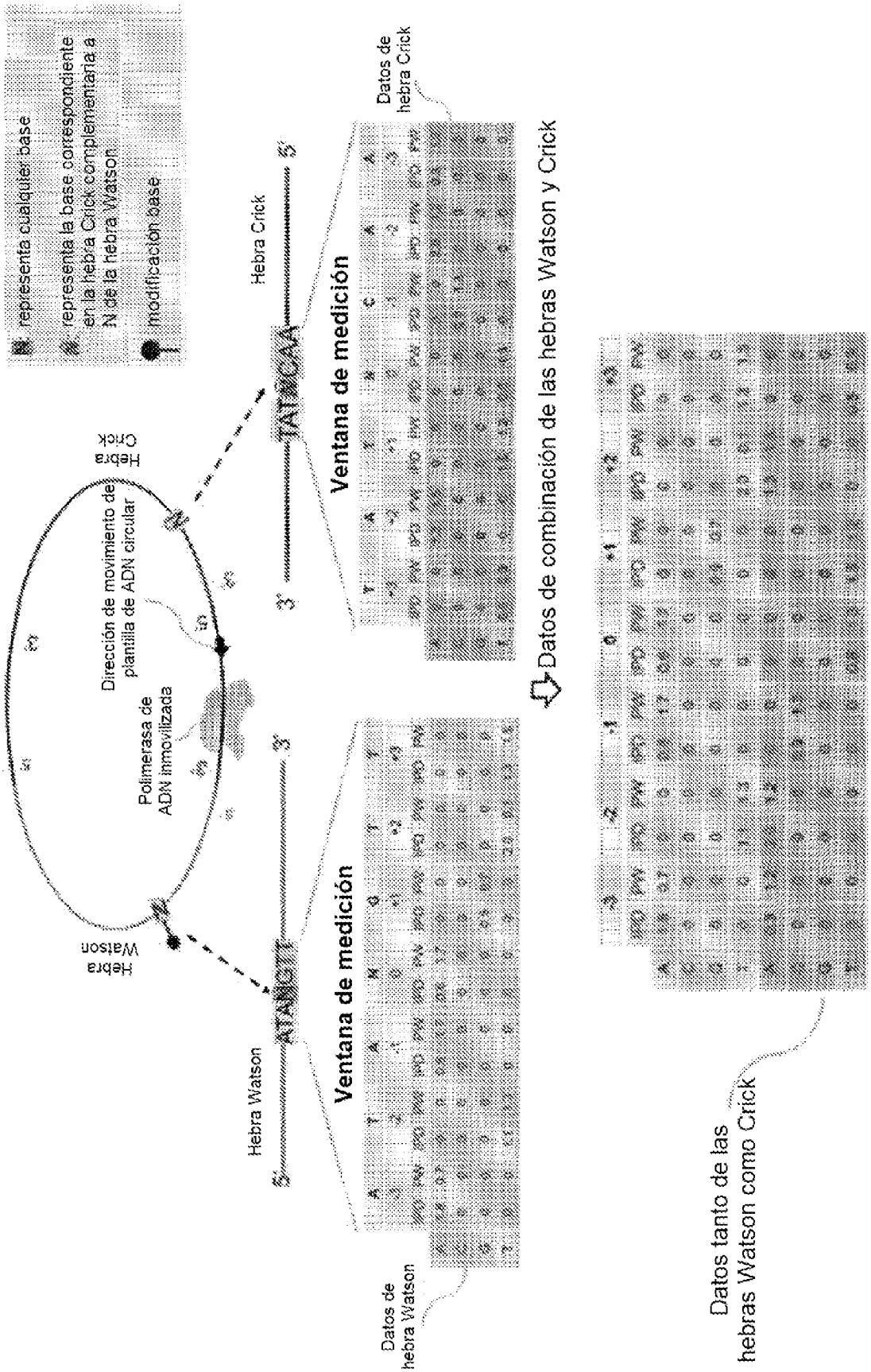

N representa cualquier base
representa la base correspondiente en la hebra Crick complementaria a N de la hebra Watson
modificación base
Hebra Crick
Dirección de movimiento de plantilla de ADN circular
Polimerasa de ADN inmovilizada
Hebra Watson
Hebra Watson
5'
ATAGCTT
3'
Ventana de medición
Datos de hebra Watson
Hebra Crick
5'
3'
TATGCAA
Ventana de medición
Datos de hebra Crick
Datos de combinación de las hebras Watson y Crick
Datos tanto de las hebras Watson como Crick

FIG. 6

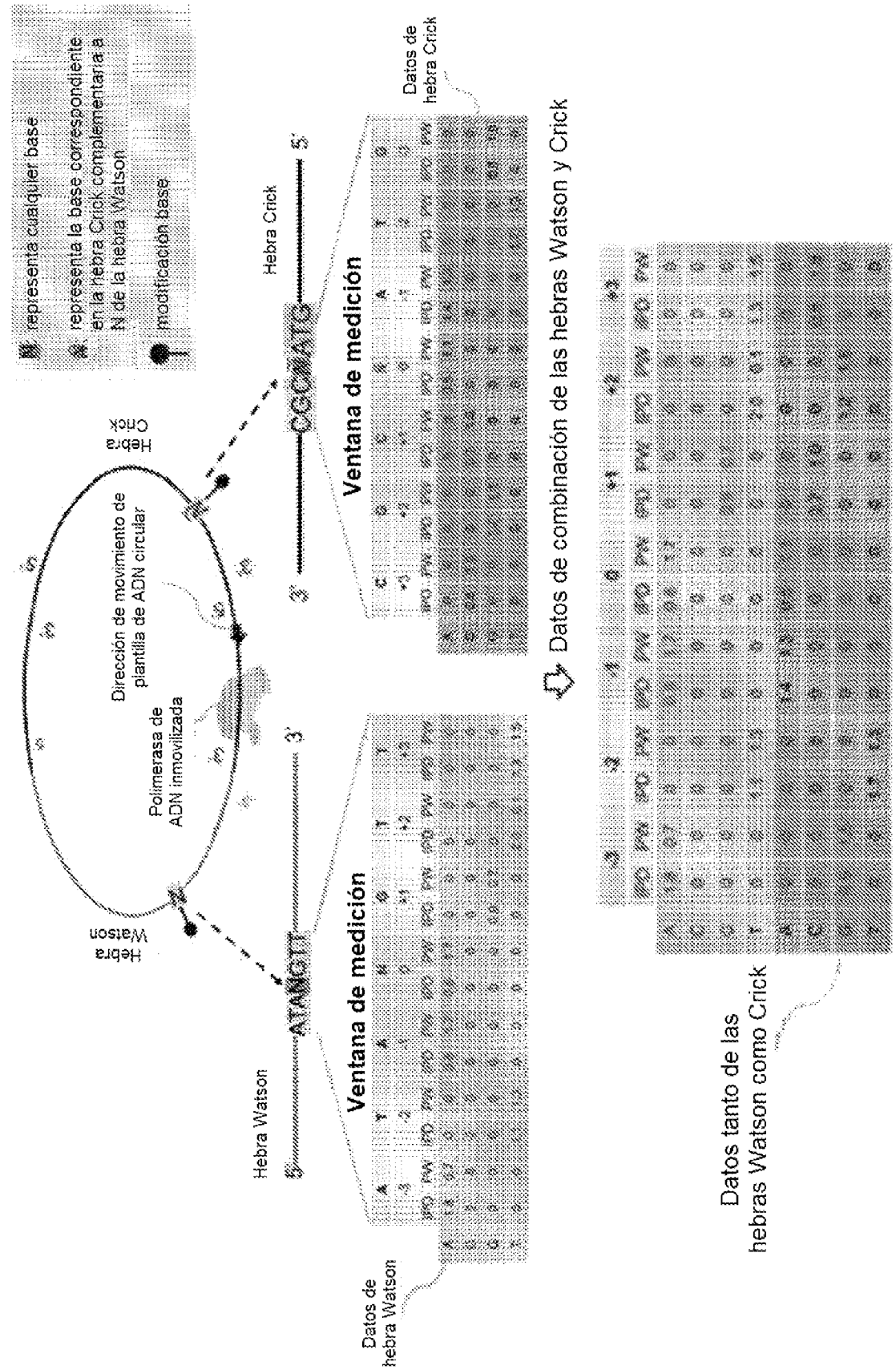

representa cualquier base
representa la base correspondiente en la hebra Crick complementaria a N de la hebra Watson
modificación base
Hebra Crick
Hebra Watson
Dirección de movimiento de plantilla de ADN circular
Polimerasa de ADN inmovilizada
5'
3'
ATAMGTT
CGCAMATG
Ventana de medición
Ventana de medición
Datos de hebra Watson
Datos de hebra Crick
Datos de combinación de las hebras Watson y Crick
Datos tanto de las hebras Watson como Crick

FIG. 7

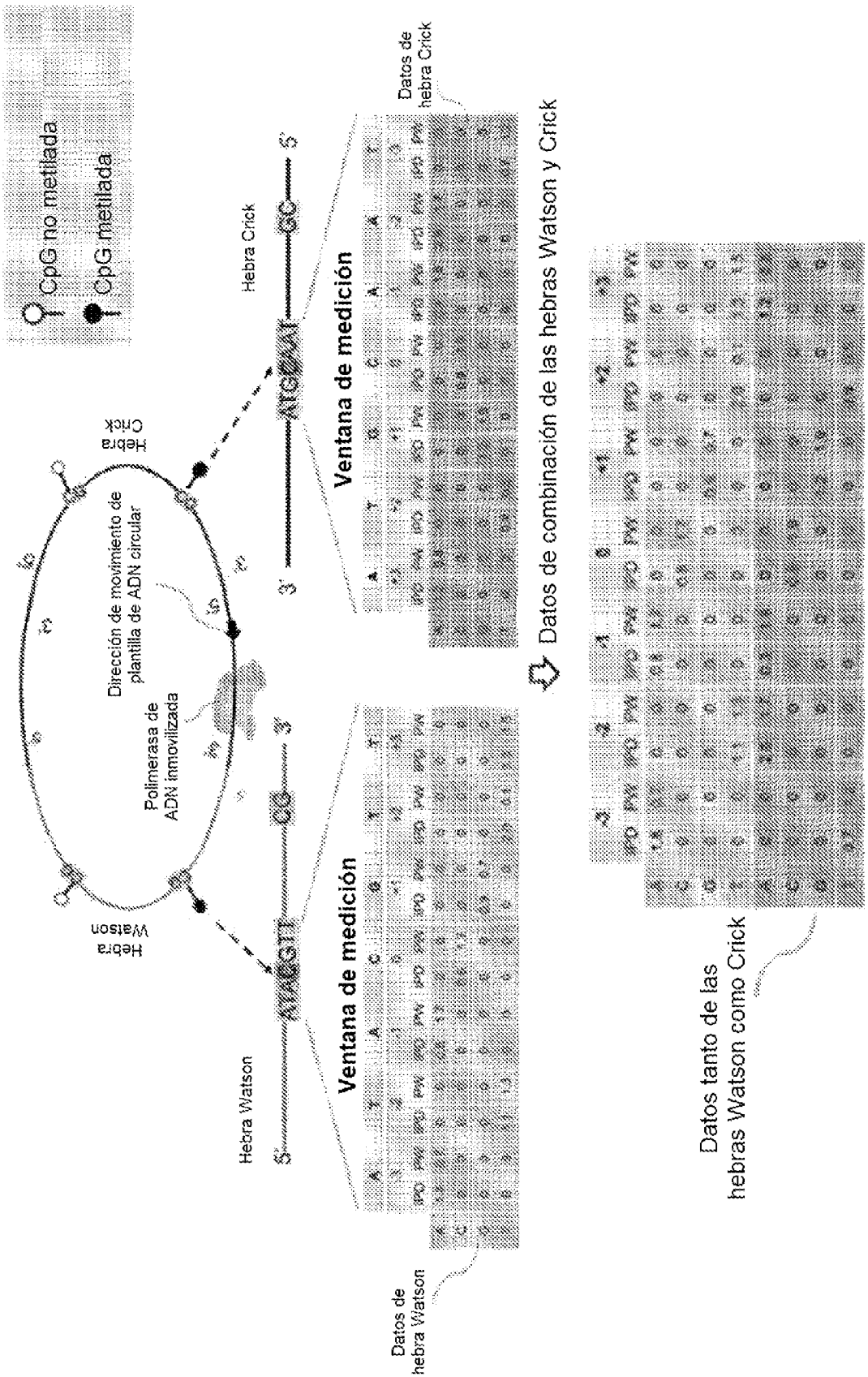

CpG no metilada
CpG metilada
Hebra Watson
Hebra Crick
Dirección de movimiento de plantilla de ADN circular
Polimerasa de ADN inmovilizada
Hebra Watson
Hebra Crick
Ventana de medición
Ventana de medición
Datos de hebra Watson
Datos de hebra Crick
Datos de combinación de las hebras Watson y Crick
Datos tanto de las hebras Watson como Crick

FIG. 8

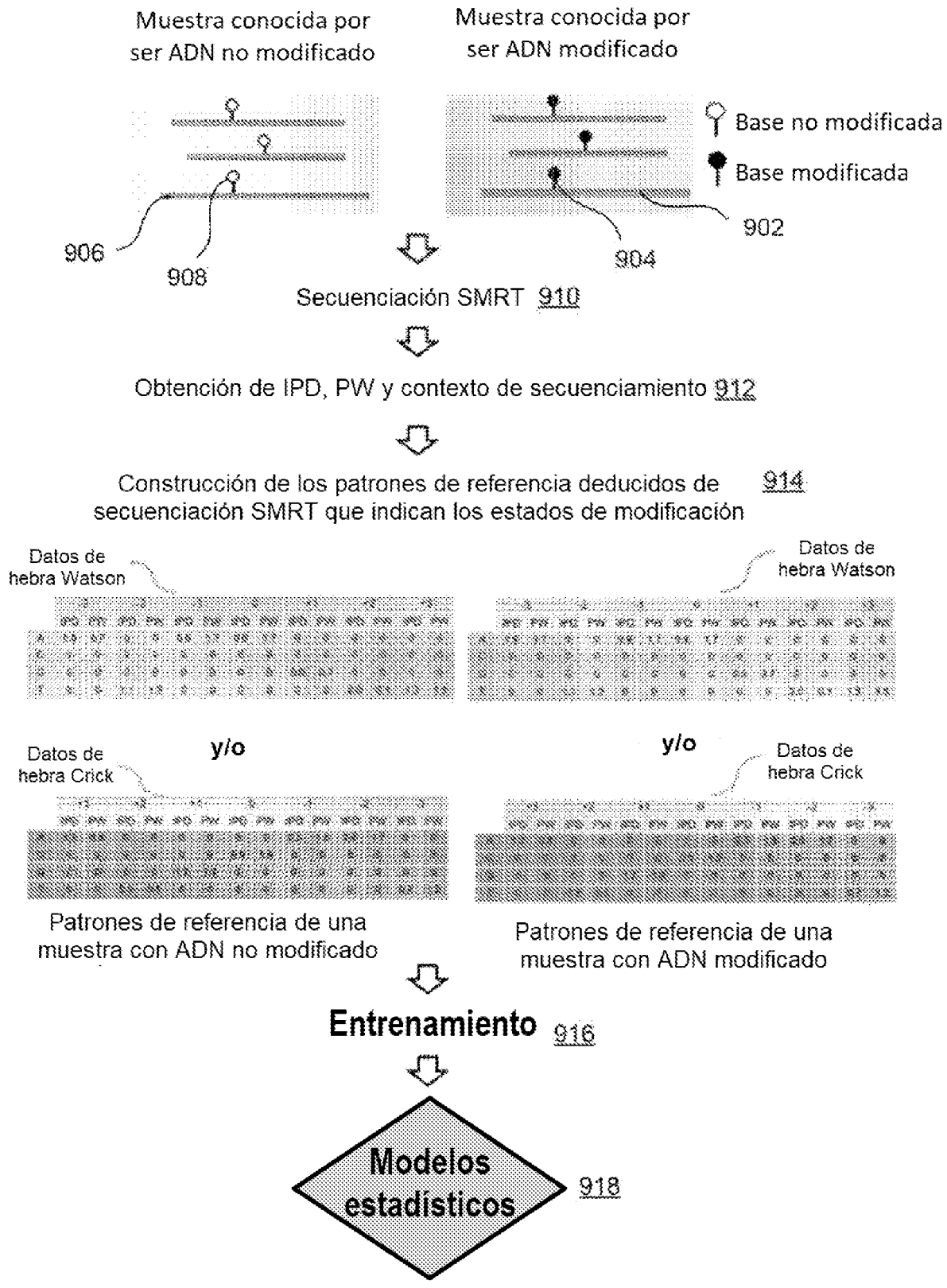

Muestra conocida por ser ADN no modificado
Muestra conocida por ser ADN modificado
Base no modificada
Base modificada
906
908
902
904
Secuenciación SMRT 910
Obtención de IPD, PW y contexto de secuenciamiento 912
Construcción de los patrones de referencia deducidos de secuenciación SMRT que indican los estados de modificación 914
Datos de hebra Watson
Datos de hebra Watson
y/o
Datos de hebra Crick
y/o
Datos de hebra Crick
Patrones de referencia de una muestra con ADN no modificado
Patrones de referencia de una muestra con ADN modificado
Entrenamiento 916
Modelos estadísticos
918

FIG. 9

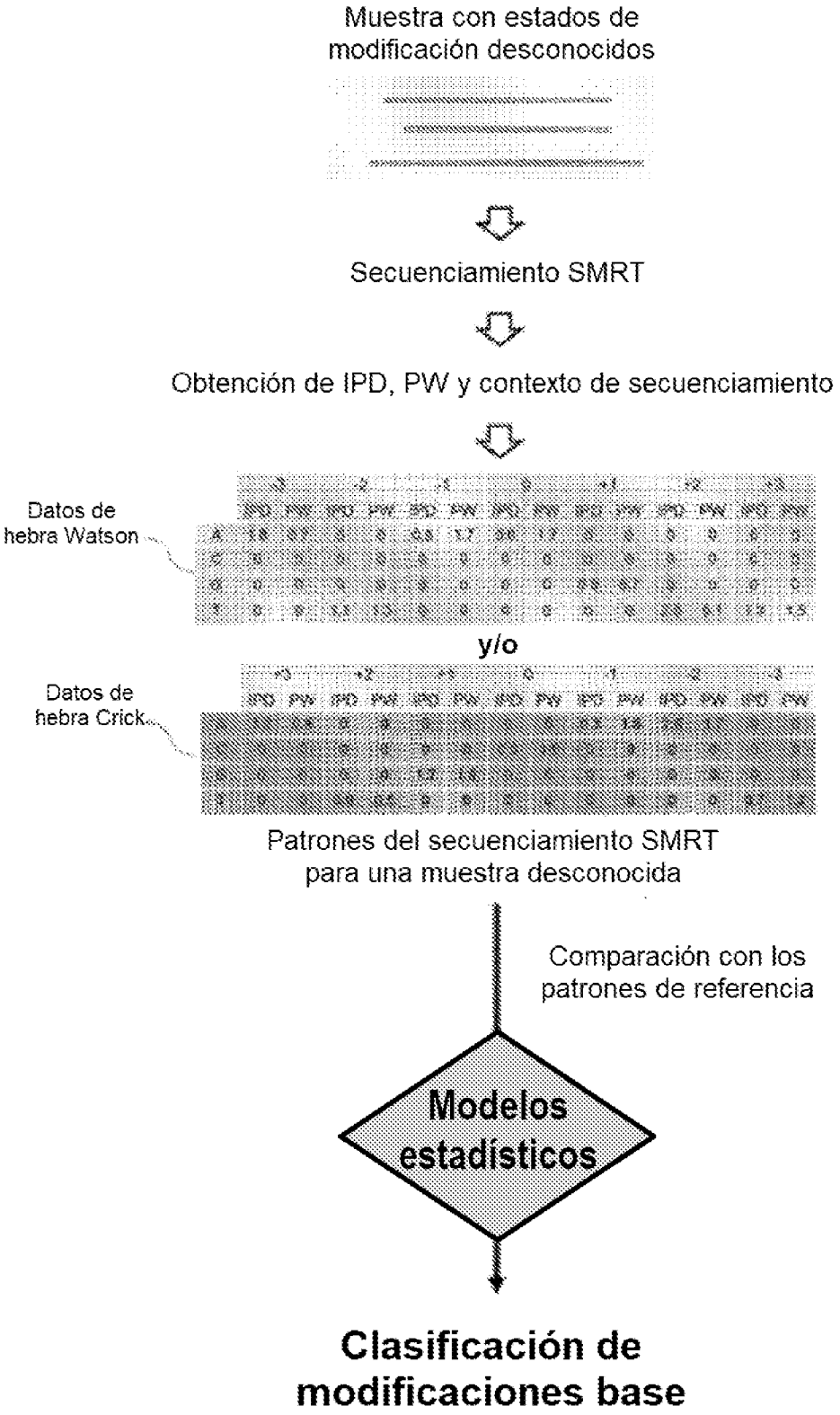

Muestra con estados de modificación desconocidos
Secuenciamiento SMRT
Obtención de IPD, PW y contexto de secuenciamiento
Datos de hebra Watson
y/o
Datos de hebra Crick
Patrones del secuenciamiento SMRT para una muestra desconocida
Comparación con los patrones de referencia
Modelos estadísticos
Clasificación de modificaciones base

FIG. 10

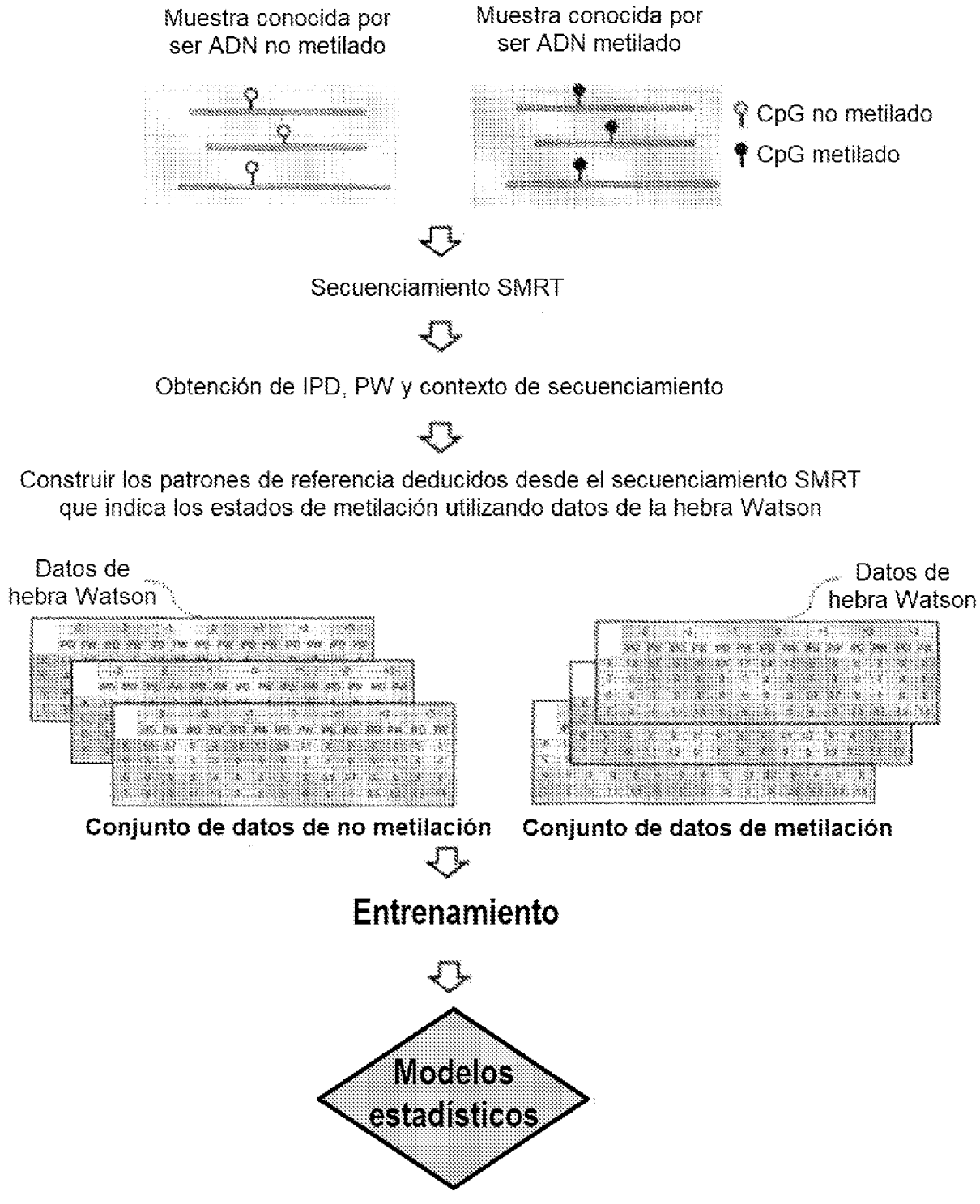

Muestra conocida por ser ADN no metilado
Muestra conocida por ser ADN metilado
CpG no metilado
CpG metilado
Secuenciamiento SMRT
Obtención de IPD, PW y contexto de secuenciamiento
Construir los patrones de referencia deducidos desde el secuenciamiento SMRT que indica los estados de metilación utilizando datos de la hebra Watson
Datos de hebra Watson
Datos de hebra Watson
Conjunto de datos de no metilación
Conjunto de datos de metilación
Entrenamiento
Modelos estadísticos

FIG. 11

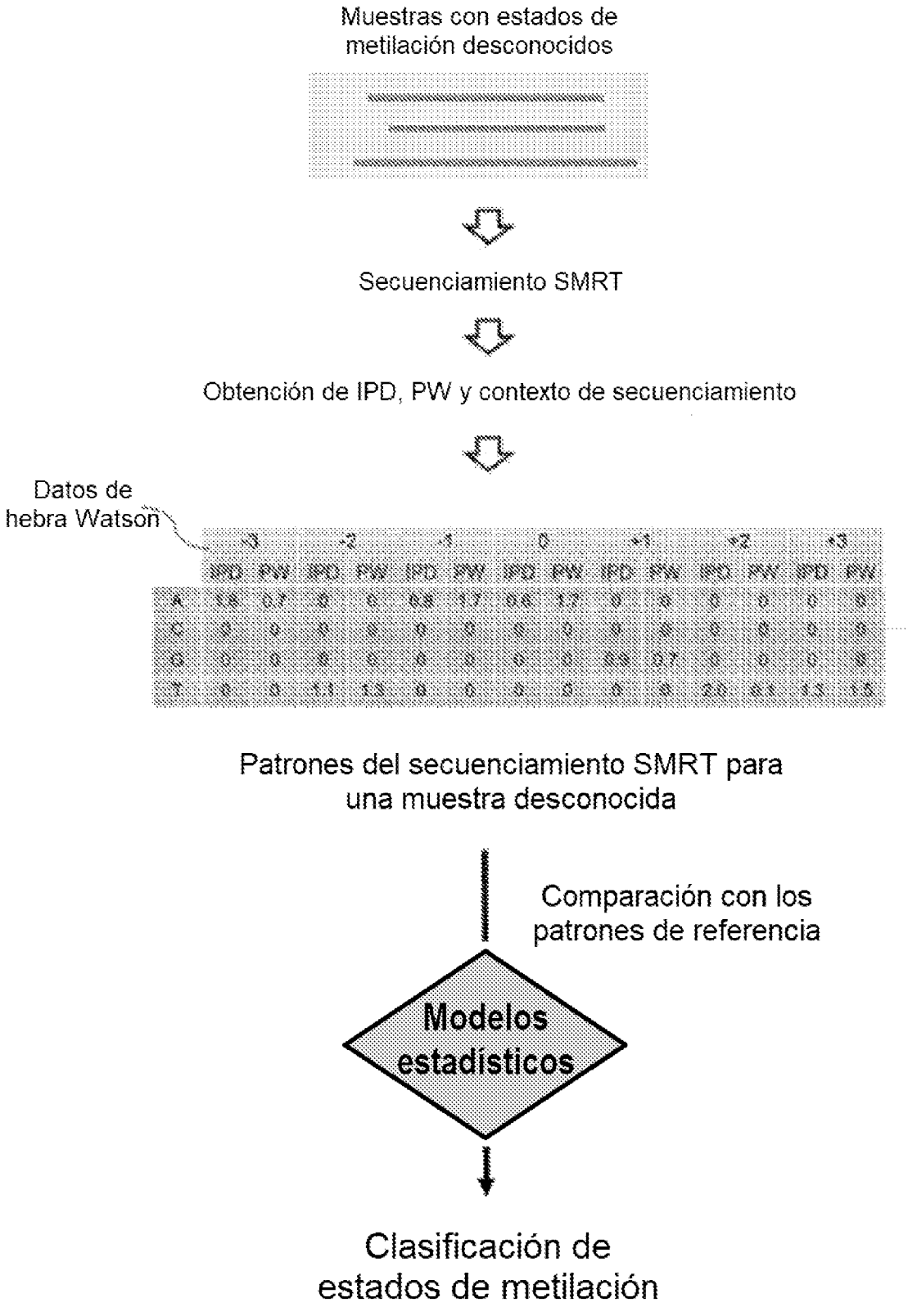

Muestras con estados de metilación desconocidos
Secuenciamiento SMRT
Obtención de IPD, PW y contexto de secuenciamiento
Datos de hebra Watson
-3
-2
-1
0
+1
+2
+3
IPD
PW
A
C
G
T
Patrones del secuenciamiento SMRT para una muestra desconocida
Comparación con los patrones de referencia
Modelos estadísticos
Clasificación de estados de metilación

FIG. 12

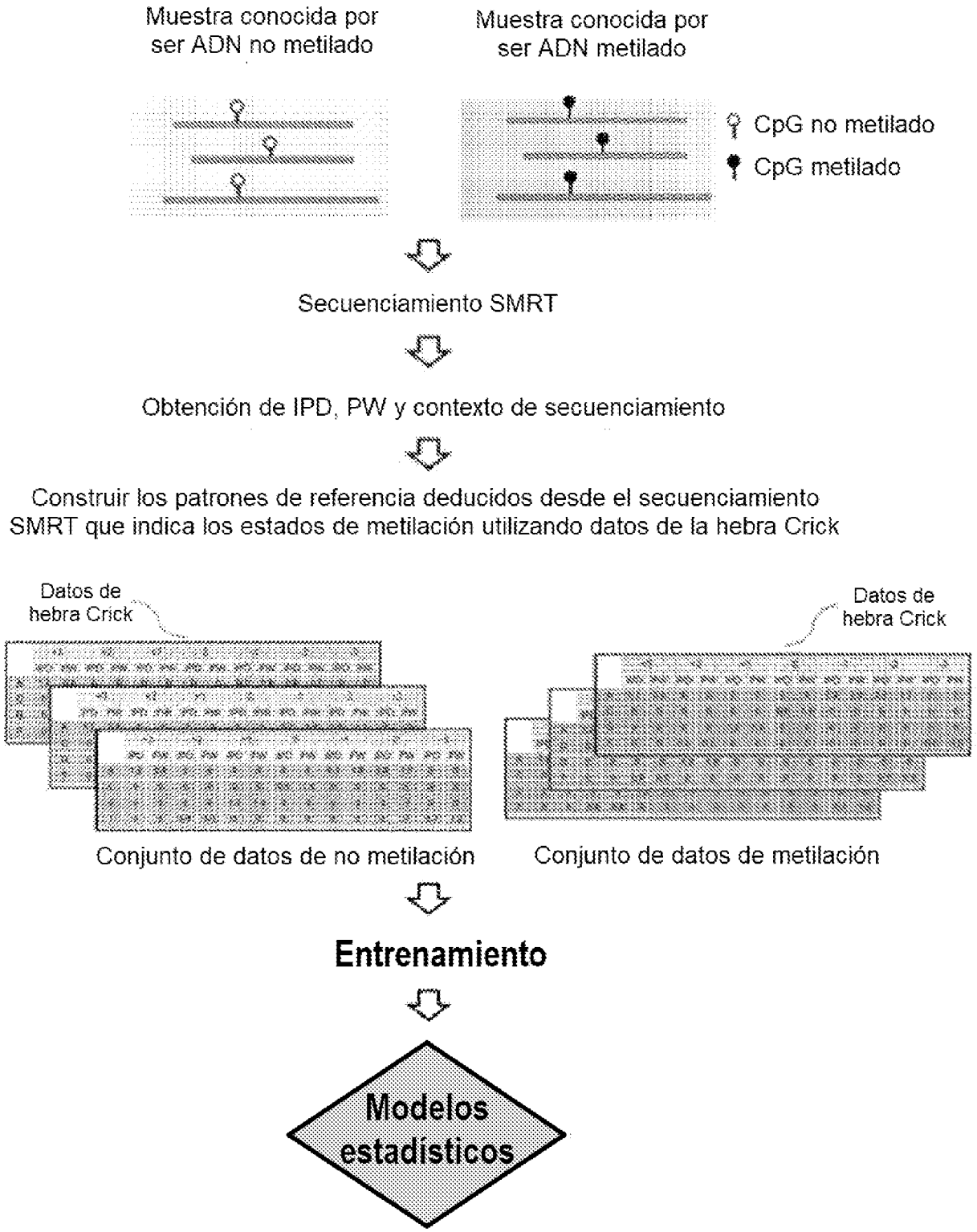

Muestra conocida por ser ADN no metilado
Muestra conocida por ser ADN metilado
CpG no metilado
CpG metilado
Secuenciamiento SMRT
Obtención de IPD, PW y contexto de secuenciamiento
Construir los patrones de referencia deducidos desde el secuenciamiento SMRT que indica los estados de metilación utilizando datos de la hebra Crick
Datos de hebra Crick
Datos de hebra Crick
Conjunto de datos de no metilación
Conjunto de datos de metilación
Entrenamiento
Modelos estadísticos

FIG. 13

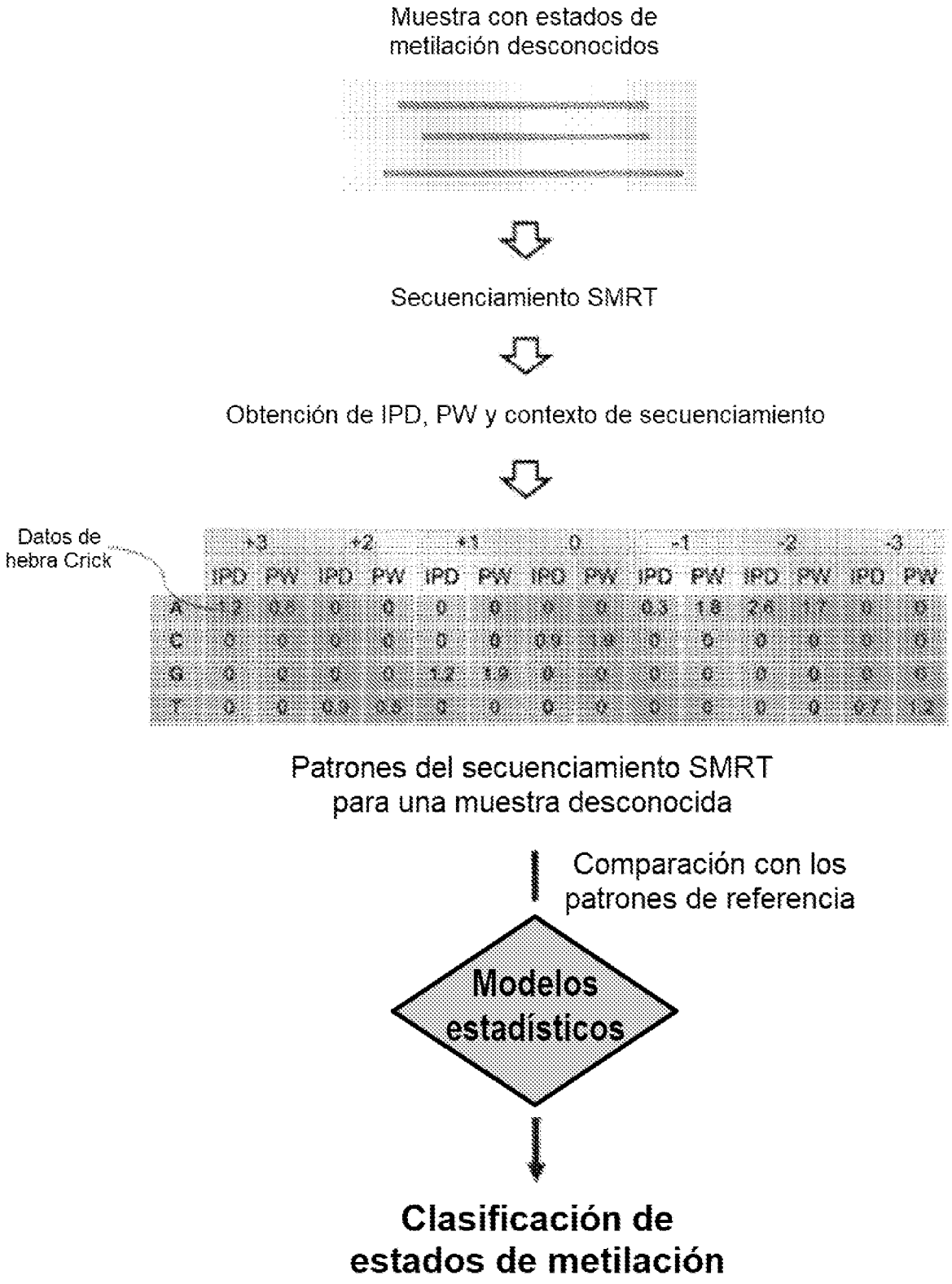

Muestra con estados de metilación desconocidos
Secuenciamiento SMRT
Obtención de IPD, PW y contexto de secuenciamiento
Datos de hebra Crick
+3
+2
+1
0
-1
-2
-3
IPD
PW
A
C
G
T
Patrones del secuenciamiento SMRT para una muestra desconocida
Comparación con los patrones de referencia
Modelos estadísticos
Clasificación de estados de metilación

FIG.14

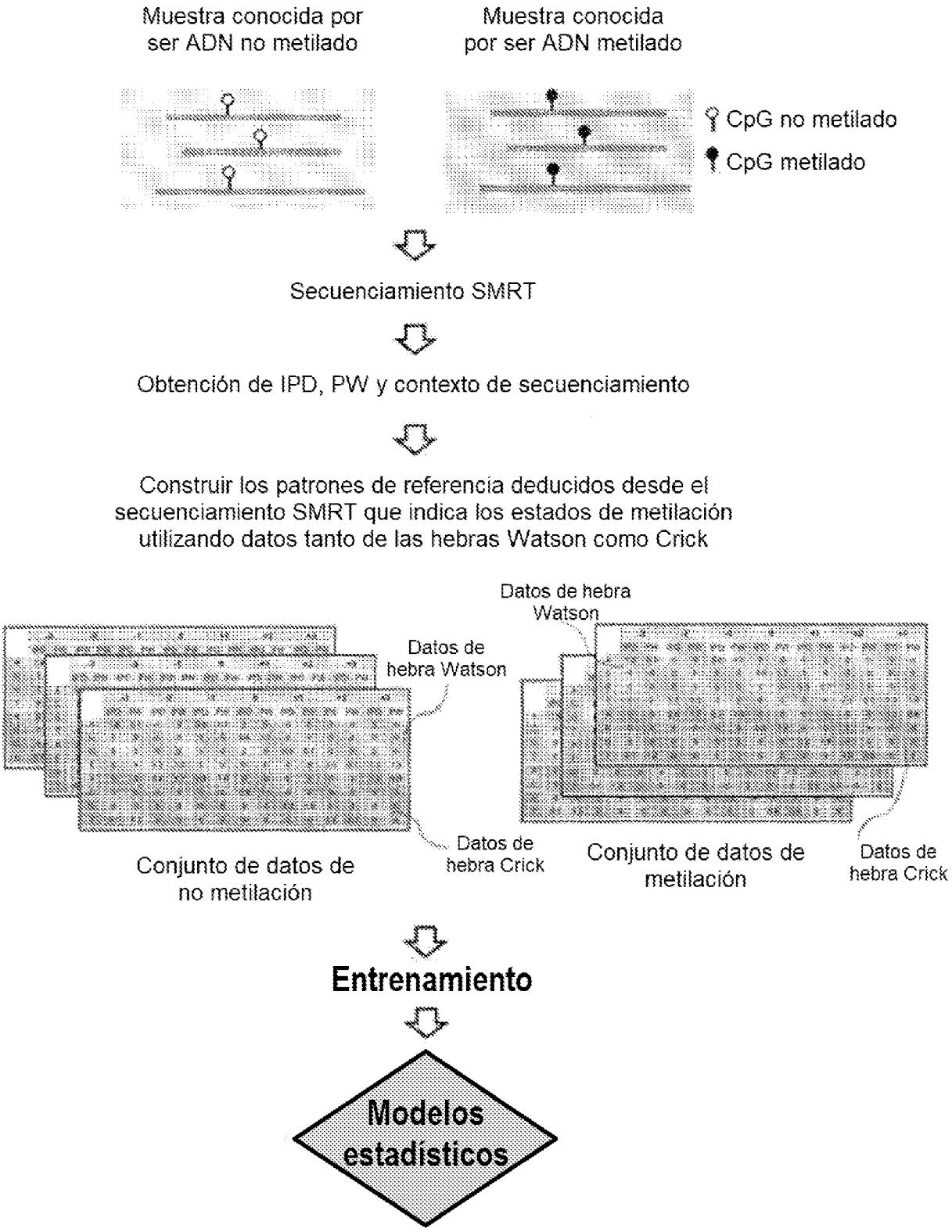

Muestra conocida por ser ADN no metilado
Muestra conocida por ser ADN metilado
CpG no metilado
CpG metilado
Secuenciamiento SMRT
Obtención de IPD, PW y contexto de secuenciamiento
Construir los patrones de referencia deducidos desde el secuenciamiento SMRT que indica los estados de metilación utilizando datos tanto de las hebras Watson como Crick
Datos de hebra Watson
Datos de hebra Watson
Datos de hebra Crick
Conjunto de datos de no metilación
Datos de hebra Crick
Conjunto de datos de metilación
Entrenamiento
Modelos estadísticos

FIG. 15

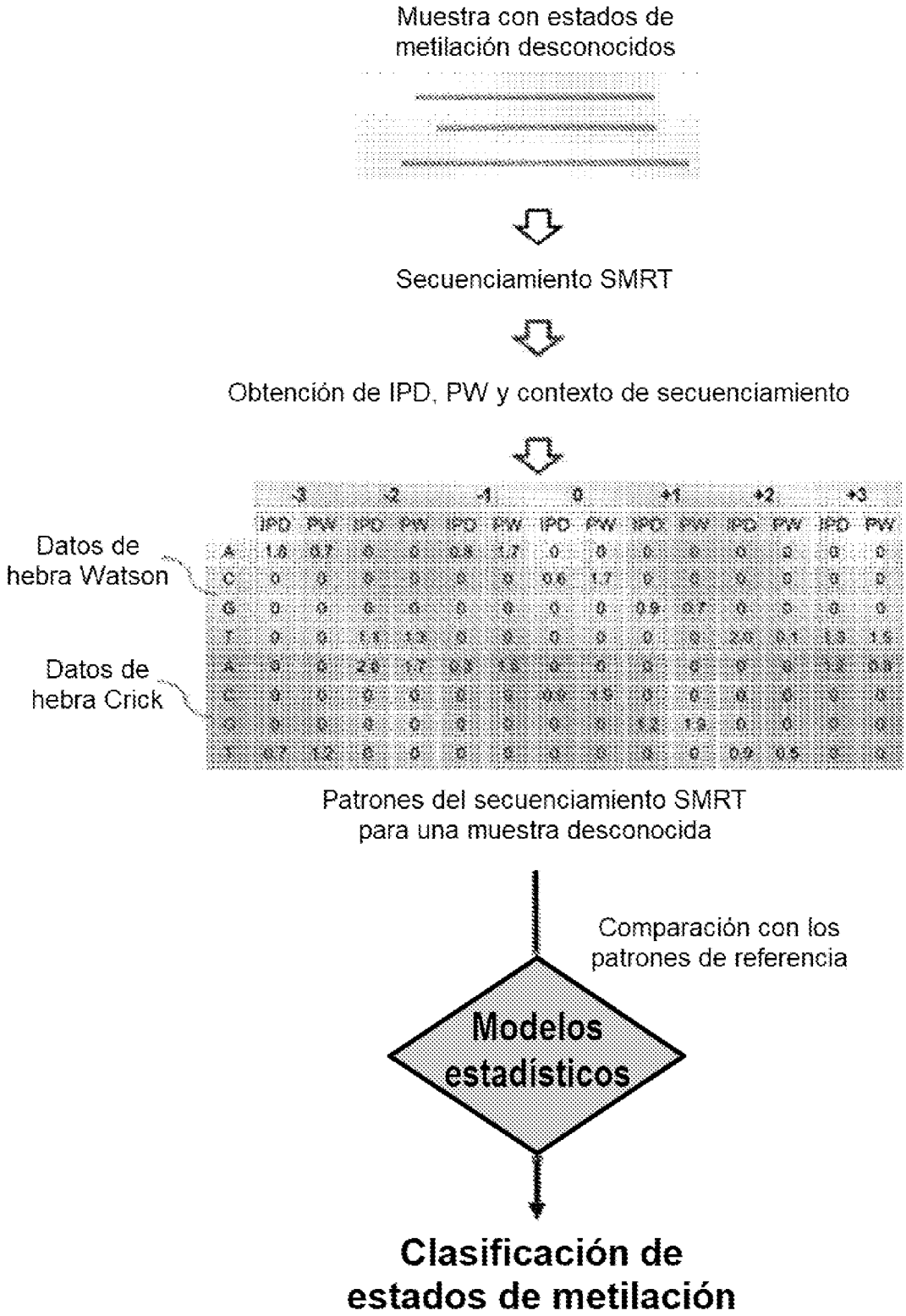

Muestra con estados de metilación desconocidos
Secuenciamiento SMRT
Obtención de IPD, PW y contexto de secuenciamiento
Datos de hebra Watson
Datos de hebra Crick
-3 -2 -1 0 +1 +2 +3
IPD PW IPD PW IPD PW IPD PW IPD PW IPD PW IPD PW
A 1.8 0.7 0 0 0.8 1.7 0 0 0 0 0 0 0 0
C 0 0 0 0 0 0 0.6 1.7 0 0 0 0 0 0
G 0 0 0 0 0 0 0 0 0.9 0.7 0 0 0 0
T 0 0 1.8 1.3 0 0 0 0 0 0 2.0 0.1 1.3 1.5
A 0 0 2.6 1.7 0.8 1.8 0 0 0 0 0 0 1.2 0.8
C 0 0 0 0 0 0 0.6 1.5 0 0 0 0 0 0
G 0 0 0 0 0 0 0 0 1.2 1.9 0 0 0 0
T 0.7 1.2 0 0 0 0 0 0 0 0 0.9 0.5 0 0
Patrones del secuenciamiento SMRT para una muestra desconocida
Comparación con los patrones de referencia
Modelos estadísticos
Clasificación de estados de metilación

FIG. 16

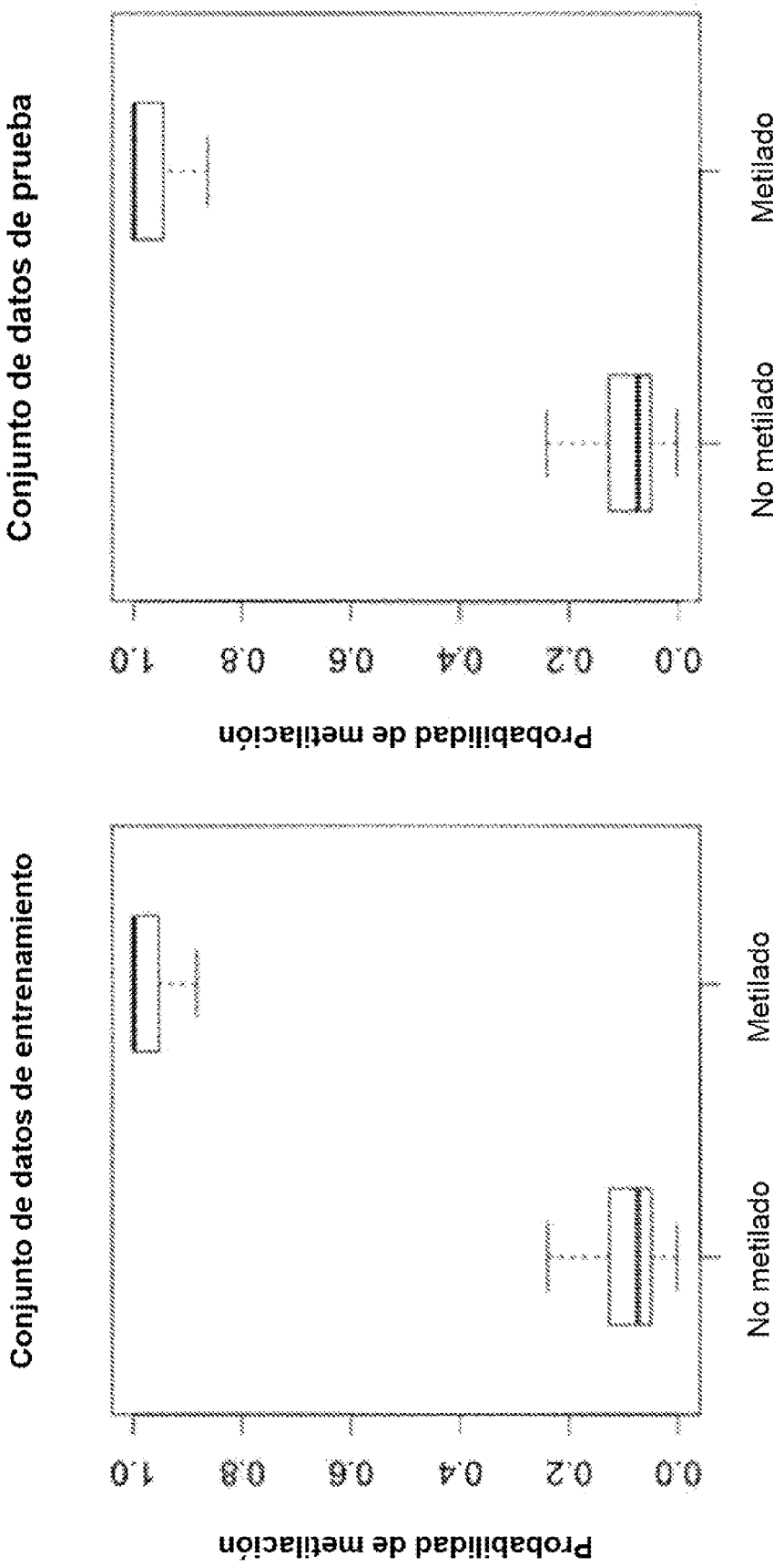


FIG. 17A

FIG. 17B

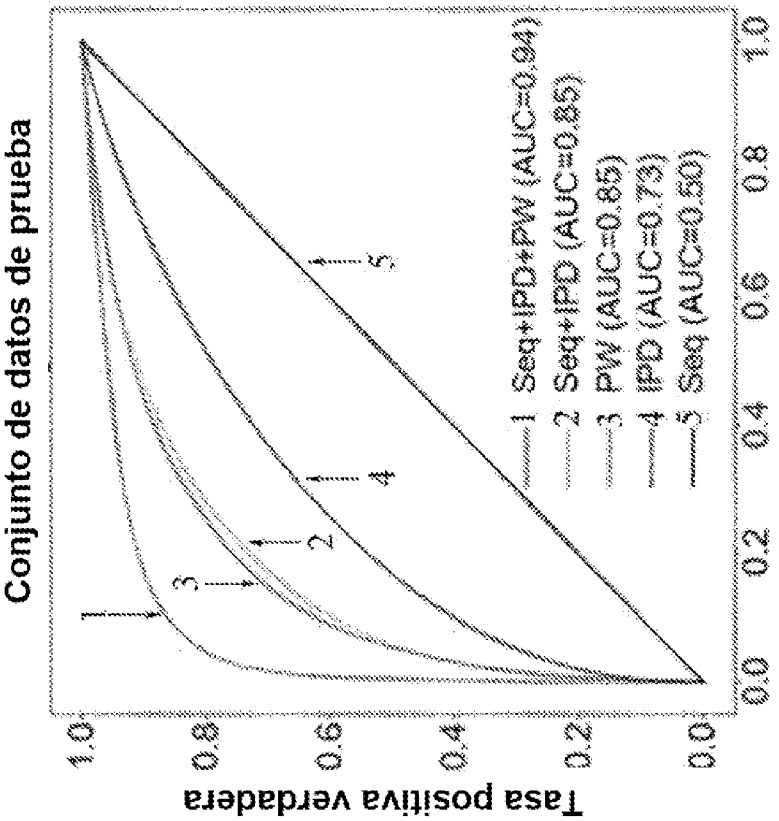


FIG. 18B

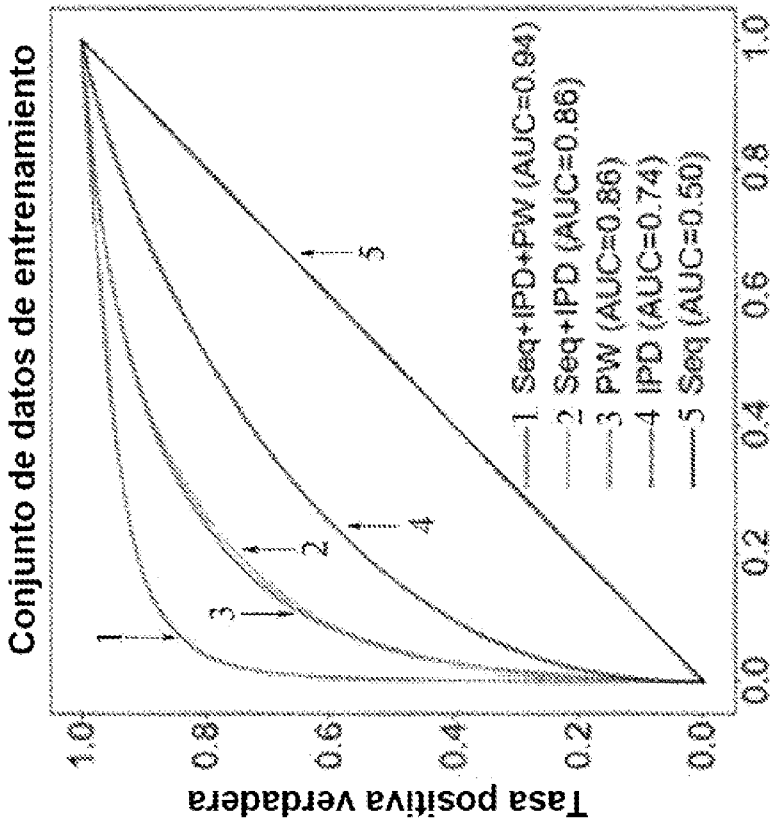


FIG. 18A

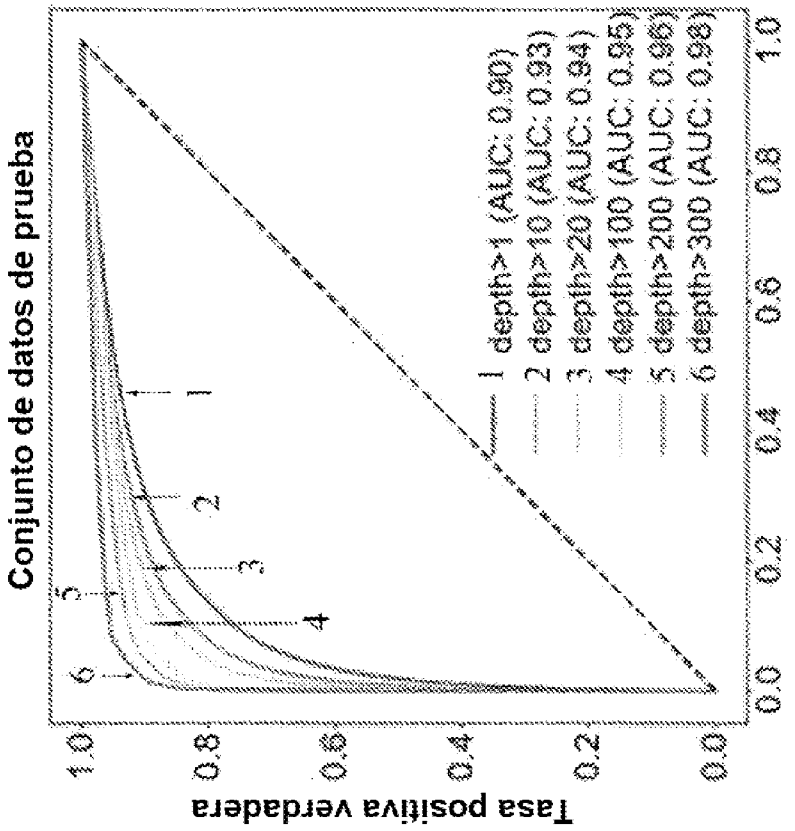


FIG. 19B

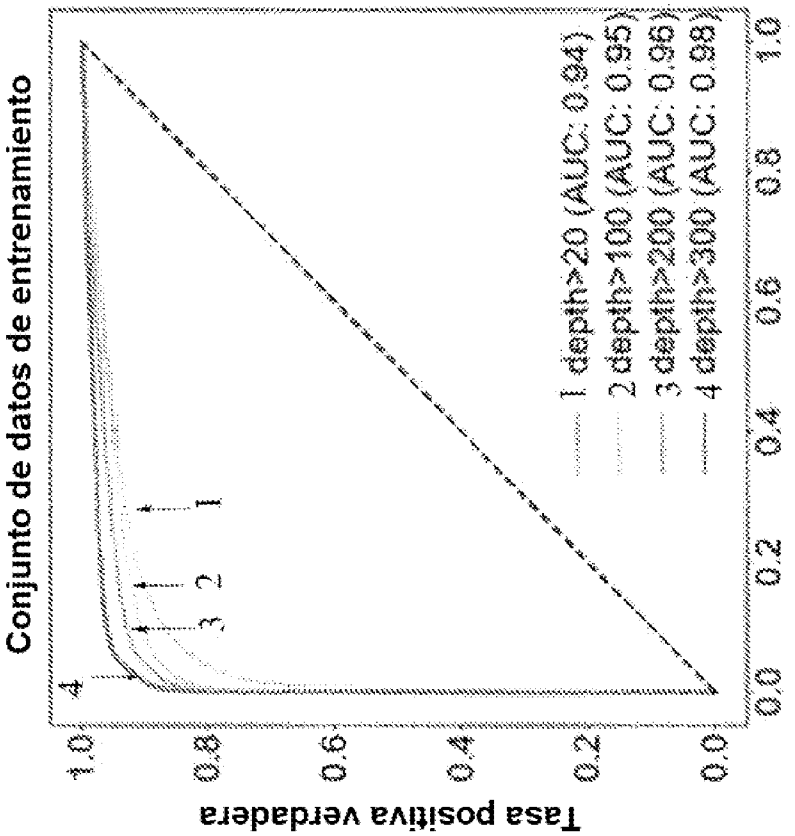


FIG. 19A

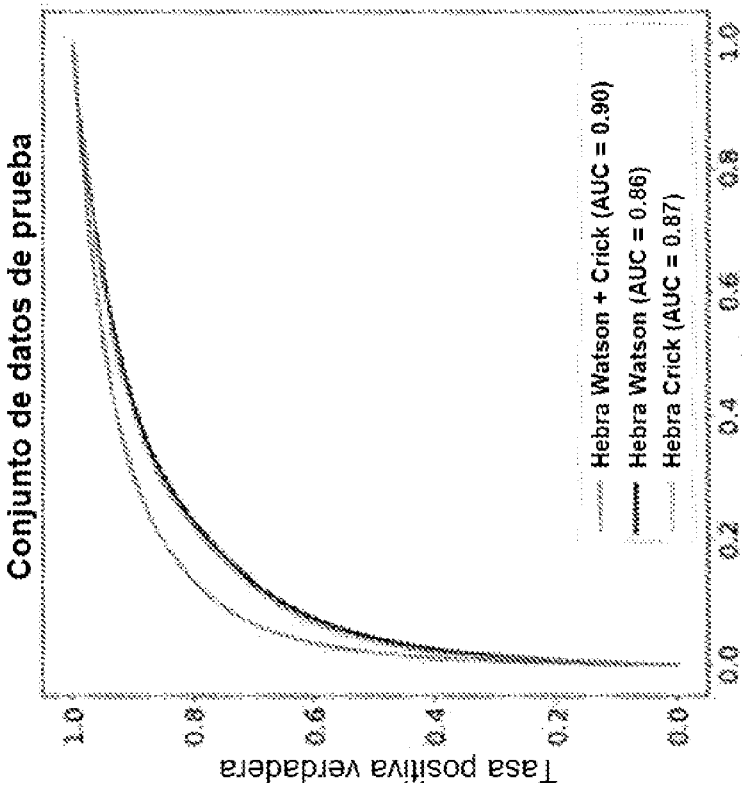


FIG. 20B

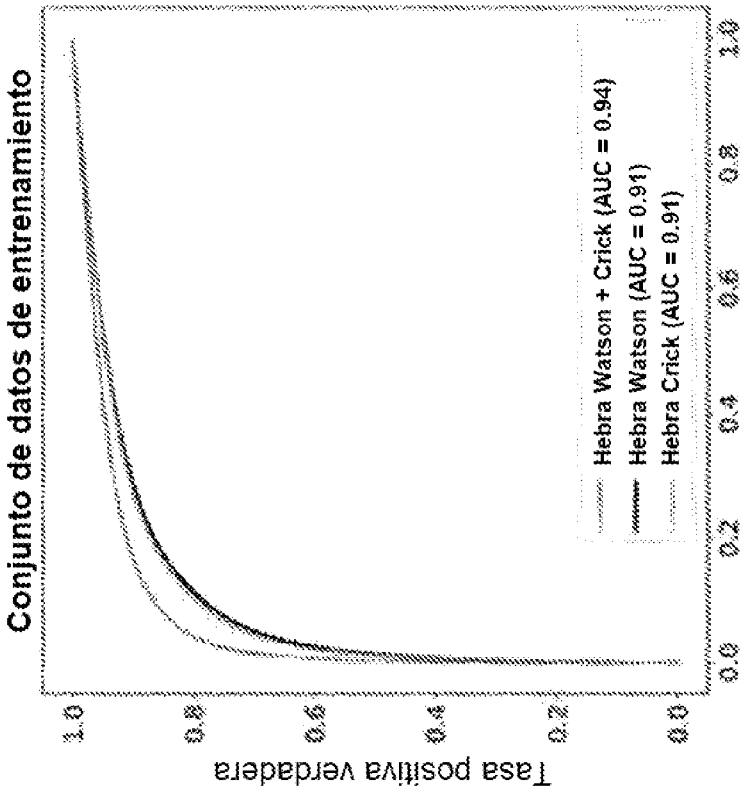


FIG. 20A

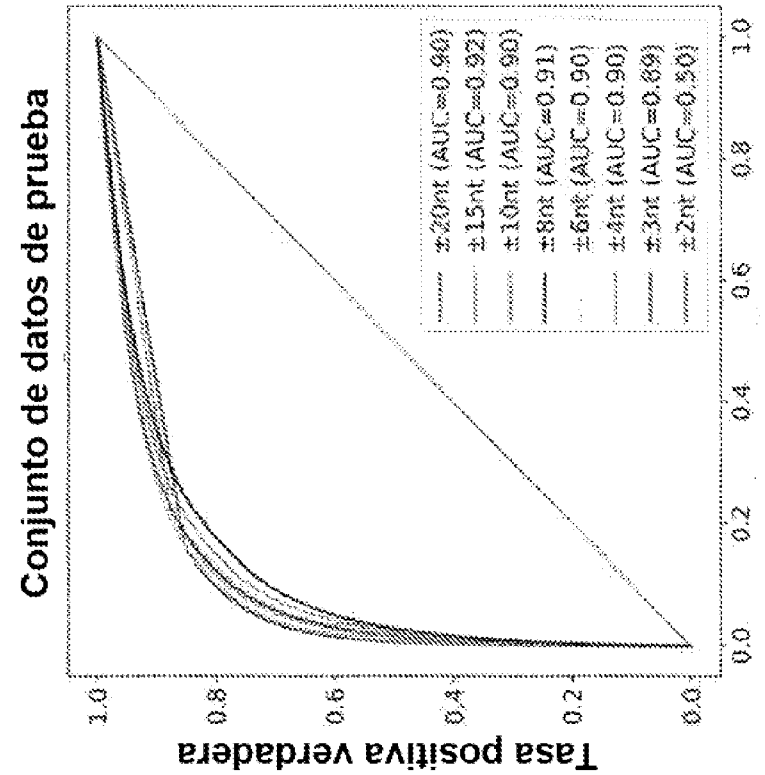


FIG. 21B

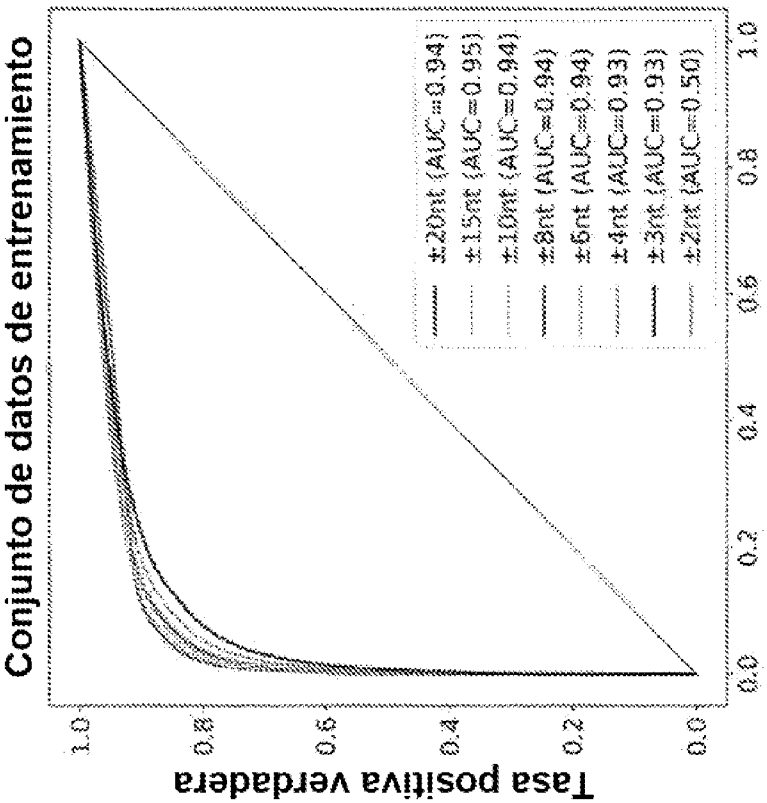


FIG. 21A

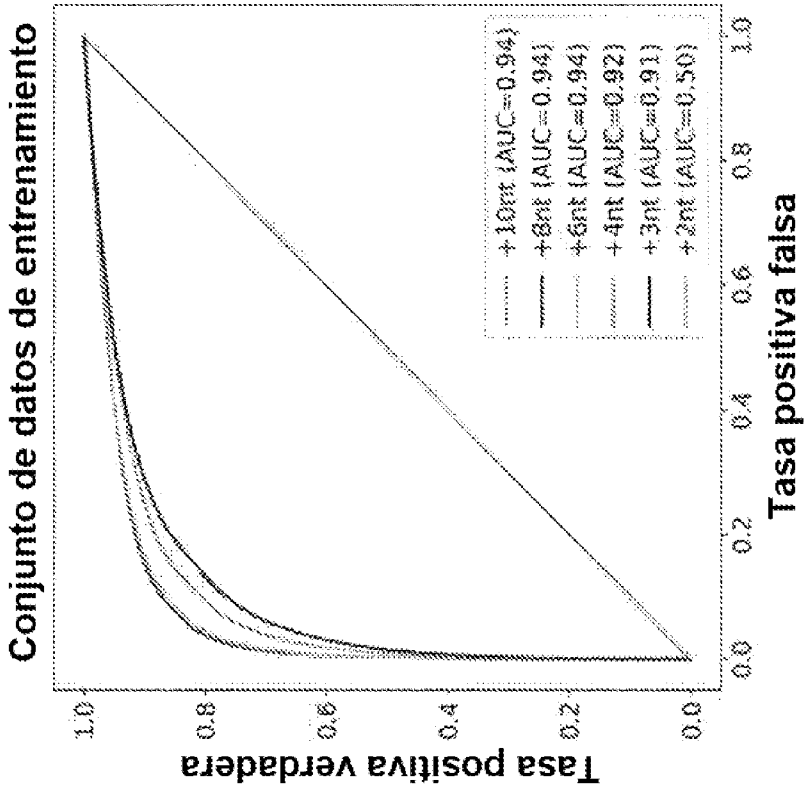


FIG. 22A

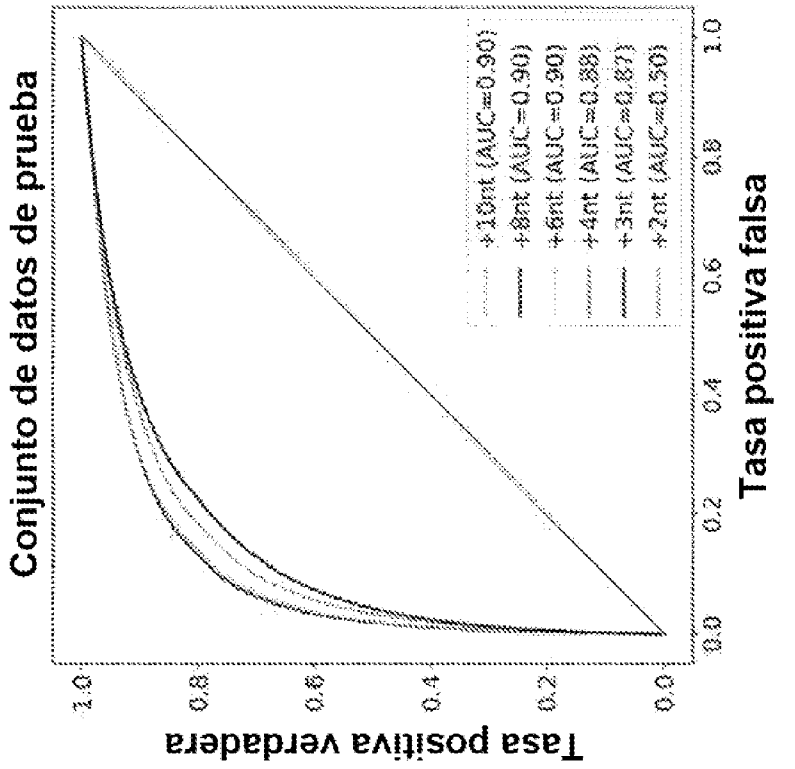


FIG. 22B

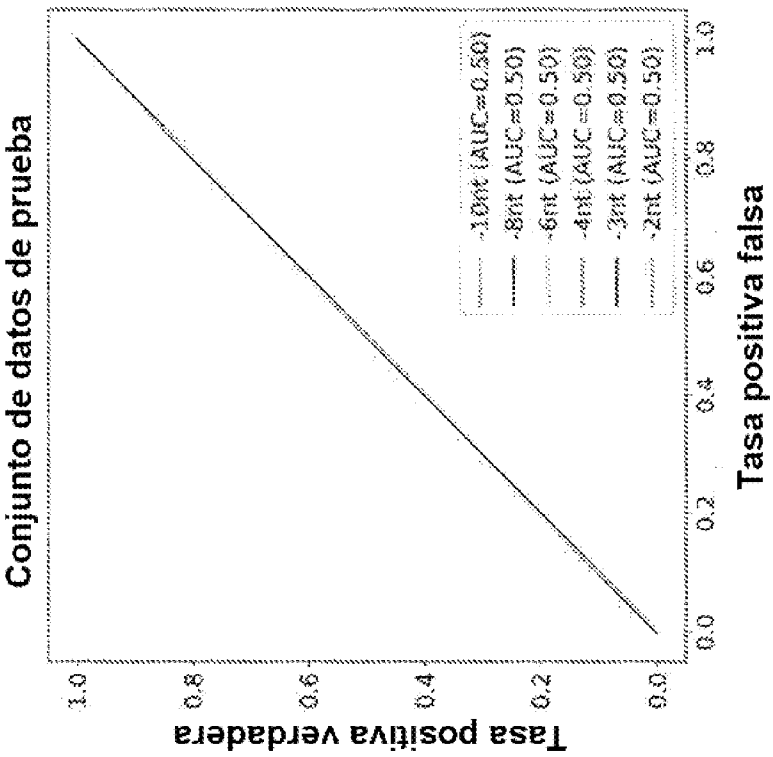


FIG. 23B

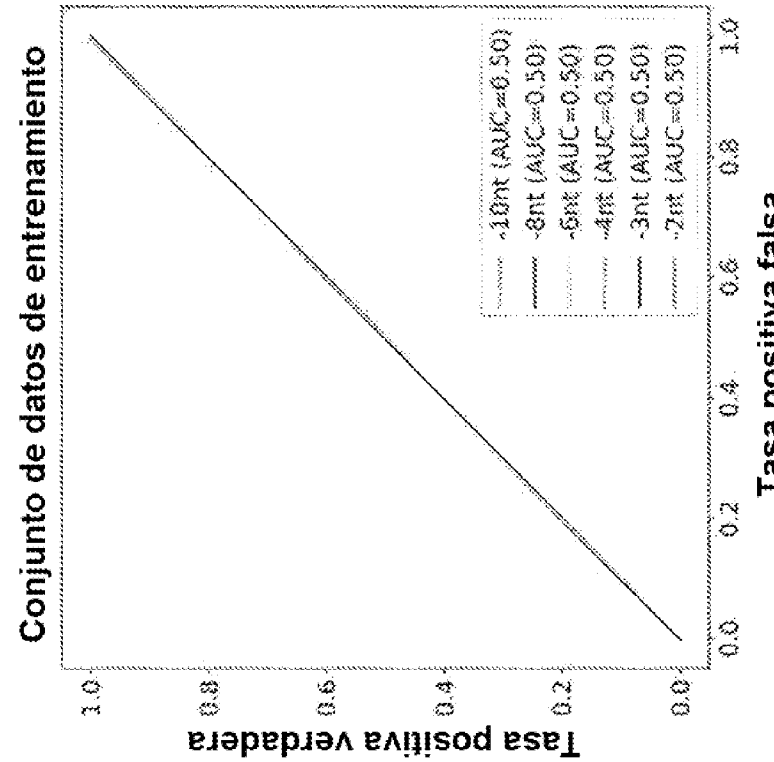


FIG. 23A

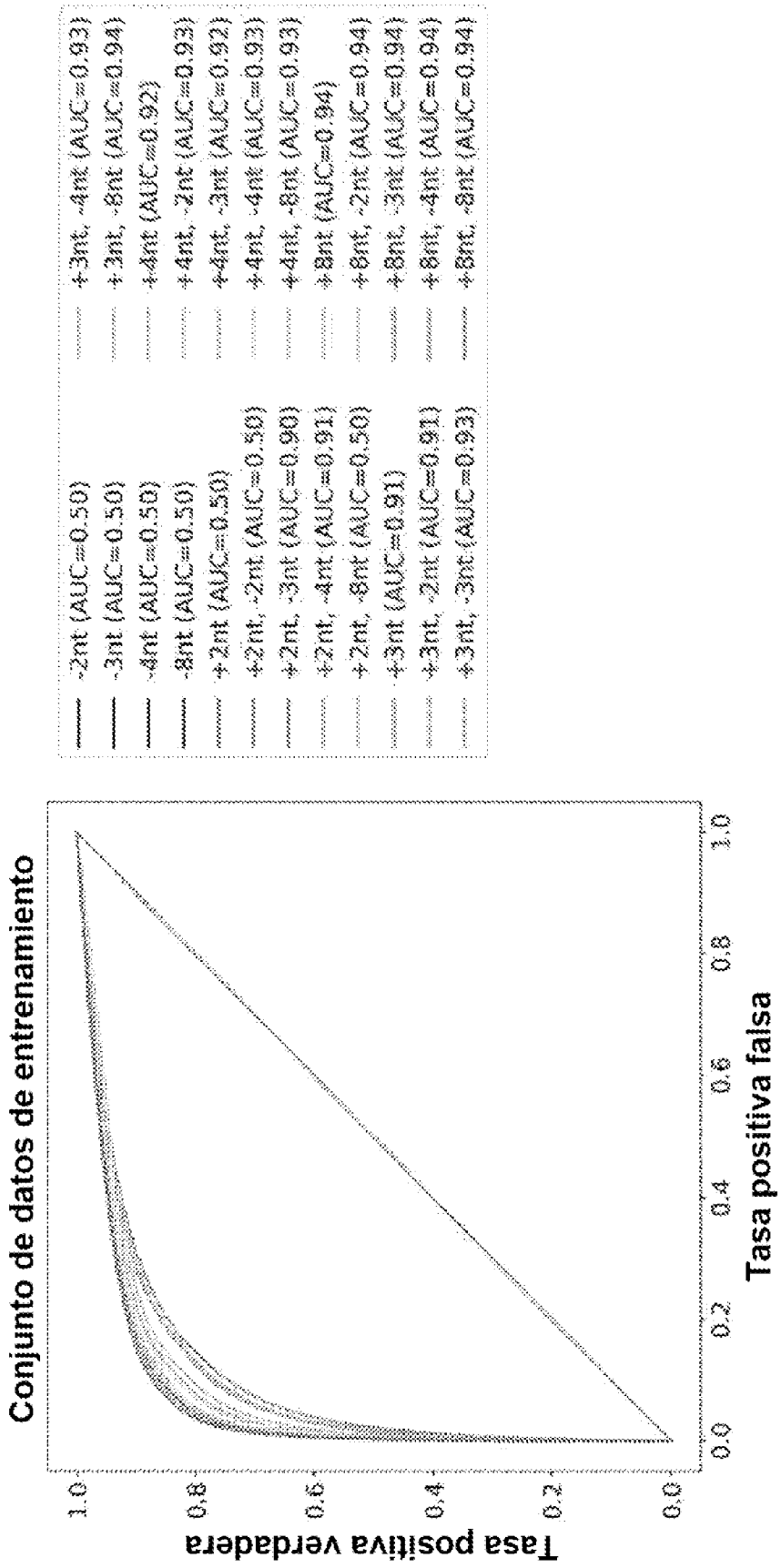

Conjunto de datos de entrenamiento
Tasa positiva verdadera
Tasa positiva falsa
0.0
0.2
0.4
0.6
0.8
1.0
-2nt (AUC=0.50)
-3nt (AUC=0.50)
-4nt (AUC=0.50)
-8nt (AUC=0.50)
+2nt (AUC=0.50)
+2nt, -2nt (AUC=0.50)
+2nt, -3nt (AUC=0.90)
+2nt, -4nt (AUC=0.91)
+2nt, -8nt (AUC=0.50)
+3nt (AUC=0.91)
+3nt, -2nt (AUC=0.91)
+3nt, -3nt (AUC=0.93)
+3nt, -4nt (AUC=0.93)
+3nt, -8nt (AUC=0.94)
+4nt (AUC=0.92)
+4nt, -2nt (AUC=0.93)
+4nt, -3nt (AUC=0.92)
+4nt, -4nt (AUC=0.93)
+4nt, -8nt (AUC=0.93)
+8nt (AUC=0.94)
+8nt, -2nt (AUC=0.94)
+8nt, -3nt (AUC=0.94)
+8nt, -4nt (AUC=0.94)
+8nt, -8nt (AUC=0.94)

FIG. 24

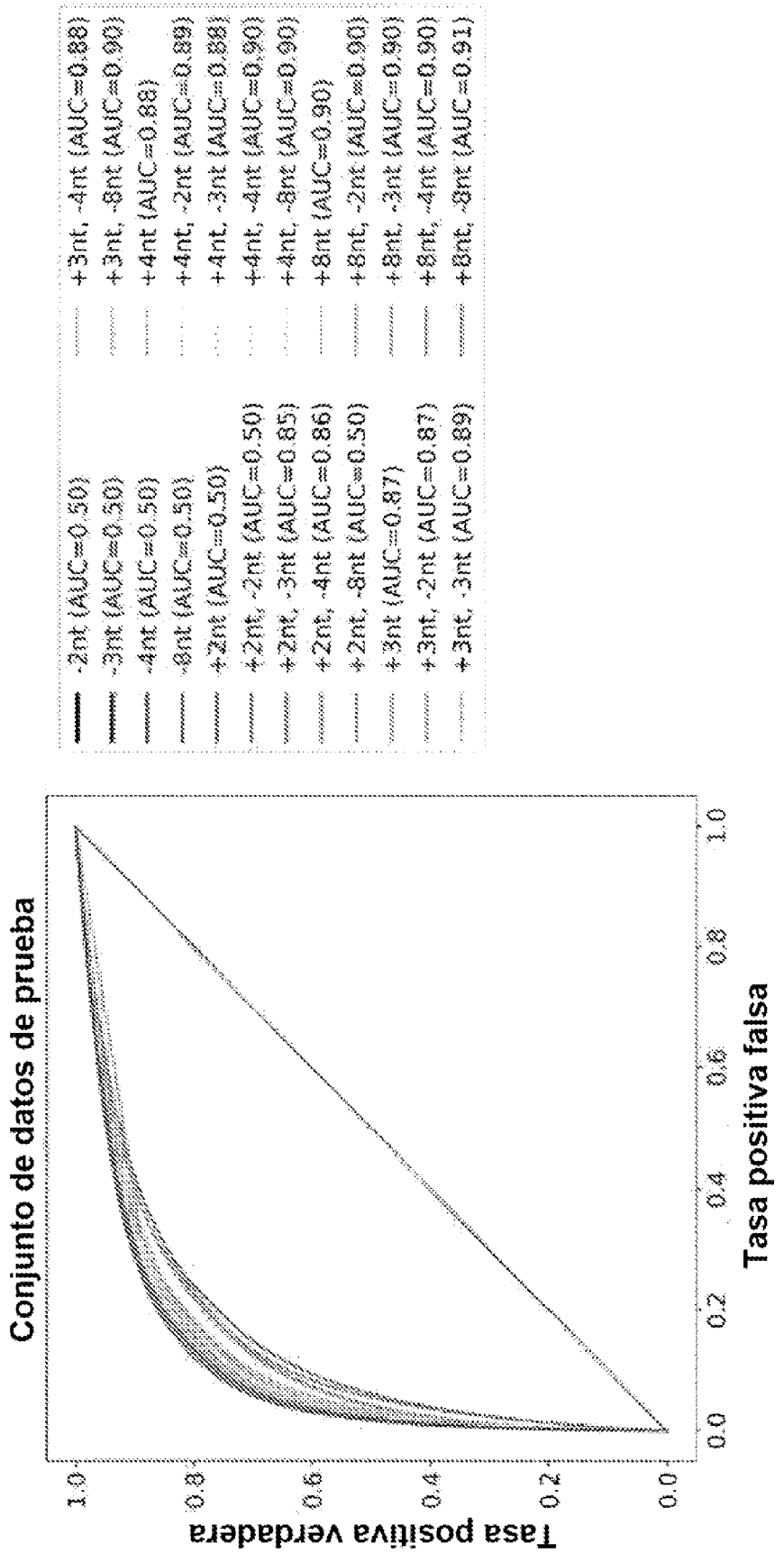

Conjunto de datos de prueba
Tasa positiva verdadera
Tasa positiva falsa
0.0
0.2
0.4
0.6
0.8
1.0
-2nt (AUC=0.50)
-3nt (AUC=0.50)
-4nt (AUC=0.50)
-8nt (AUC=0.50)
+2nt (AUC=0.50)
+2nt, -2nt (AUC=0.50)
+2nt, -3nt (AUC=0.85)
+2nt, -4nt (AUC=0.86)
+2nt, -8nt (AUC=0.50)
+3nt (AUC=0.87)
+3nt, -2nt (AUC=0.87)
+3nt, -3nt (AUC=0.89)
+3nt, -4nt (AUC=0.88)
+3nt, -8nt (AUC=0.90)
+4nt (AUC=0.88)
+4nt, -2nt (AUC=0.89)
+4nt, -3nt (AUC=0.88)
+4nt, -4nt (AUC=0.90)
+4nt, -8nt (AUC=0.90)
+8nt (AUC=0.90)
+8nt, -2nt (AUC=0.90)
+8nt, -3nt (AUC=0.90)
+8nt, -4nt (AUC=0.90)
+8nt, -8nt (AUC=0.91)

FIG. 25

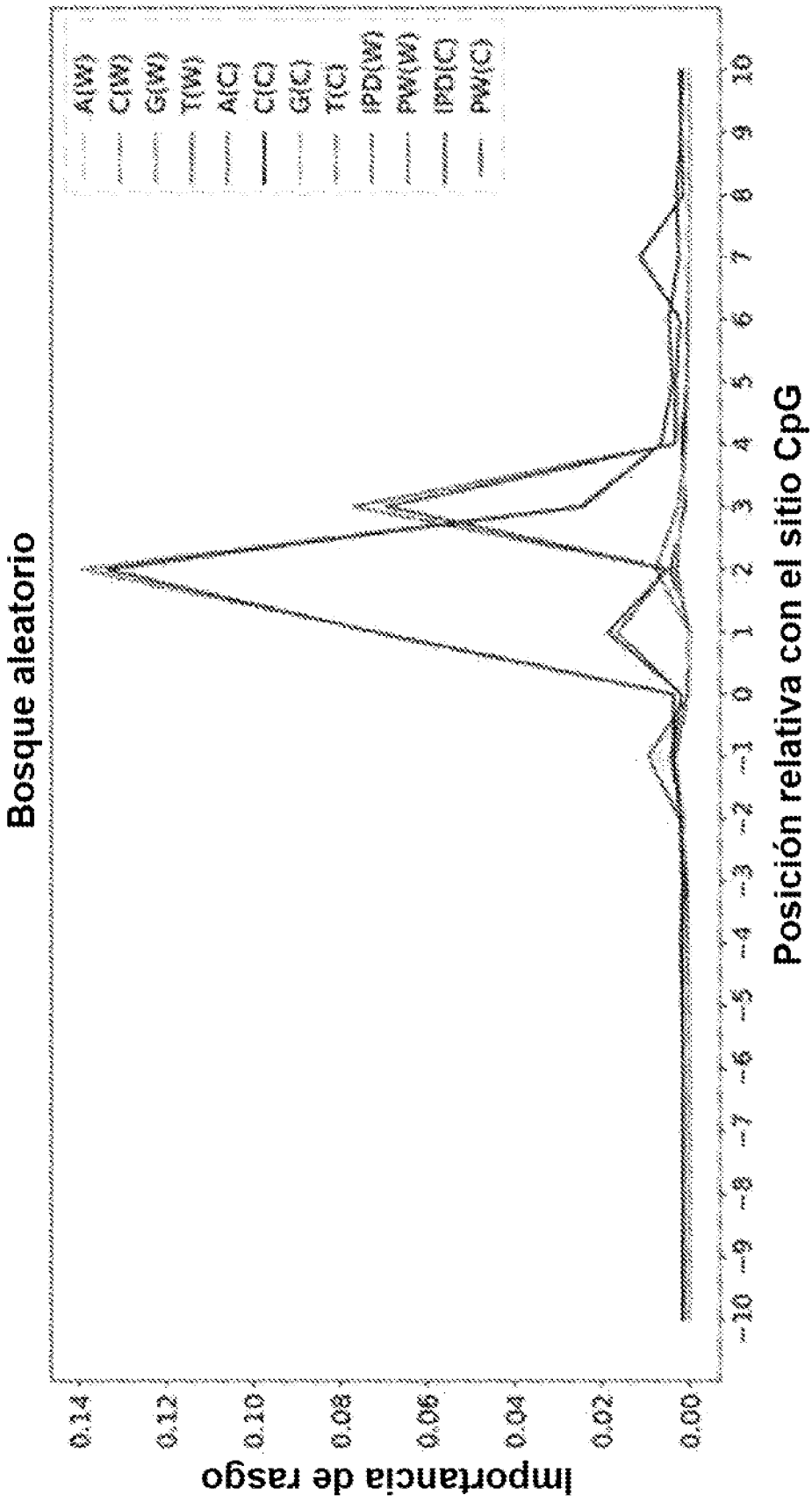

Bosque aleatorio
importancia de rasgo
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
A(W)
C(W)
G(W)
T(W)
A(C)
C(C)
G(C)
T(C)
PD(W)
PW(W)
PD(C)
PW(C)
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10
Posición relativa con el sitio CpG

FIG. 26

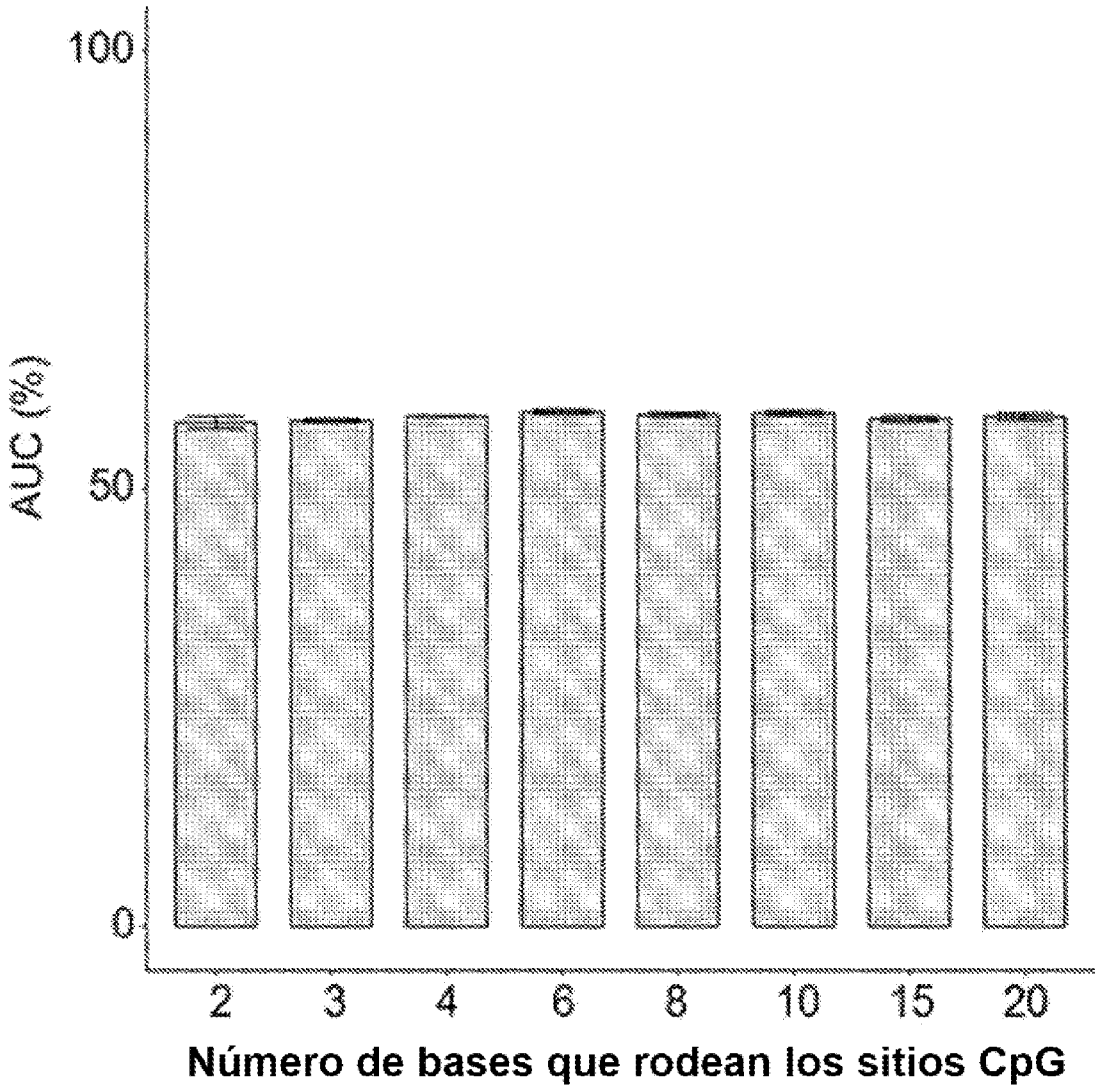

100
50
0
AUC (%)
2
3
4
6
8
10
15
20
Número de bases que rodean los sitios CpG

FIG. 27

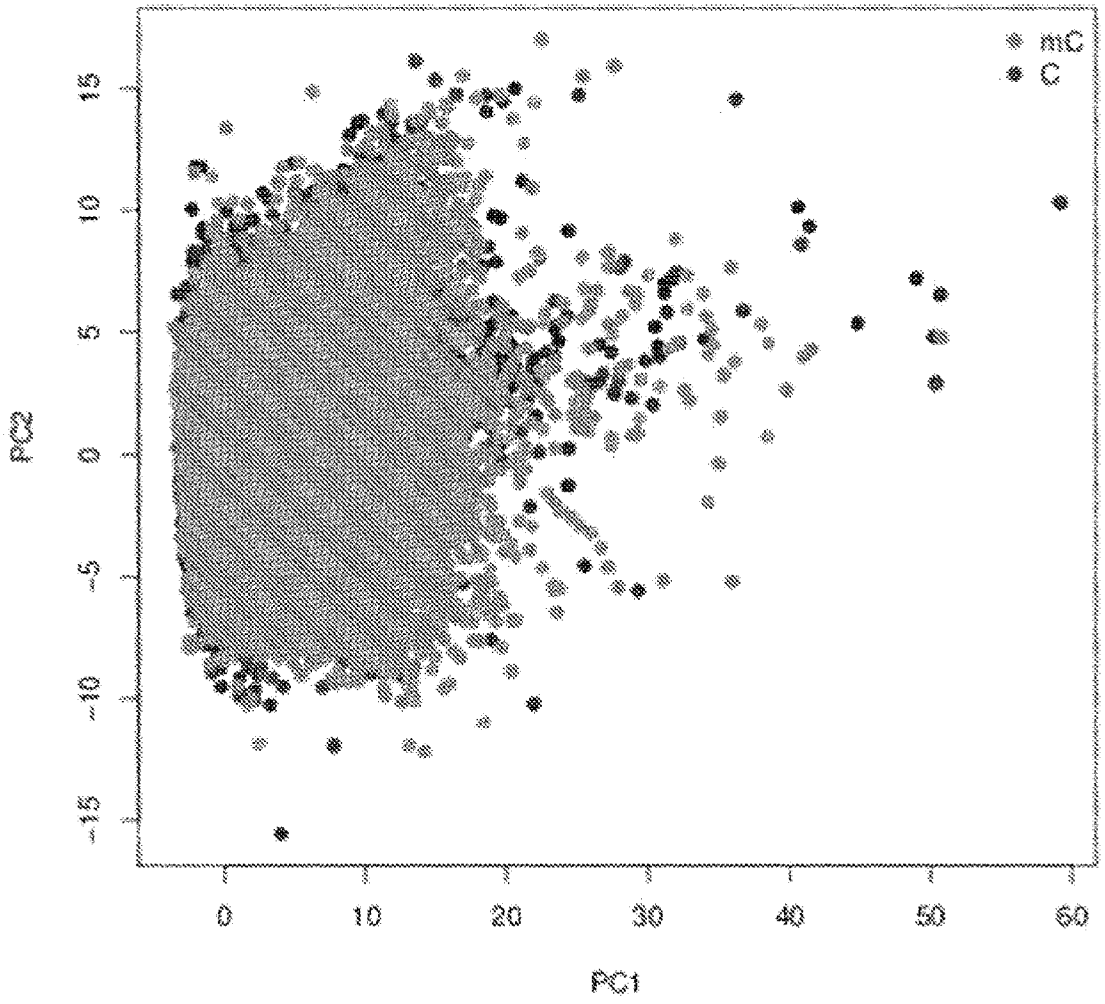


FIG. 28

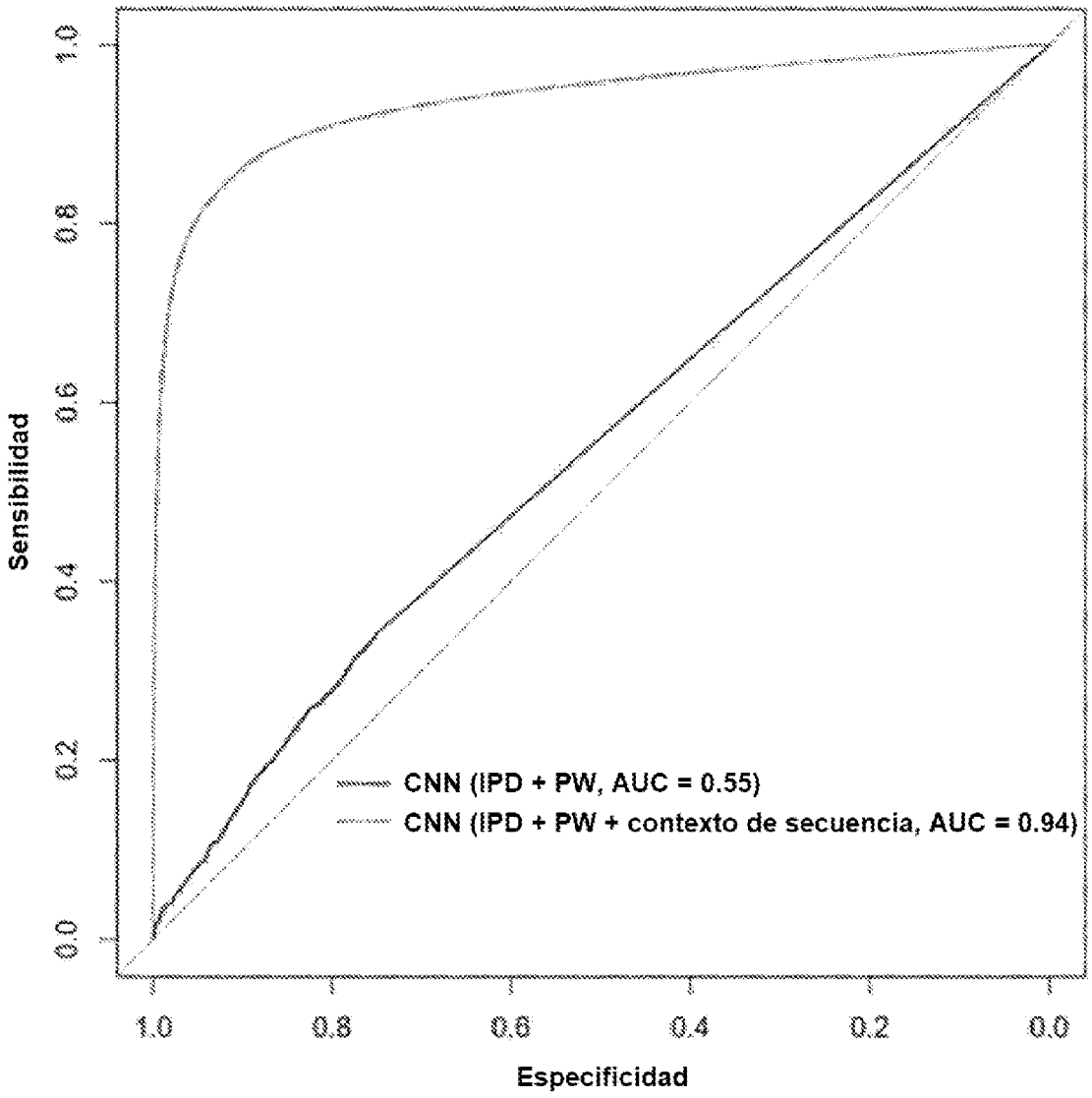

Sensibilidad
1.0
0.8
0.6
0.4
0.2
0.0
CNN (IPD + PW, AUC = 0.55)
CNN (IPD + PW + contexto de secuencia, AUC = 0.94)
1.0
0.8
0.6
0.4
0.2
0.0
Especificidad

FIG. 29

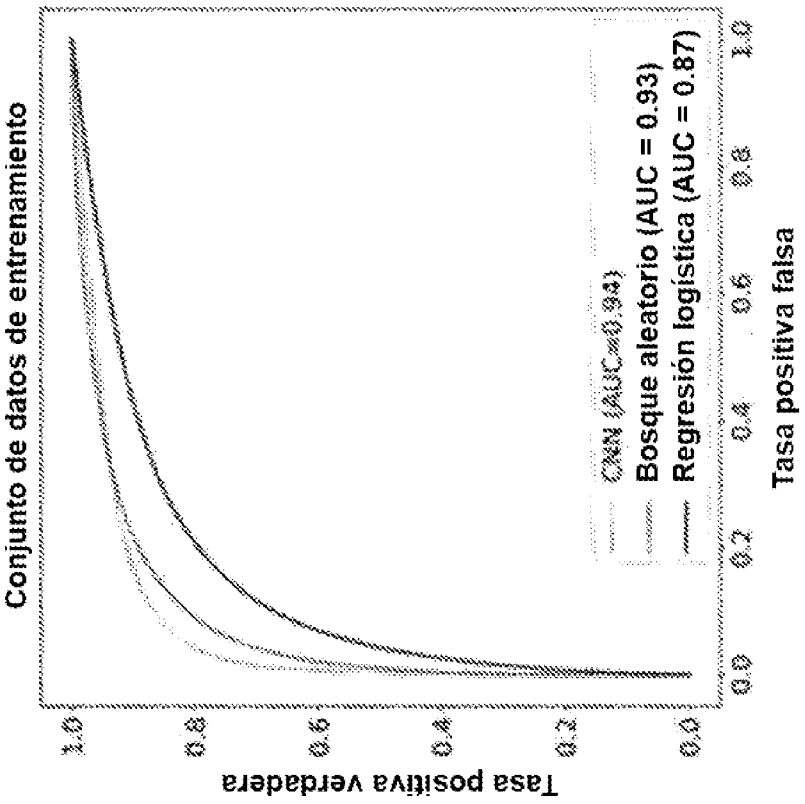


FIG. 30A

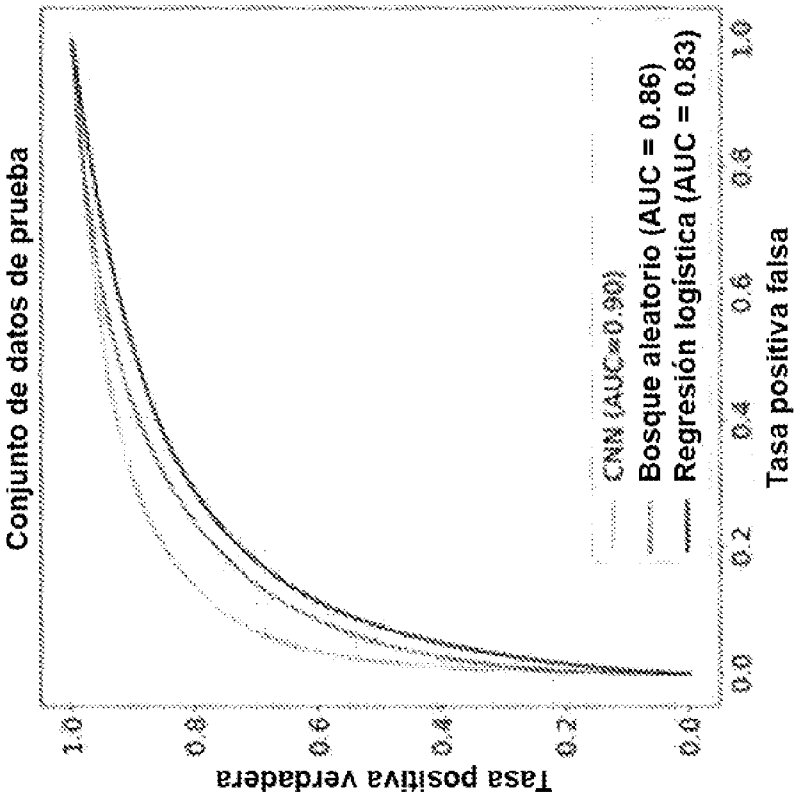
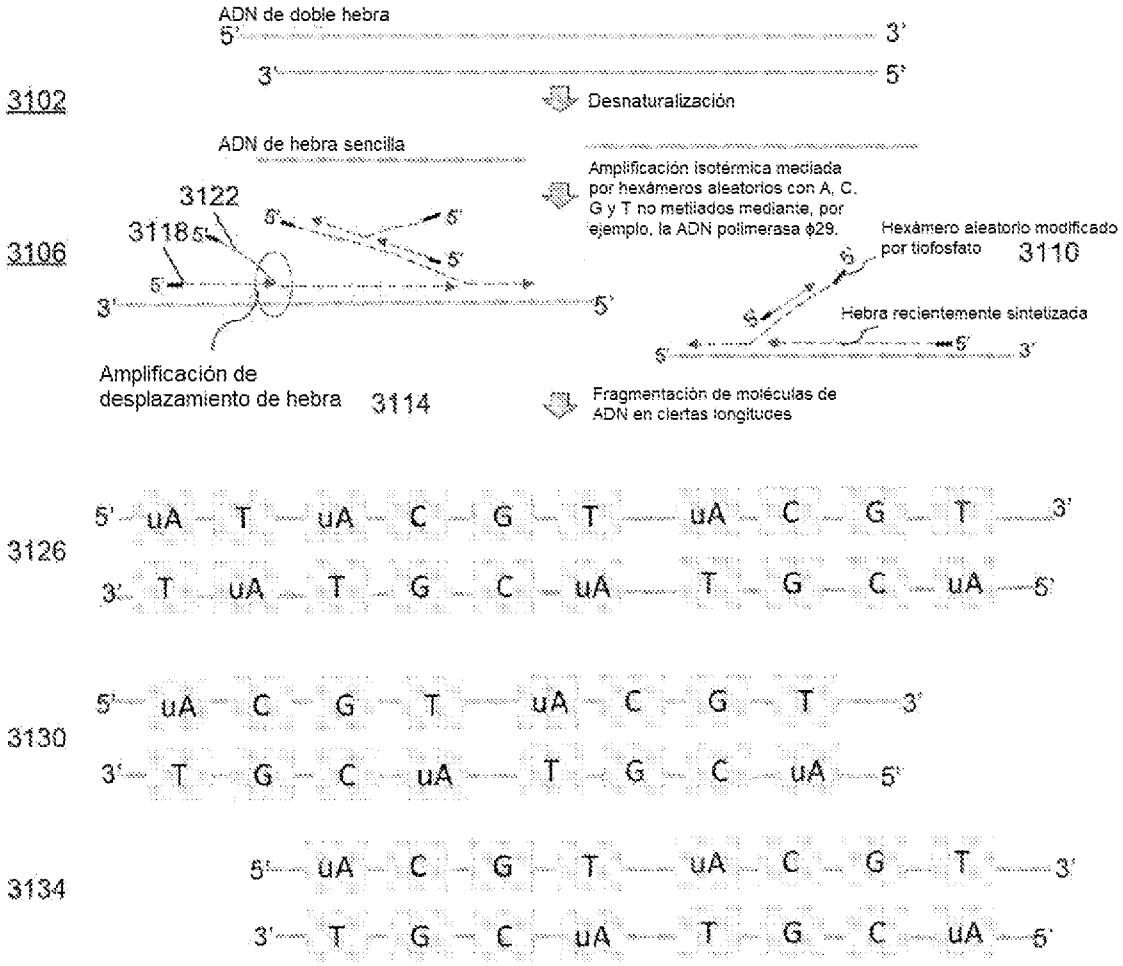
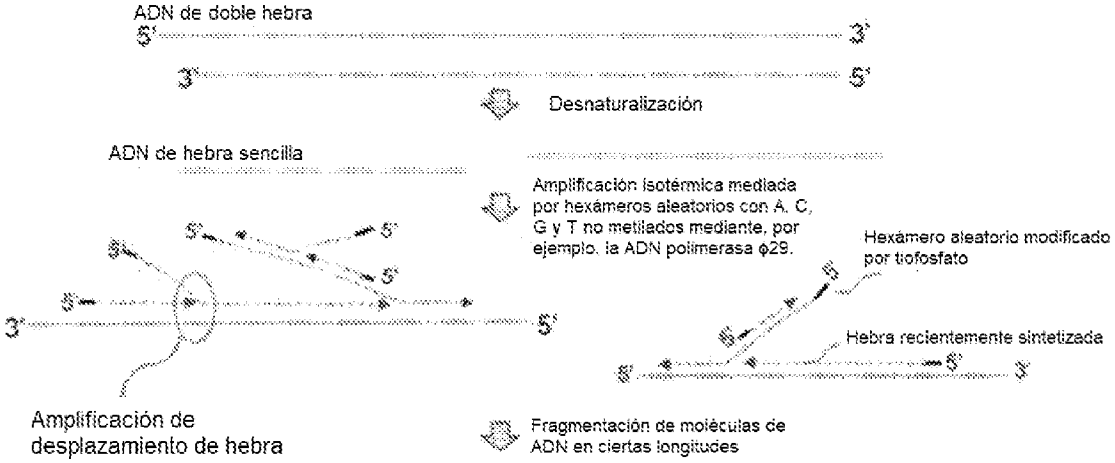


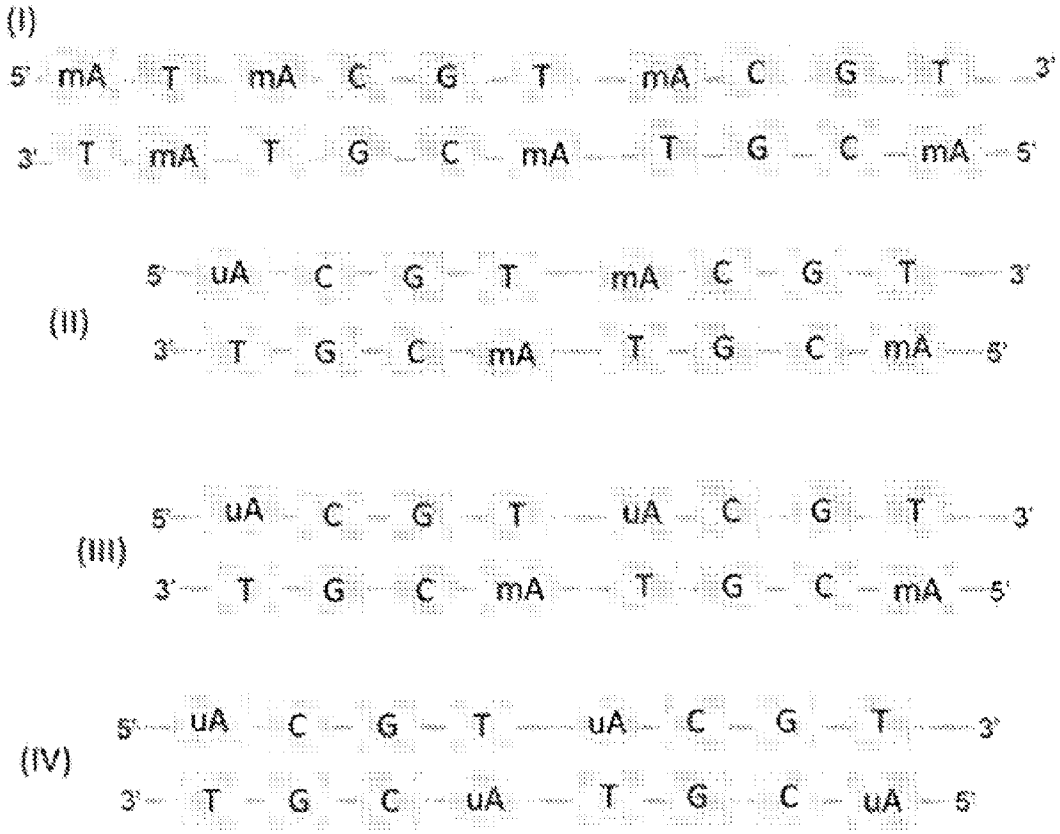
FIG. 30B



Productos de ADN amplificados de genoma completo

FIG. 31A





Productos de ADN amplificados de genoma completo

FIG. 31B

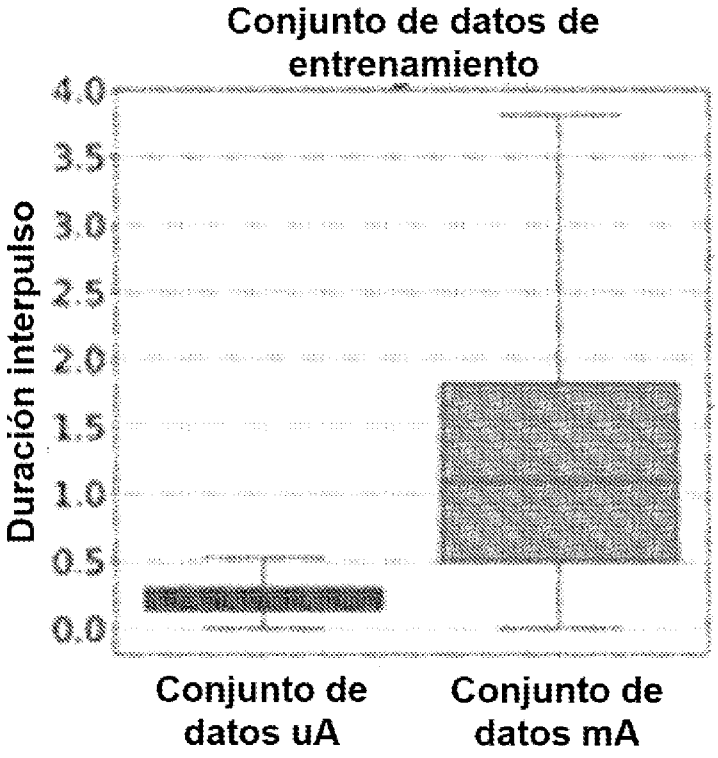


FIG. 32A

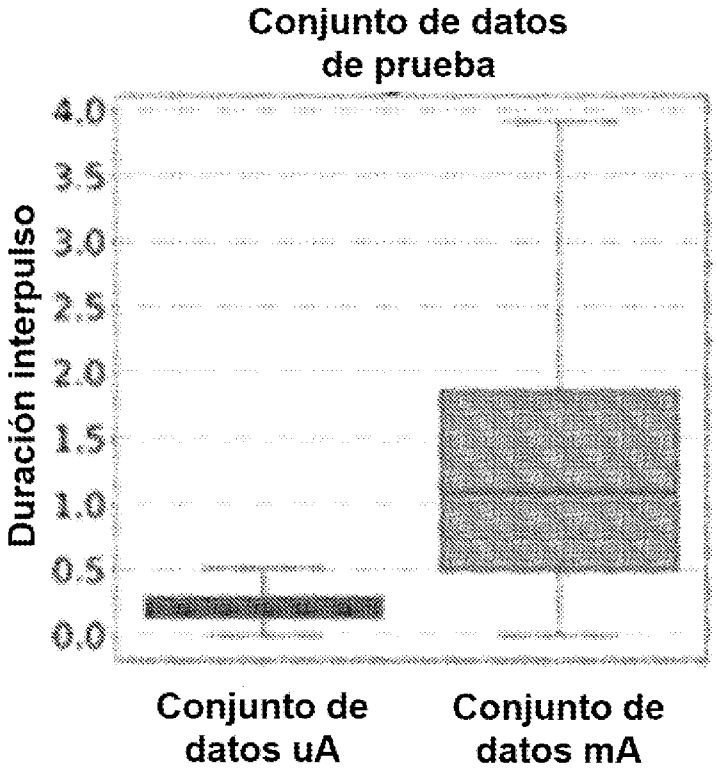


FIG. 32B

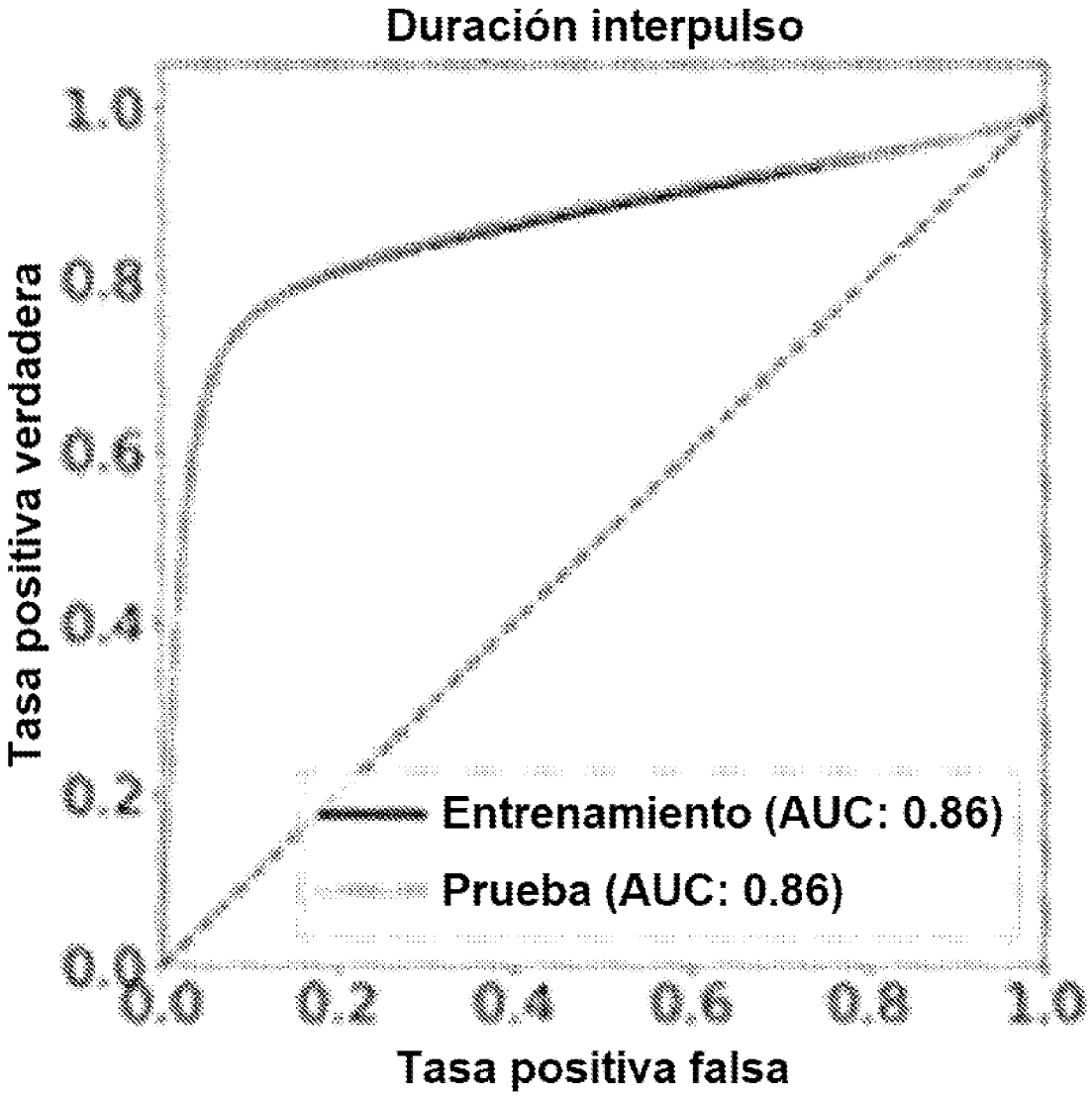

Duración interpulso
Tasa positiva verdadera
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Tasa positiva falsa
Entrenamiento (AUC: 0.86)
Prueba (AUC: 0.86)

FIG. 32C

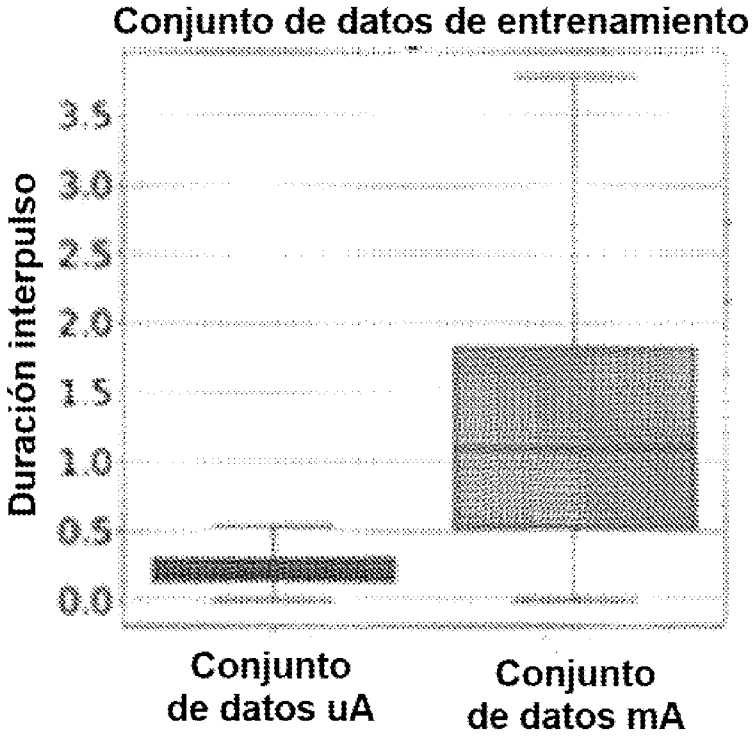


FIG. 33A

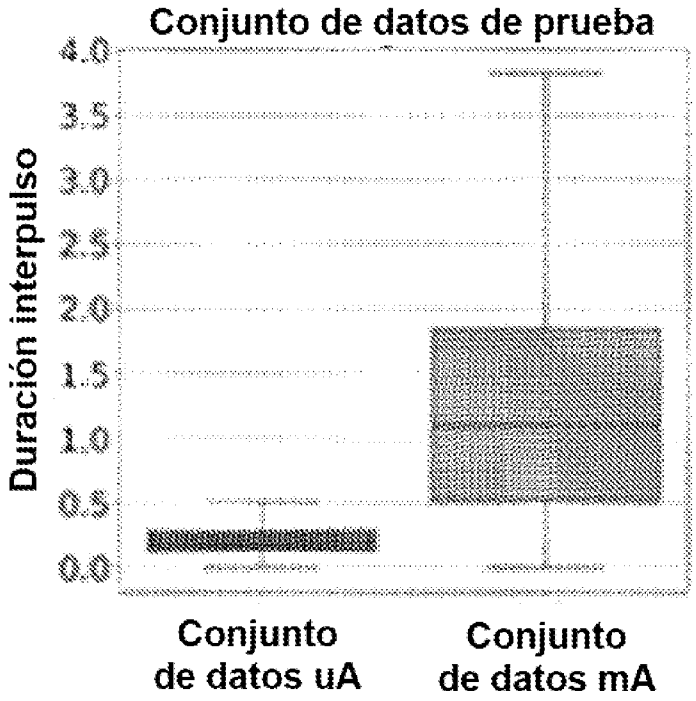


FIG. 33B

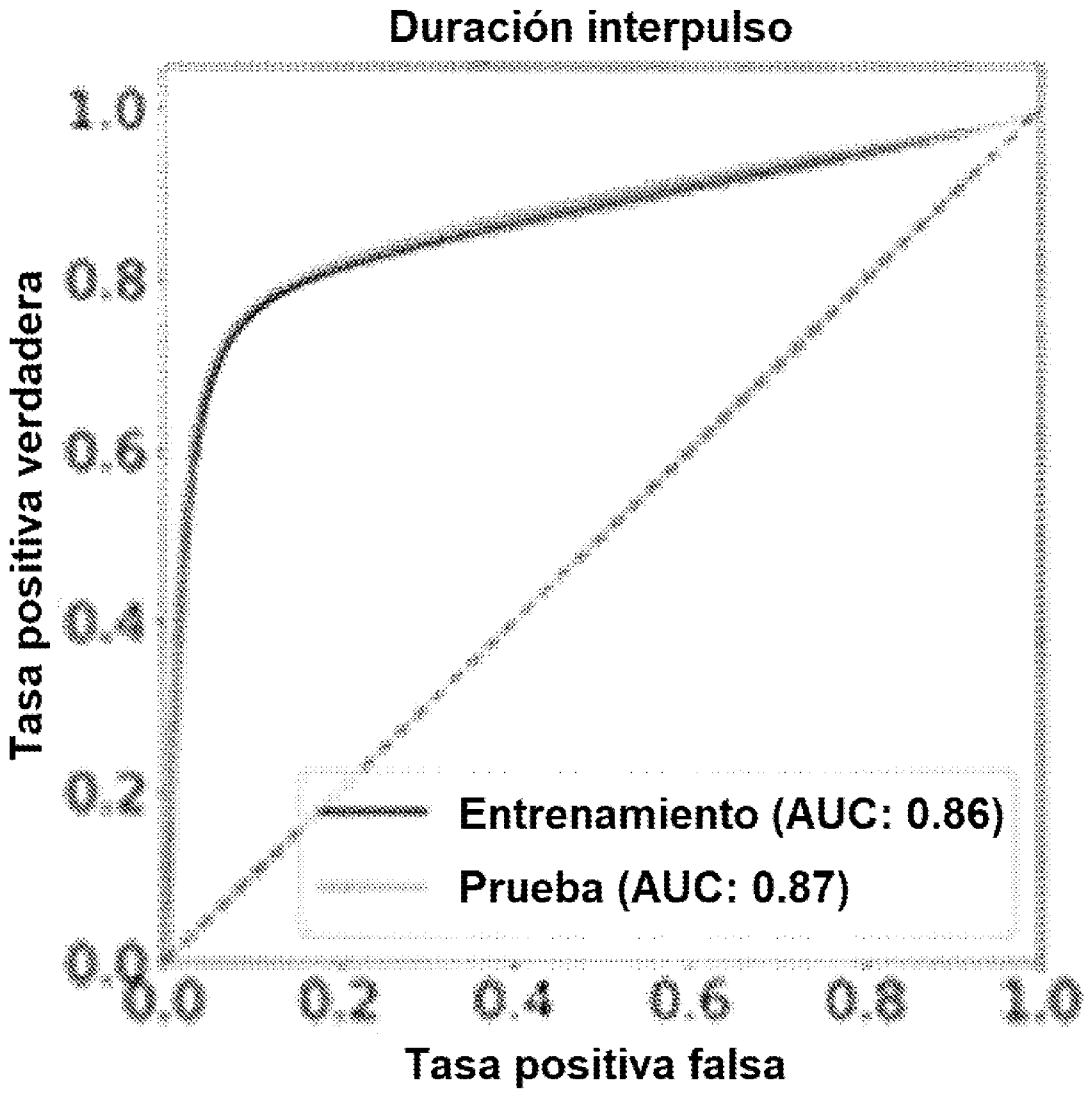

Duración interpulso
Tasa positiva verdadera
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Tasa positiva falsa
Entrenamiento (AUC: 0.86)
Prueba (AUC: 0.87)

FIG. 33C

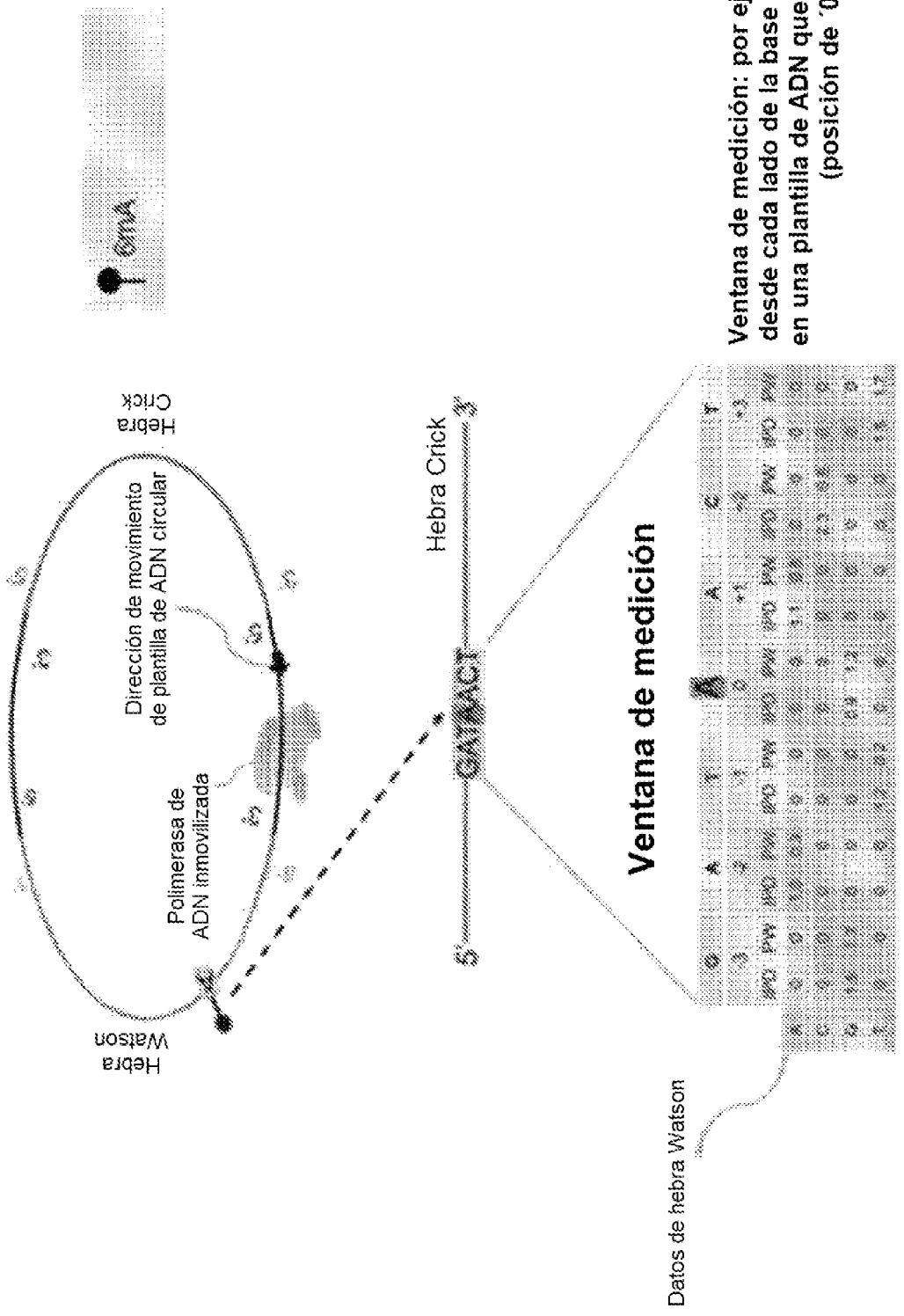
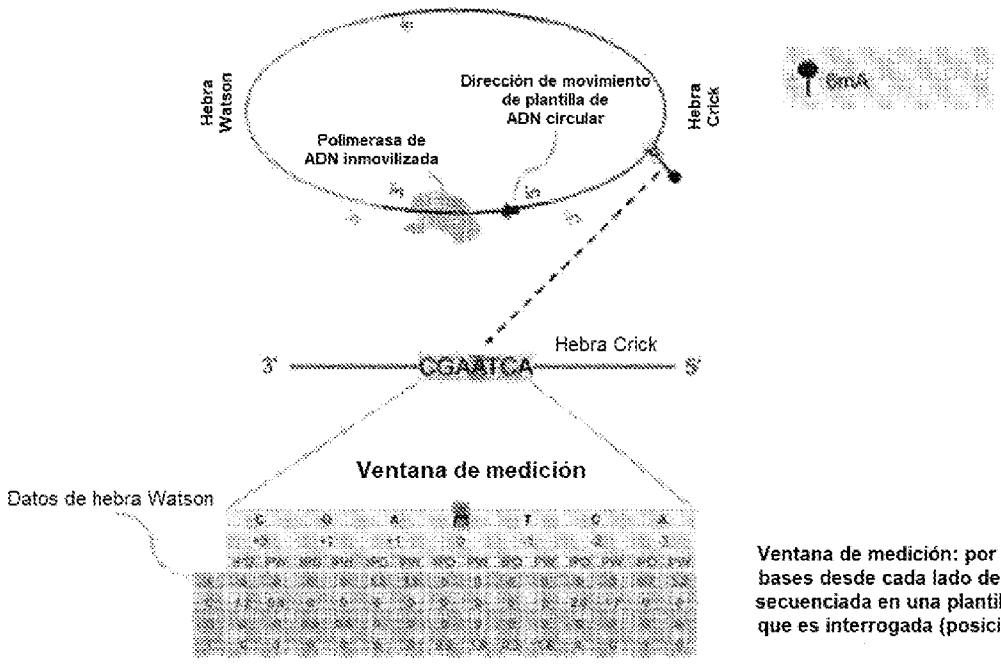

CmA
Hebra Crick
Dirección de movimiento de plantilla de ADN circular
Polimerasa de ADN inmovilizada
Hebra Watson
Hebra Crick
5'
CATTACT
3'
Ventana de medición
Ventana de medición: por ejemplo 3 bases desde cada lado de la base A secuenciada en una plantilla de ADN que es interrogada (posición de '0')
Datos de hebra Watson

FIG. 34



Ventana de medición: por ejemplo 3 bases desde cada lado de la base A secuenciada en una plantilla de ADN que es interrogada (posición de '0')

FIG.35

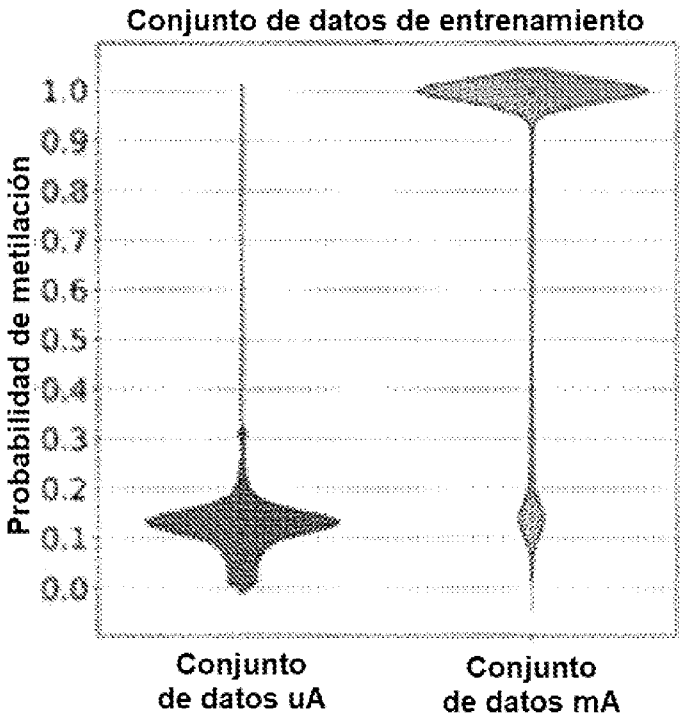


FIG. 36A

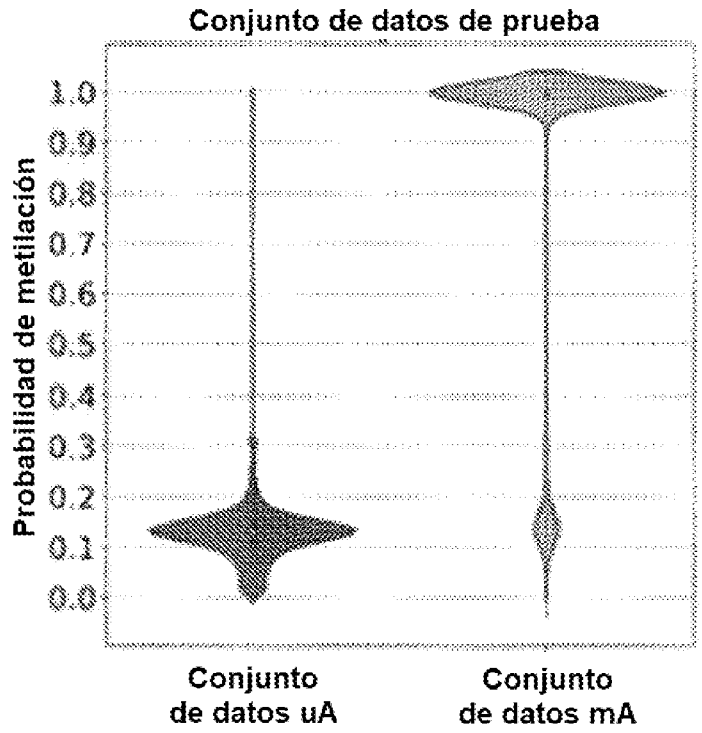


FIG. 36B

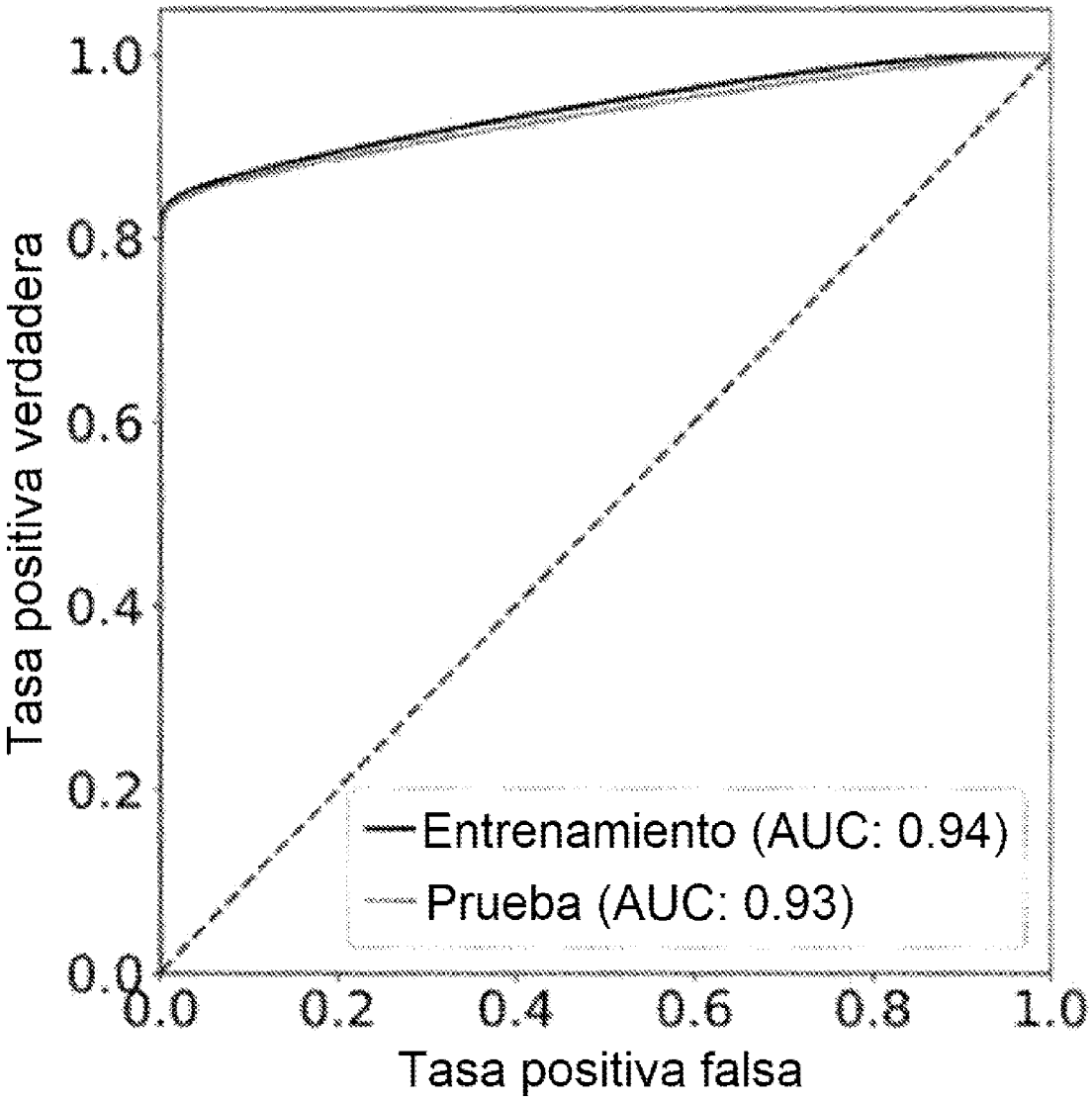

1.0
0.8
0.6
0.4
0.2
0.0
Tasa positiva verdadera
0.0
0.2
0.4
0.6
0.8
1.0
Tasa positiva falsa
Entrenamiento (AUC: 0.94)
Prueba (AUC: 0.93)

FIG. 37

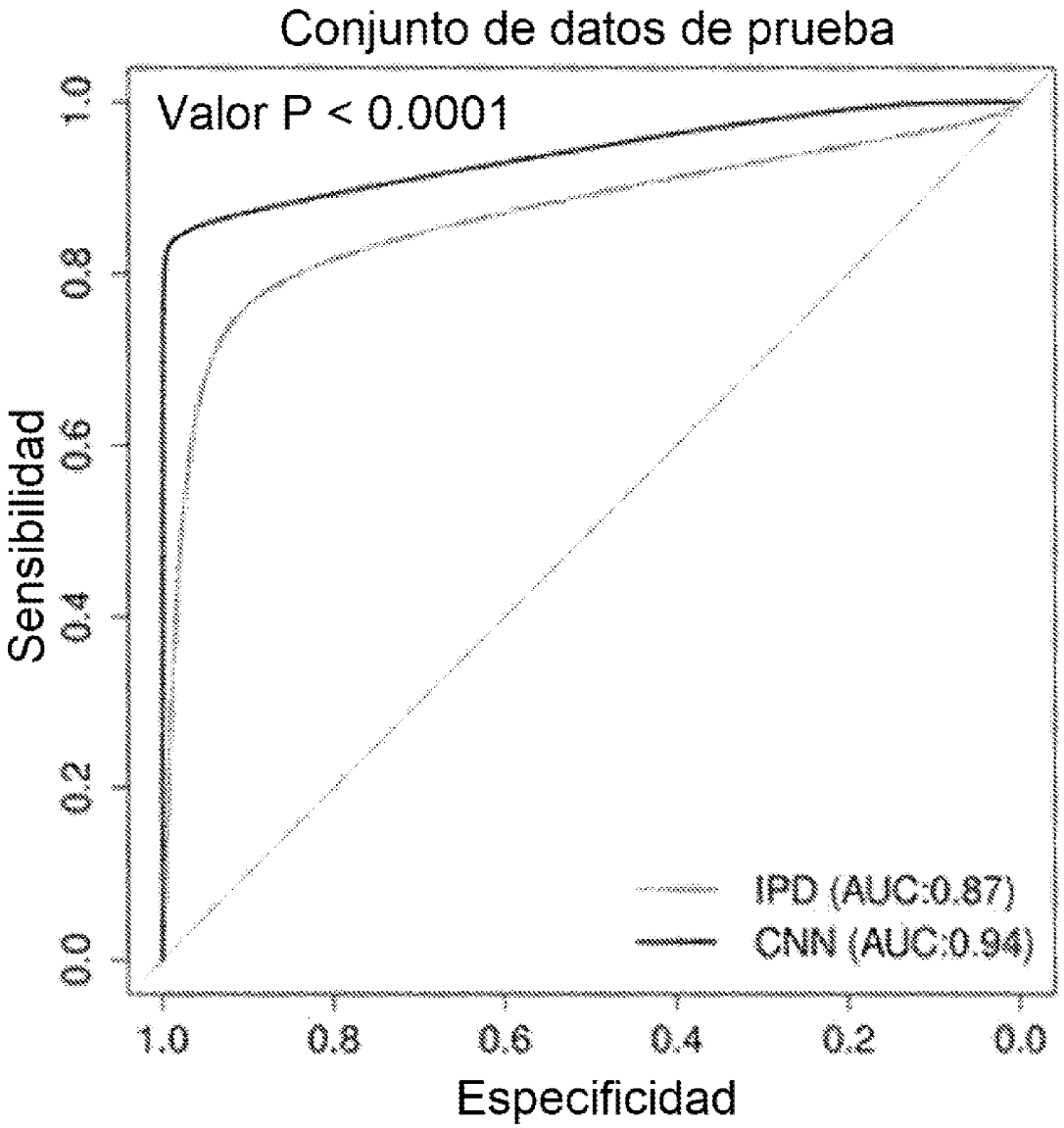

Conjunto de datos de prueba
Valor P < 0.0001
Sensibilidad
1.0
0.8
0.6
0.4
0.2
0.0
IPD (AUC:0.87)
CNN (AUC:0.94)
1.0
0.8
0.6
0.4
0.2
0.0
Especificidad

FIG. 38

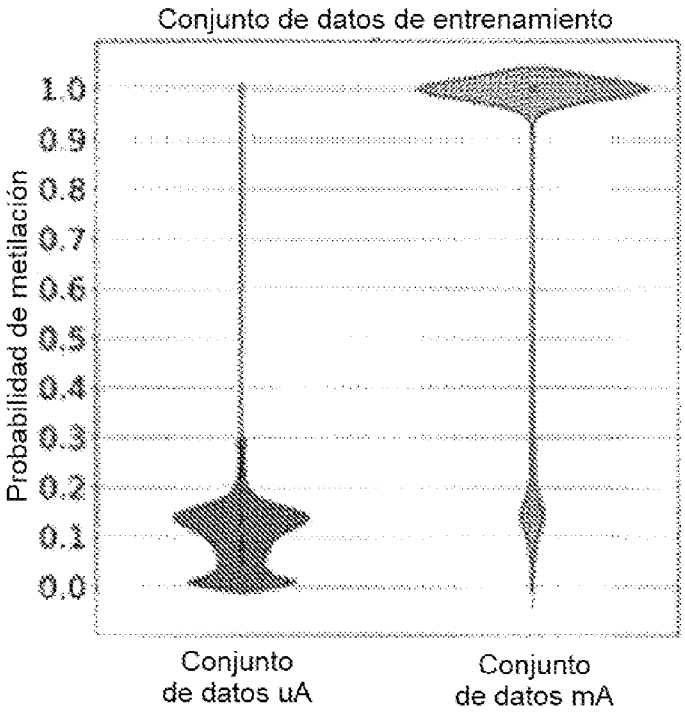


FIG. 39A

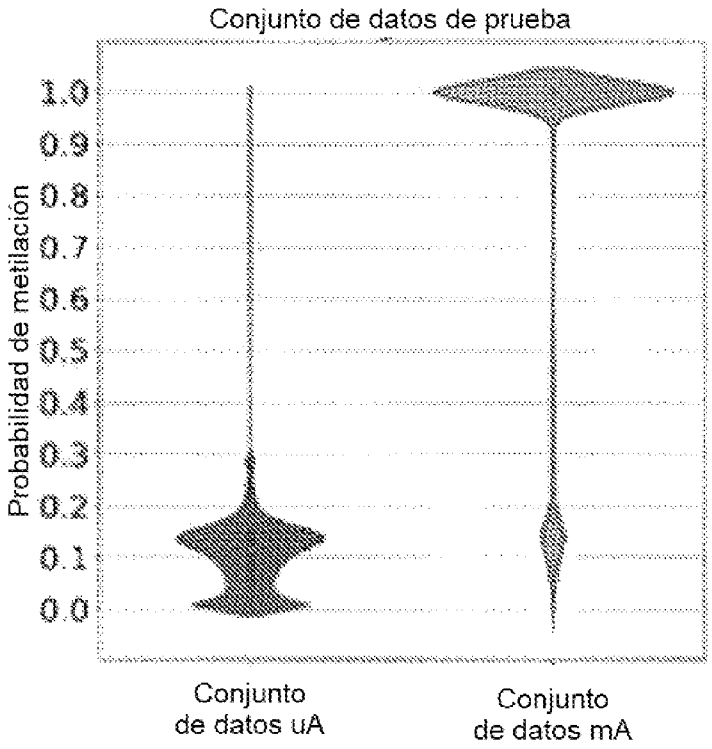


FIG. 39B

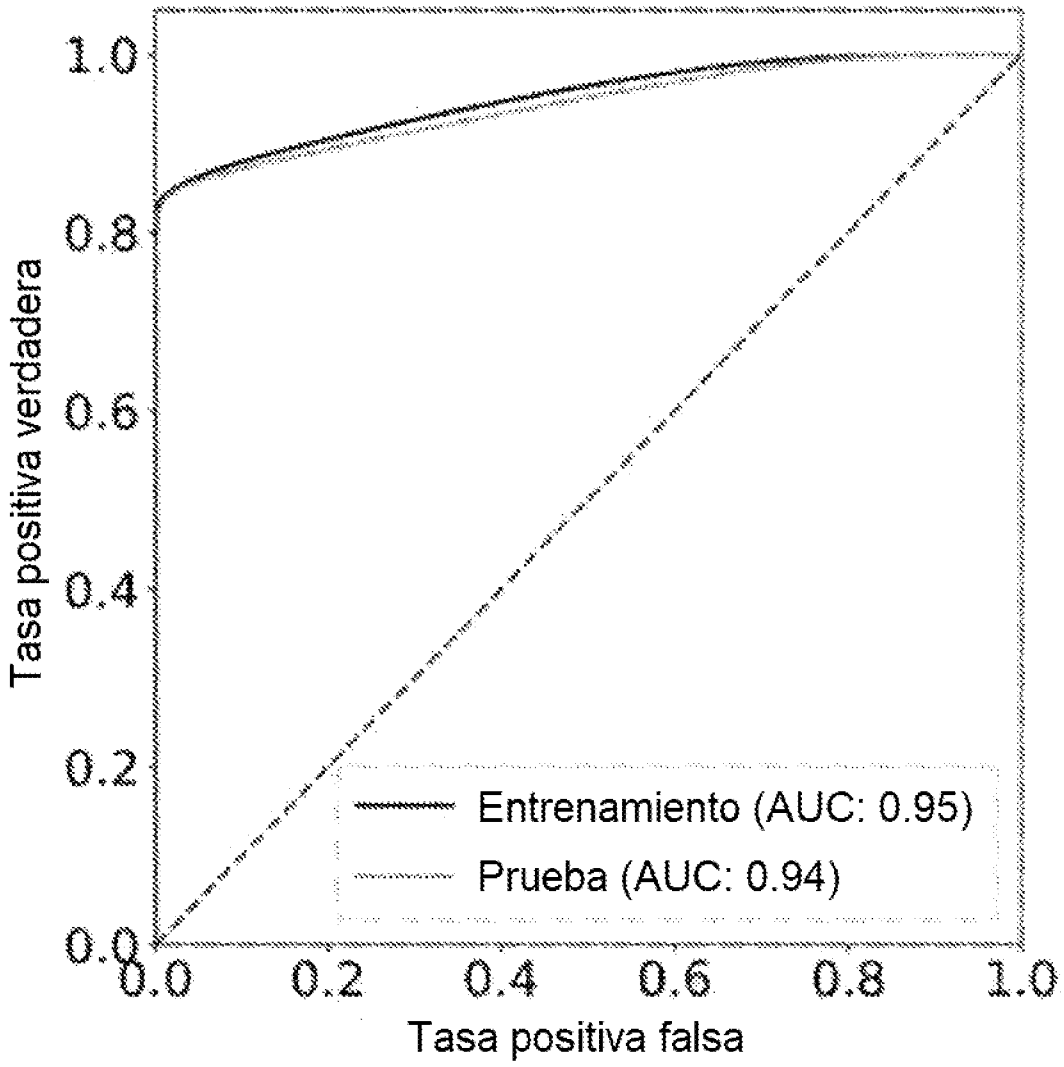

1.0
0.8
0.6
0.4
0.2
0.0
Tasa positiva verdadera
0.0
0.2
0.4
0.6
0.8
1.0
Tasa positiva falsa
Entrenamiento (AUC: 0.95)
Prueba (AUC: 0.94)

FIG. 40

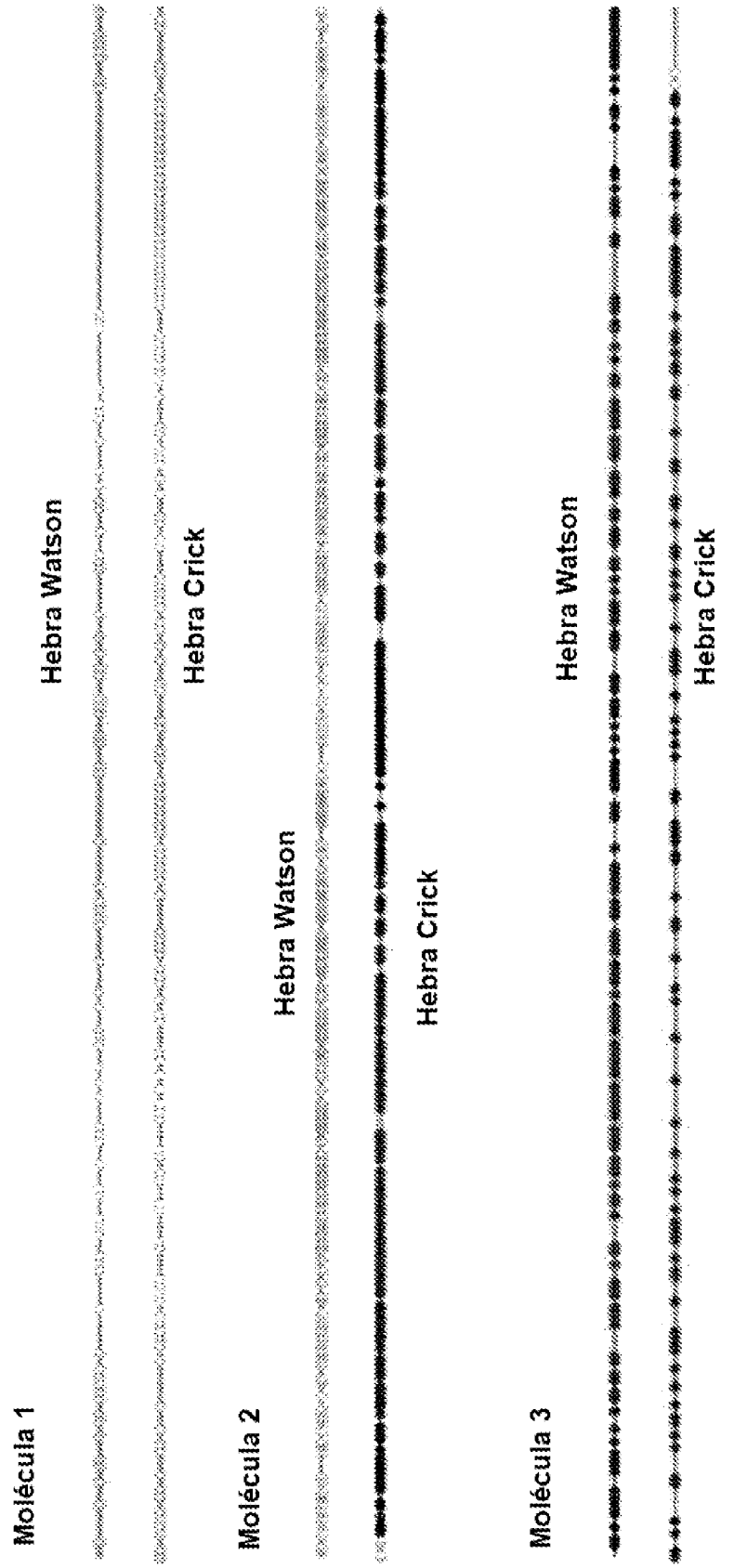

Molécula 1
Hebra Watson
Hebra Crick
Molécula 2
Hebra Watson
Hebra Crick
Molécula 3
Hebra Watson
Hebra Crick

FIG. 41

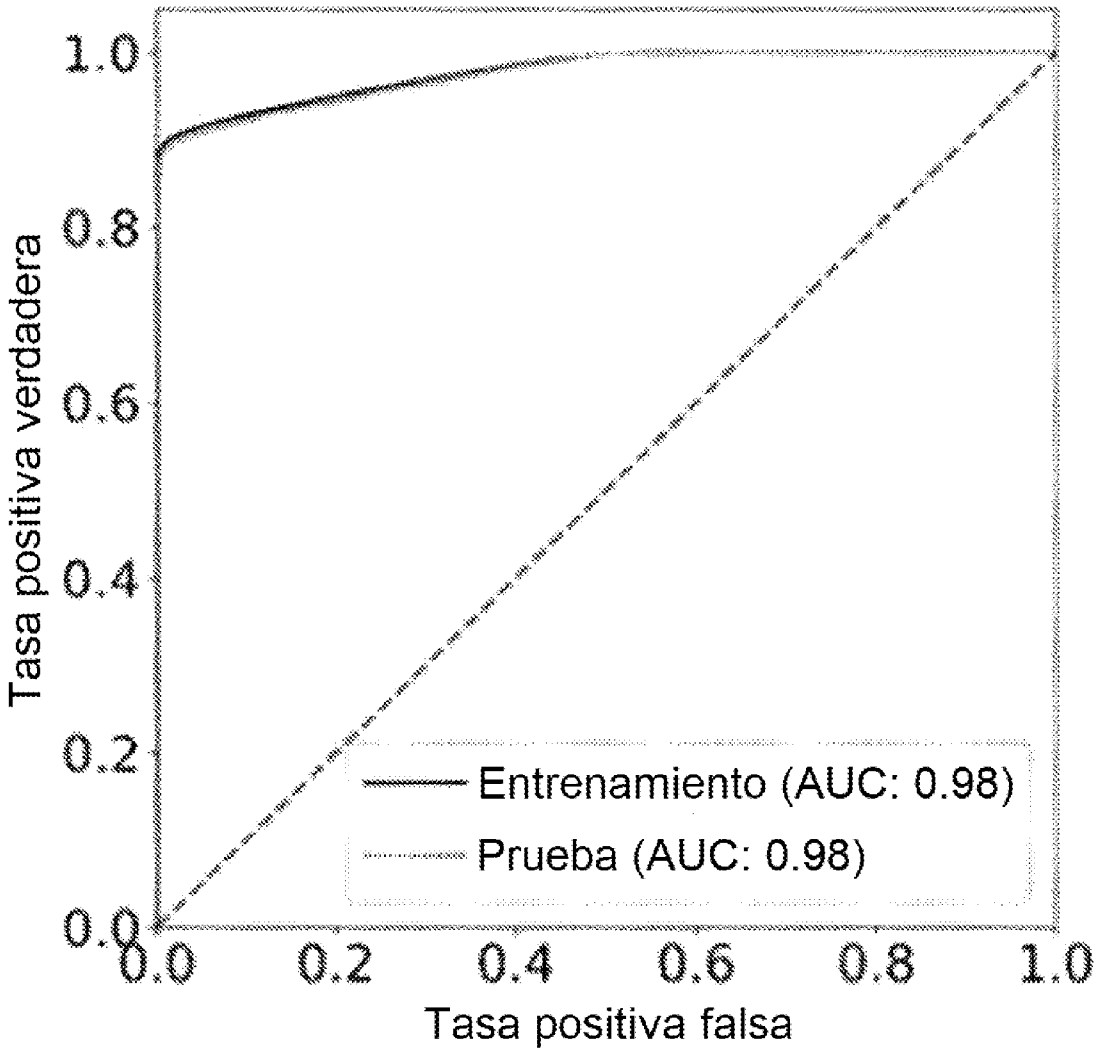

1.0
0.8
0.6
0.4
0.2
0.0
Tasa positiva verdadera
0.0
0.2
0.4
0.6
0.8
1.0
Tasa positiva falsa
Entrenamiento (AUC: 0.98)
Prueba (AUC: 0.98)

FIG. 42

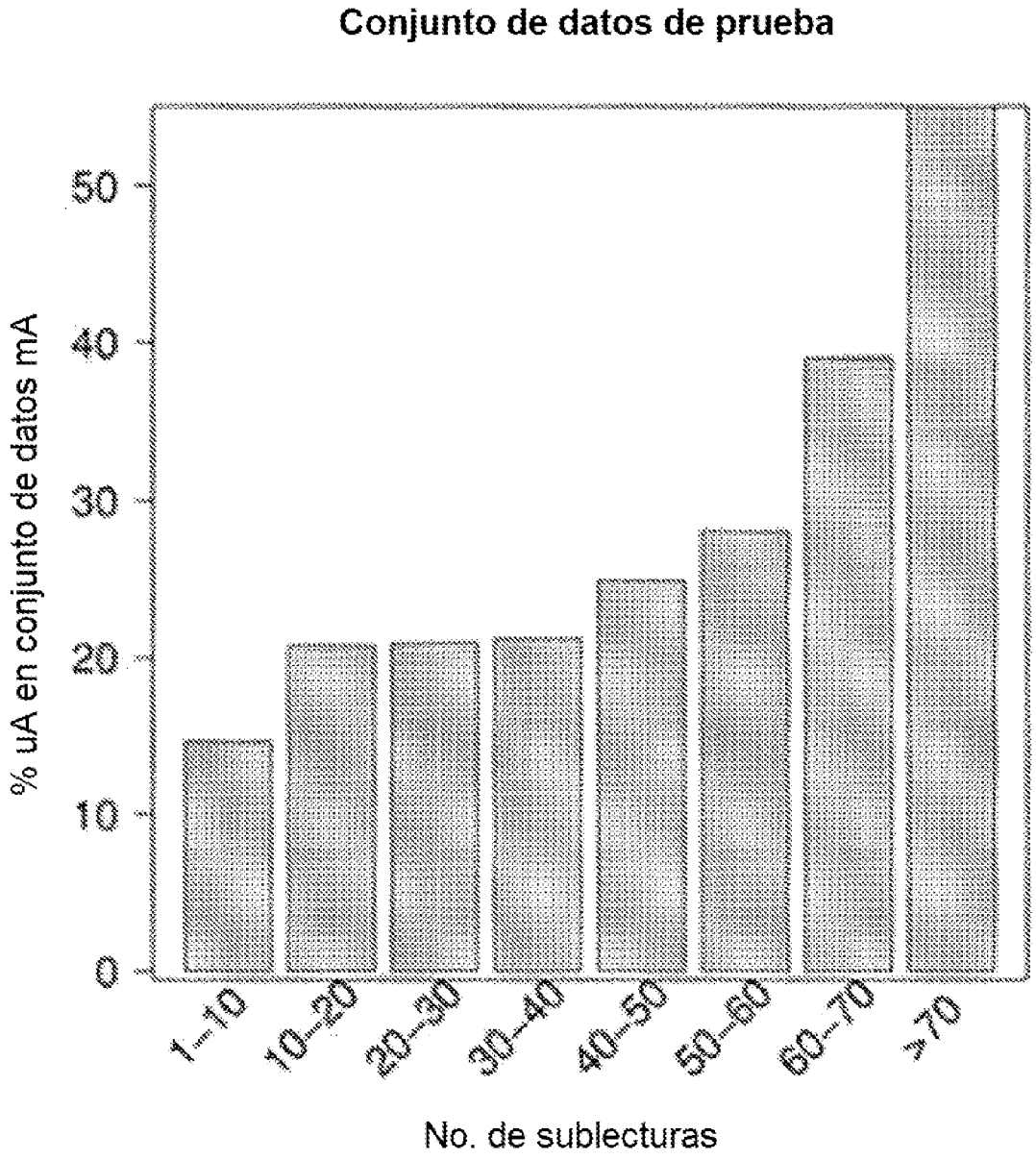

Conjunto de datos de prueba
% uA en conjunto de datos mA
0
10
20
30
40
50
1–10
10–20
20–30
30–40
40–50
50–60
60–70
>70
No. de sublecturas

FIG. 43

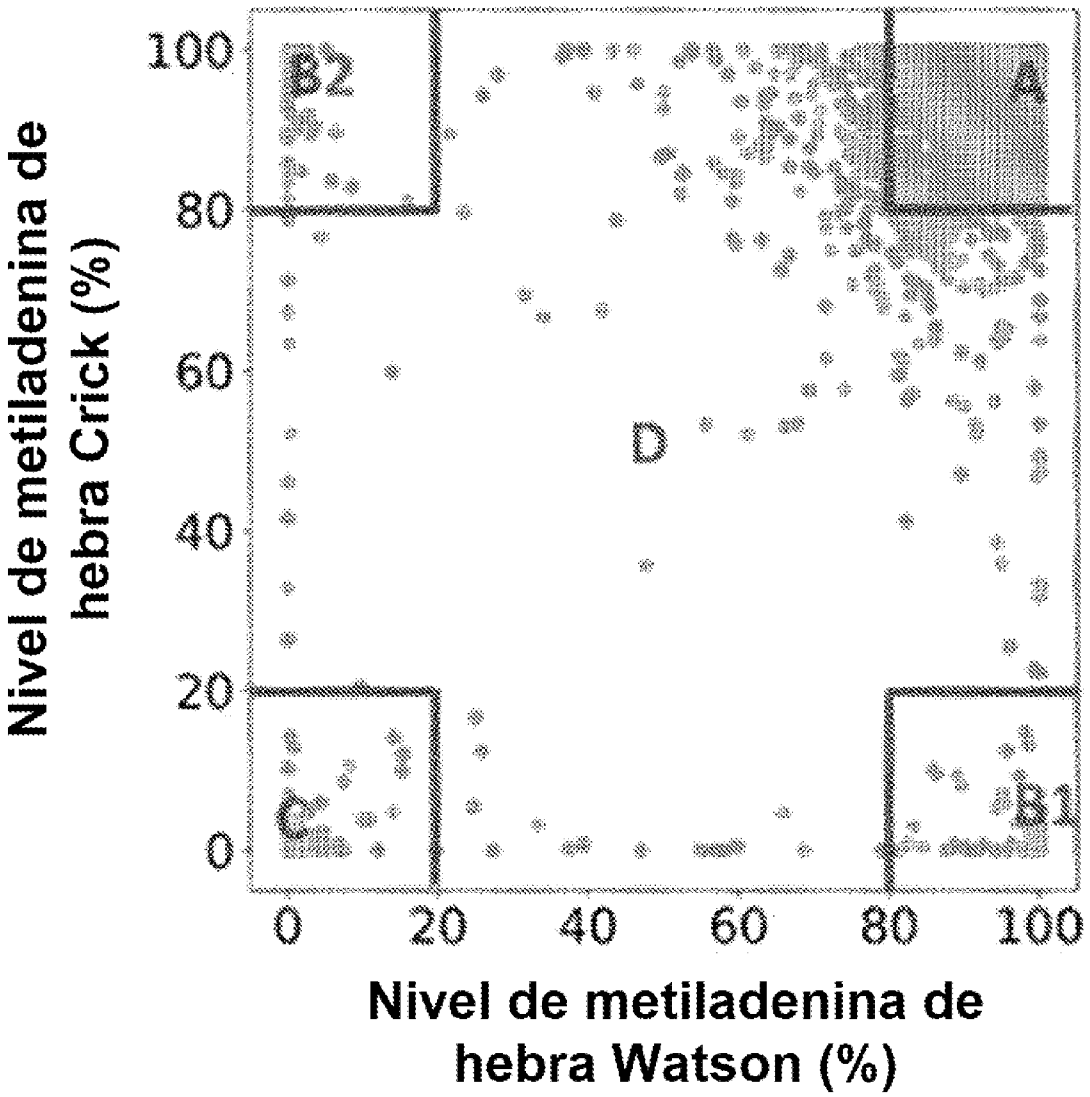

Nivel de metiladenina de hebra Crick (%)
100
80
60
40
20
0
B2
A
D
C
B1
0
20
40
60
80
100
Nivel de metiladenina de hebra Watson (%)

FIG.44

Categorías	Conjunto de datos de entrenamiento	Conjunto de datos de prueba
Completamente no metilado	283 (7.0%)	276 (7.0%)
Hemimetilado	401 (10.0%)	389 (9.8%)
Completamente metilado	3194 (79.4%)	3142 (79.4%)
Patrones de metilación de entrecruzamiento	145 (3.6%)	148 (3.7%)

FIG. 45

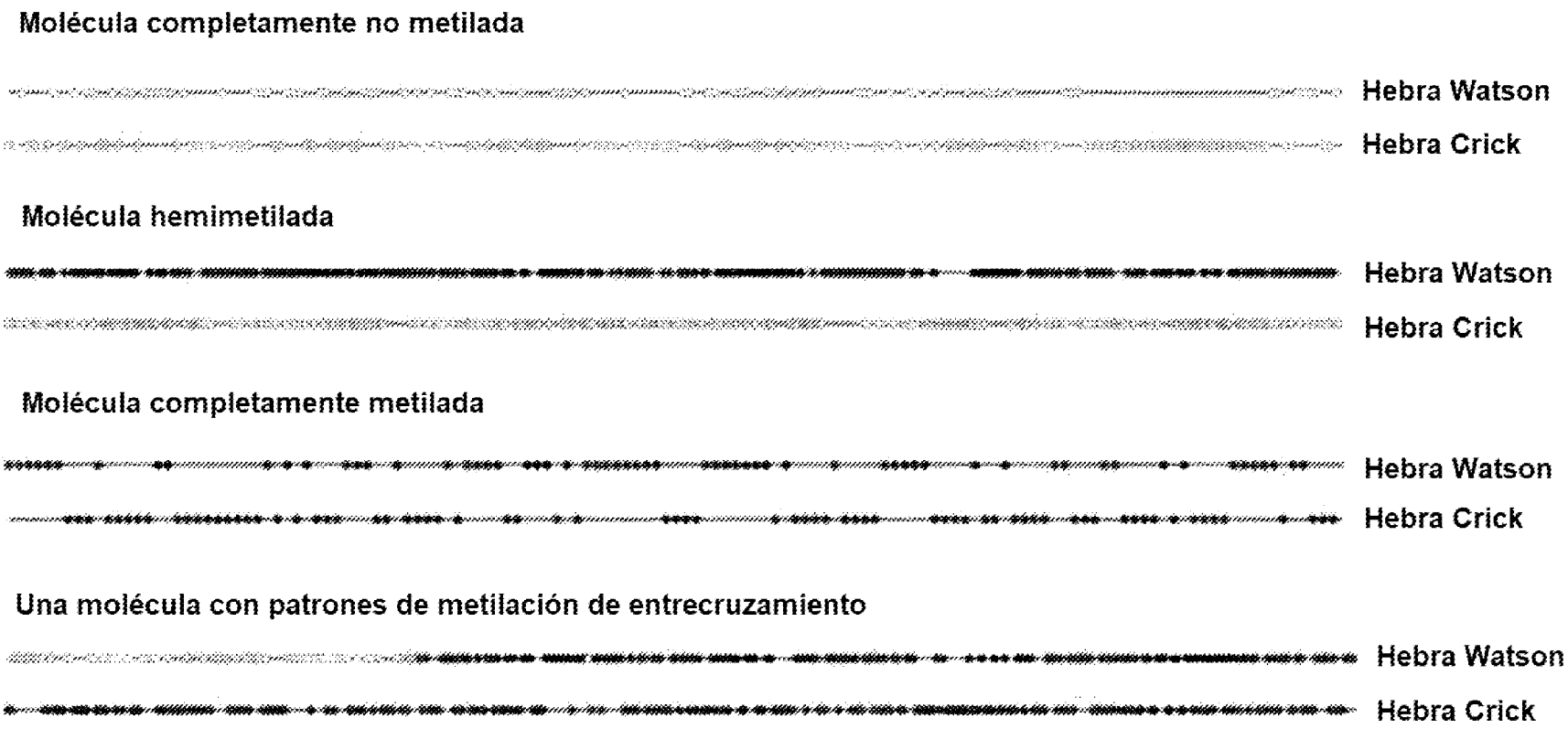

Molécula completamente no metilada
Hebra Watson
Hebra Crick
Molécula hemimetilada
Hebra Watson
Hebra Crick
Molécula completamente metilada
Hebra Watson
Hebra Crick
Una molécula con patrones de metilación de entrecruzamiento
Hebra Watson
Hebra Crick

FIG. 46

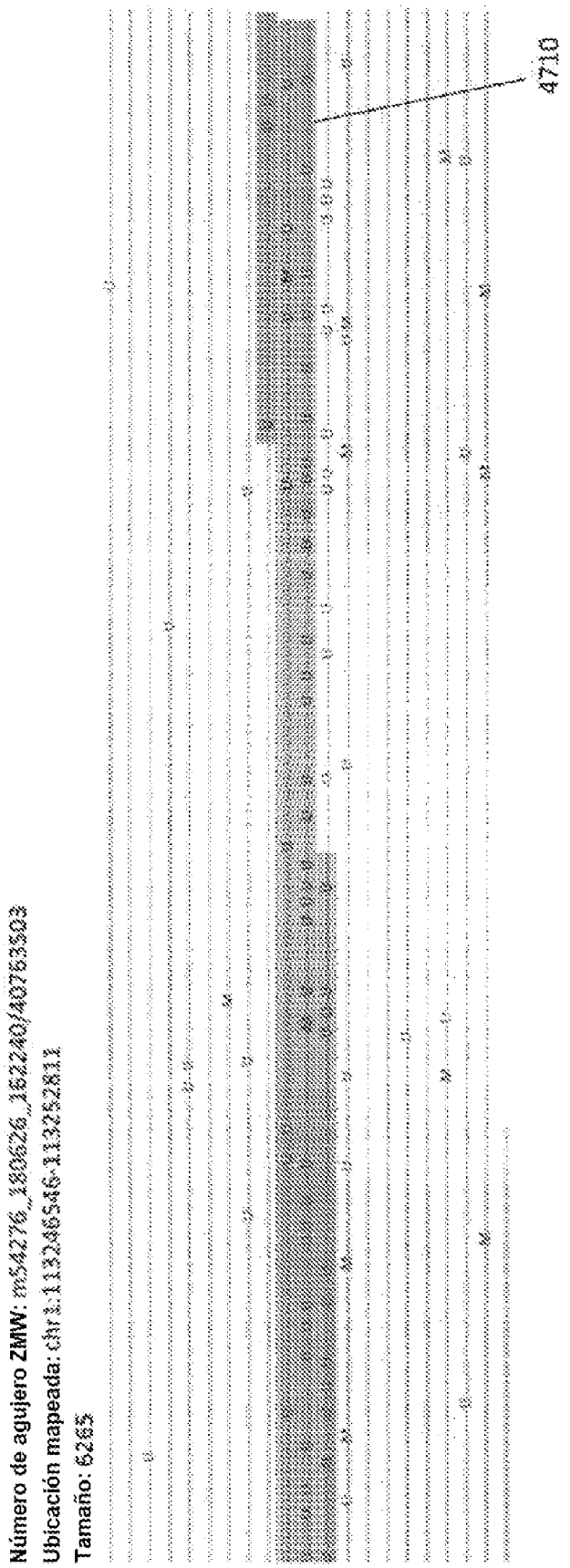
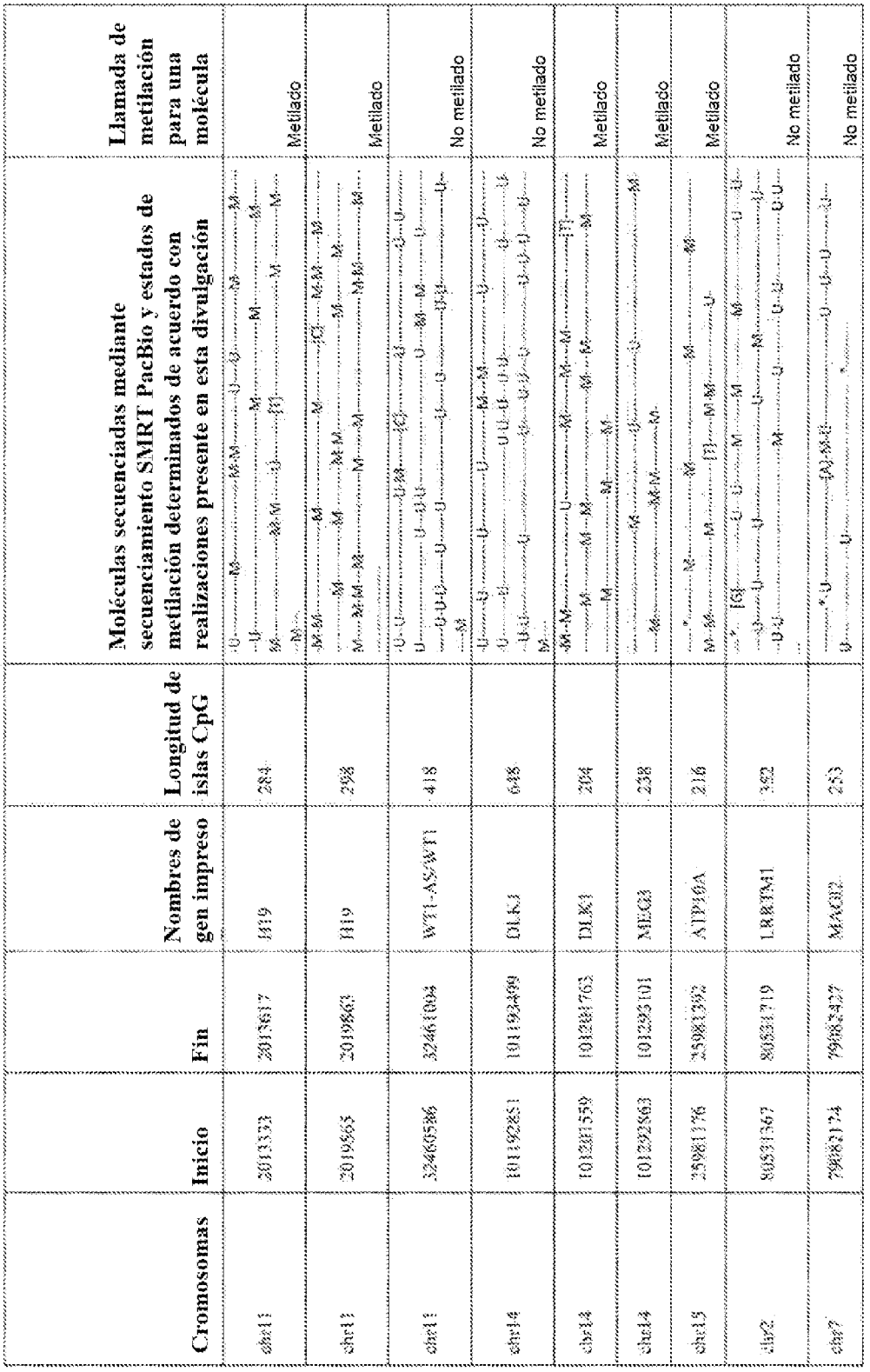

Número de agujero ZMW: m54276_180626_162240/40763503
Ubicación mapeada: chr3:113246546-113252811
Tamaño: 6265
4710

FIG. 47



Cromosomas	Inicio	Fin	Nombres de gen impreso	Longitud de islas CpG	Moléculas secuenciadas mediante secuenciamiento SMRT PacBio y estados de metilación determinados de acuerdo con realizaciones presente en esta divulgación	Llamada de metilación para una molécula
chr11	2011333	2013617	H19	284		Metilado
chr11	2019563	2019863	H19	298		Metilado
chr11	324605586	324611004	WT1-AS/WT1	418		No metilado
chr14	101192851	101192499	DLK1	648		No metilado
chr14	101281559	101281763	DLK1	204		Metilado
chr14	101292863	101293101	MEG3	238		Metilado
chr15	25981176	25983392	ATP10A	216		Metilado
chr2	80531367	80533719	LRRTM1	352		No metilado
chr7	29082174	29083427	MAC82	253		No metilado

FIG. 48

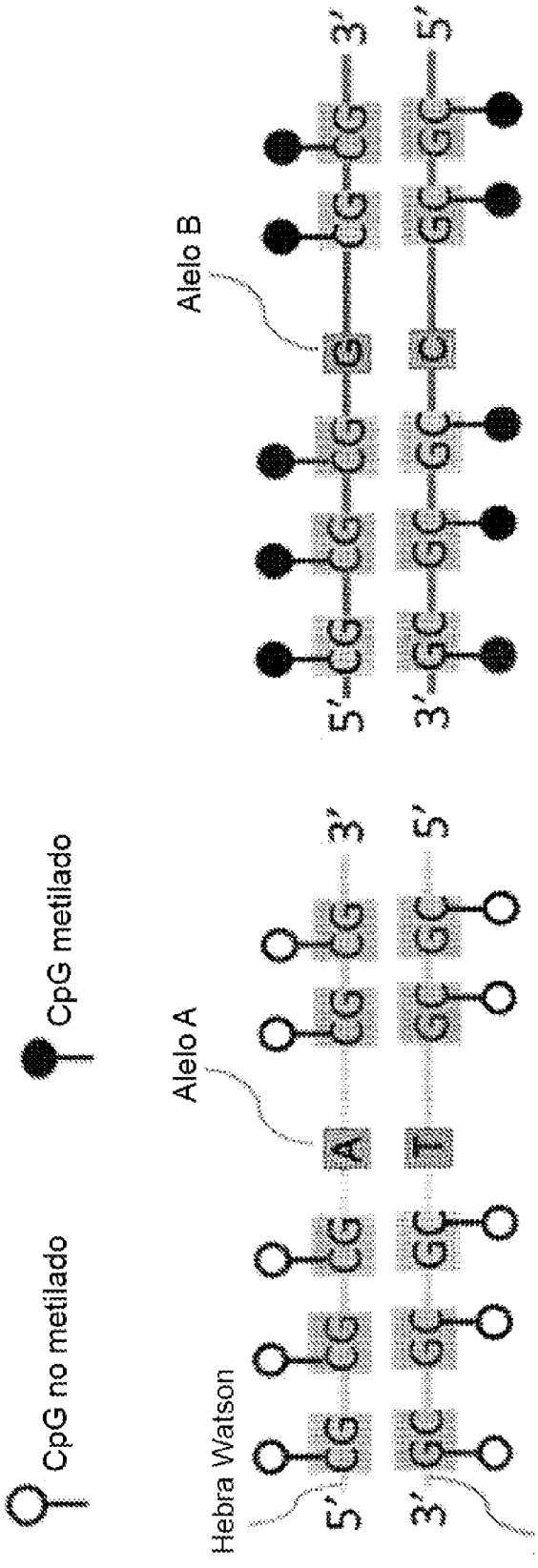

Patrones de metilación presentes en una región paternalmente impresa
CpG no metilado
CpG metilado
Alelo A
Alelo B
Hebra Watson
Hebra Crick
5'
3'
Haplotipo materno
Haplotipo paterno

FIG. 49

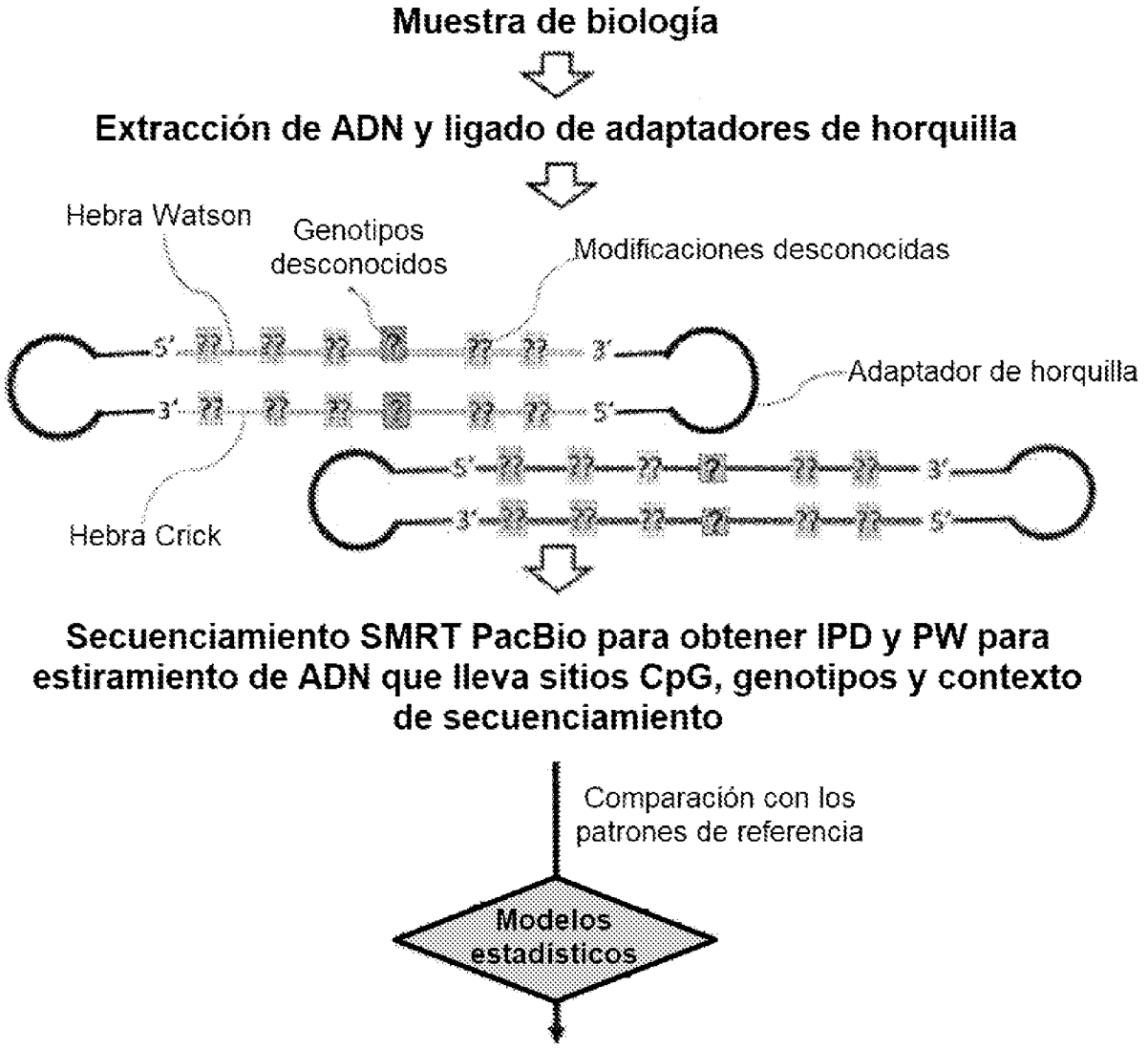

Muestra de biología
Extracción de ADN y ligado de adaptadores de horquilla
Hebra Watson
Genotipos desconocidos
Modificaciones desconocidas
Adaptador de horquilla
Hebra Crick
Secuenciamiento SMRT PacBio para obtener IPD y PW para estiramiento de ADN que lleva sitios CpG, genotipos y contexto de secuenciamiento
Comparación con los patrones de referencia
Modelos estadísticos
Clasificación de estados de metilación junto con haplotipos

FIG. 50

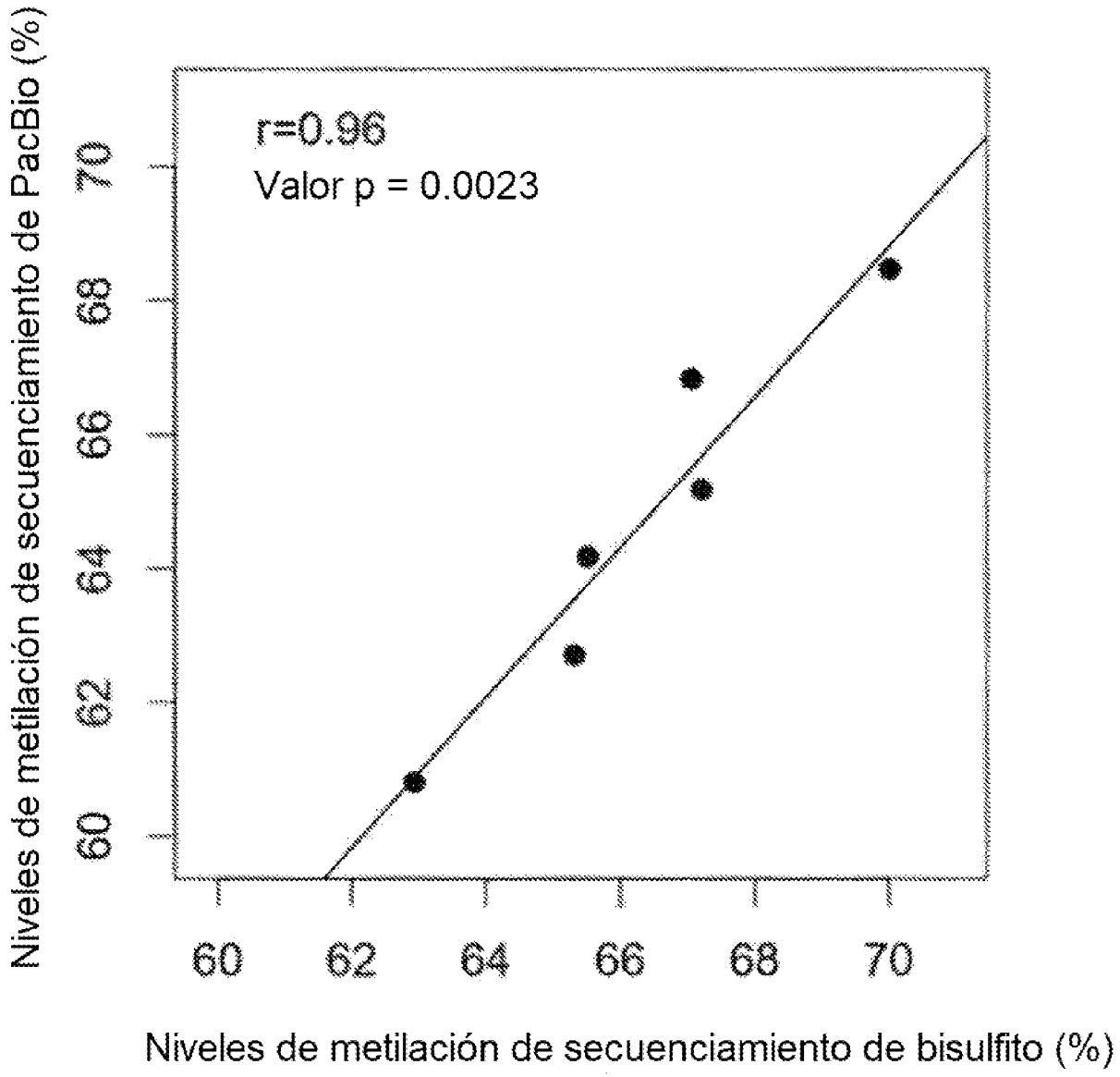

r=0.96
Valor p = 0.0023
Niveles de metilación de secuenciamiento de PacBio (%)
60
62
64
66
68
70
Niveles de metilación de secuenciamiento de bisulfito (%)

FIG. 51

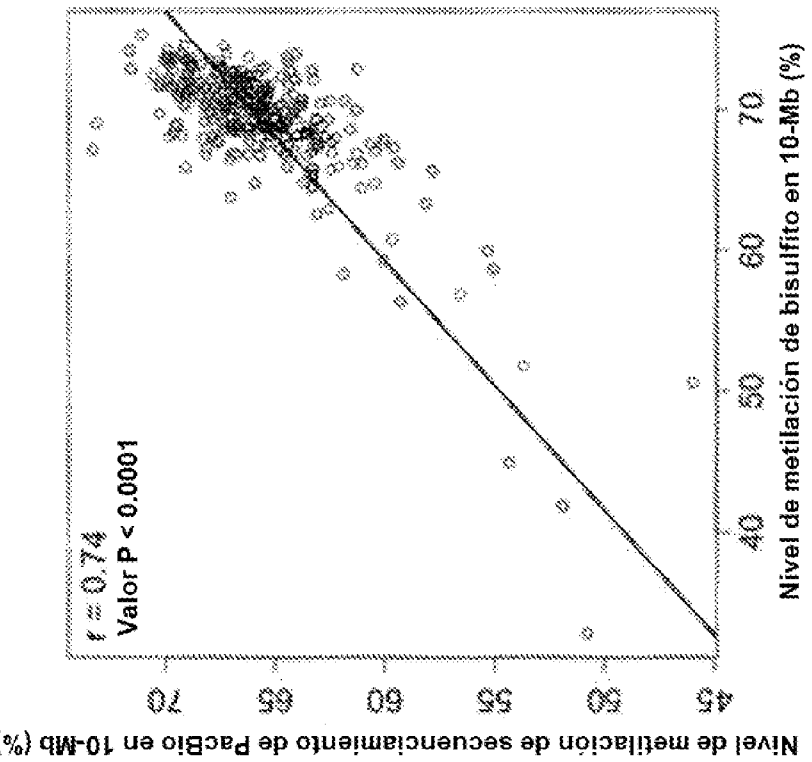


FIG. 52B

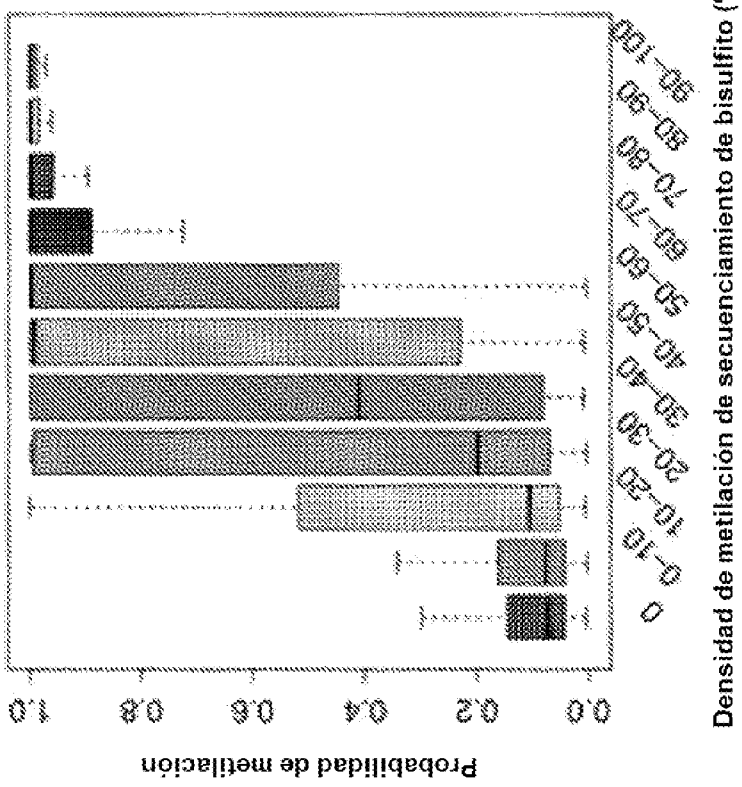


FIG. 52A

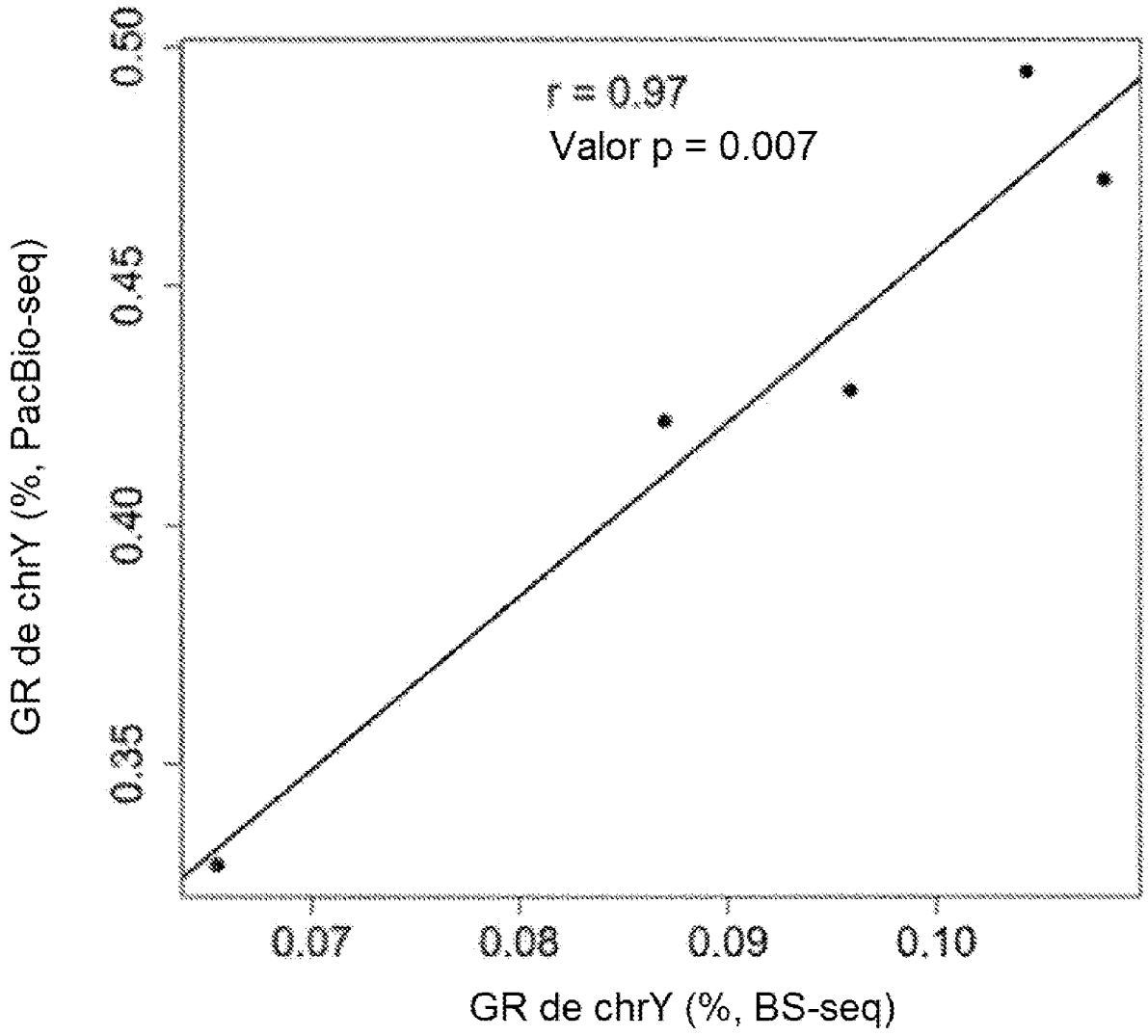

r = 0.97
Valor p = 0.007
GR de chrY (%, PacBio-seq)
0.35
0.40
0.45
0.50
0.07
0.08
0.09
0.10
GR de chrY (%, BS-seq)

FIG. 53

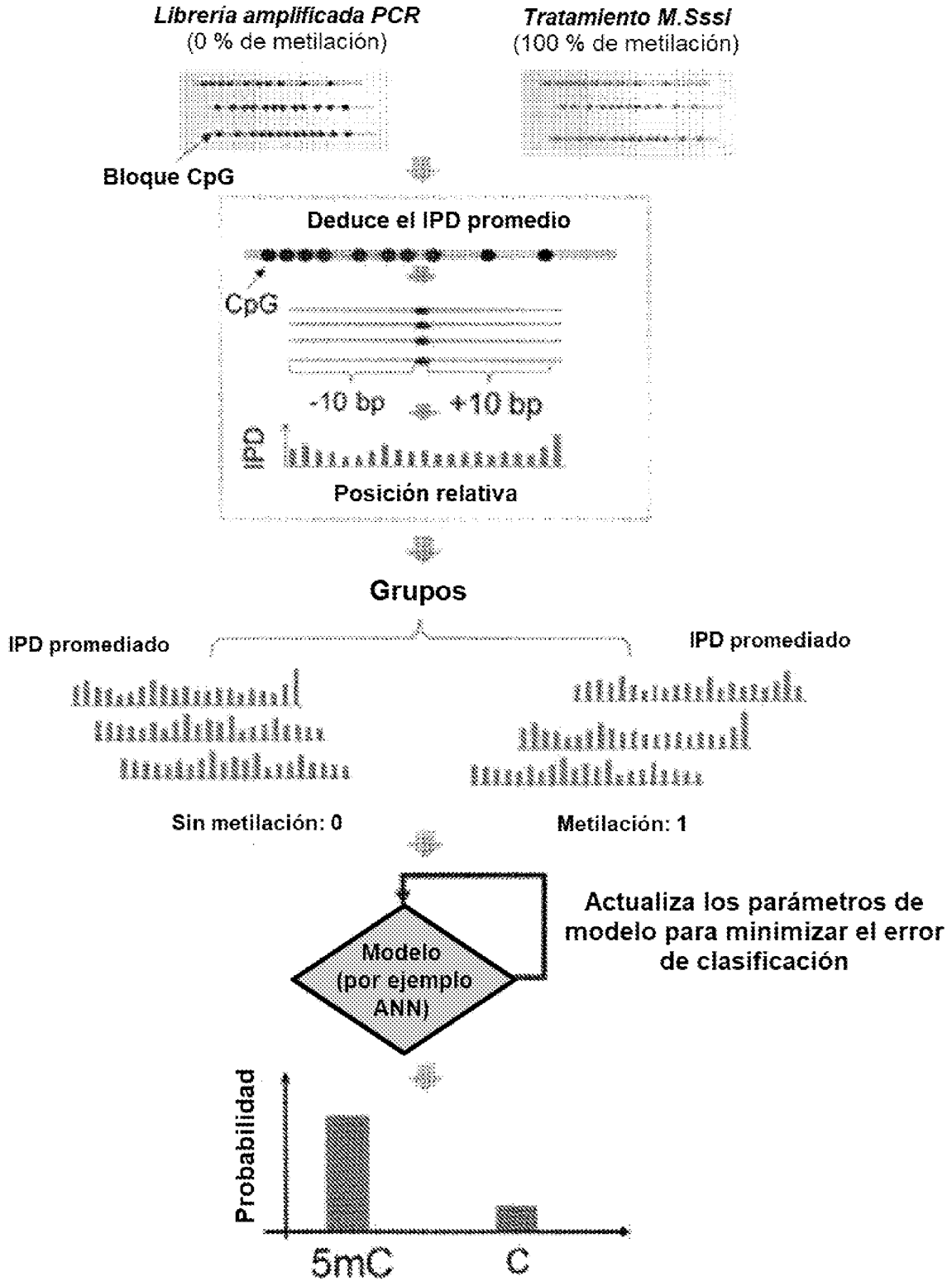

Librería amplificada PCR
(0 % de metilación)
Tratamiento M.SssI
(100 % de metilación)
Bloque CpG
Deduce el IPD promedio
CpG
-10 bp
+10 bp
IPD
Posición relativa
Grupos
IPD promediado
IPD promediado
Sin metilación: 0
Metilación: 1
Modelo
(por ejemplo
ANN)
Actualiza los parámetros de
modelo para minimizar el error
de clasificación
Probabilidad
5mC
C

FIG. 54

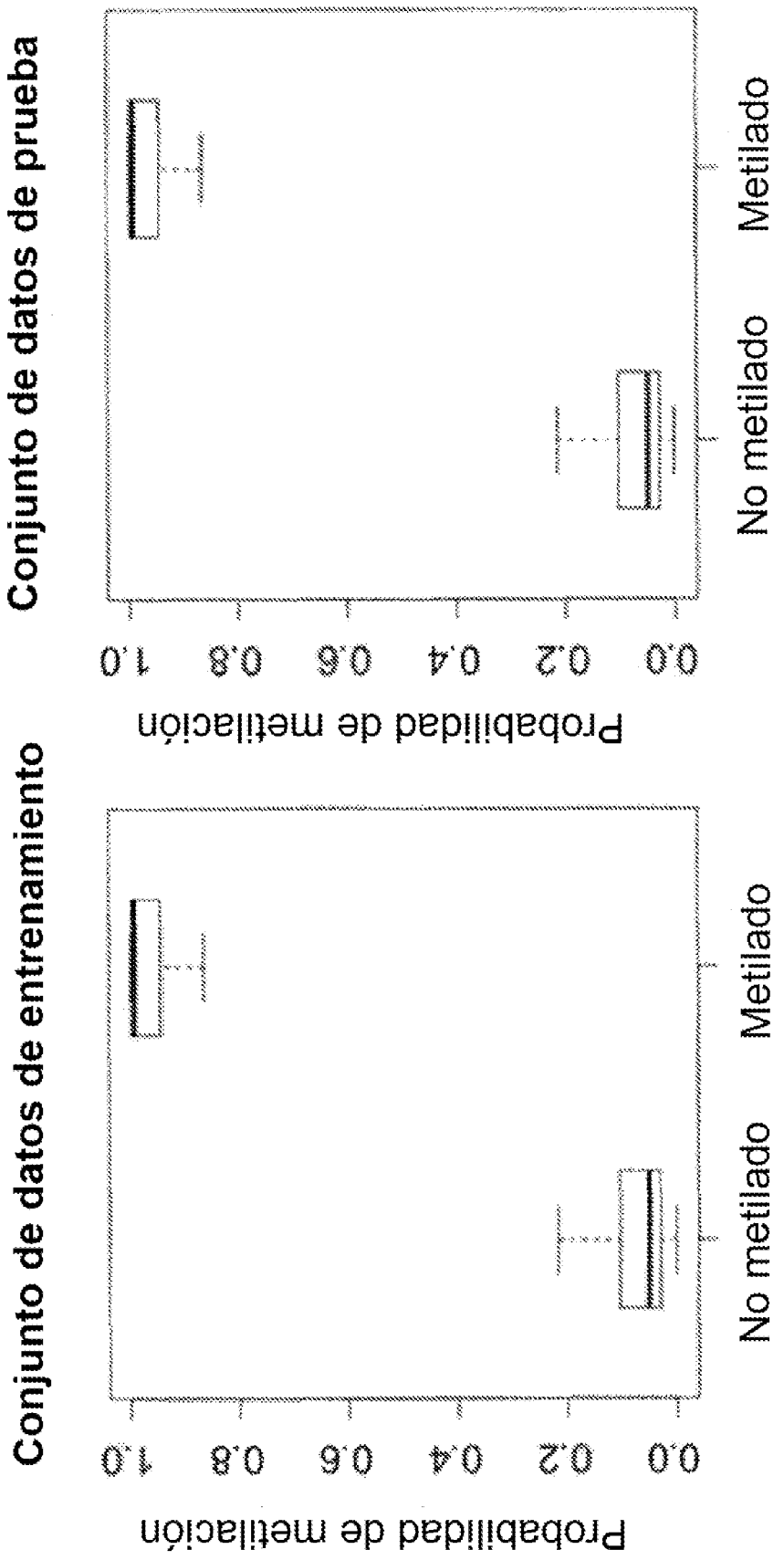


FIG. 55A

FIG. 55B

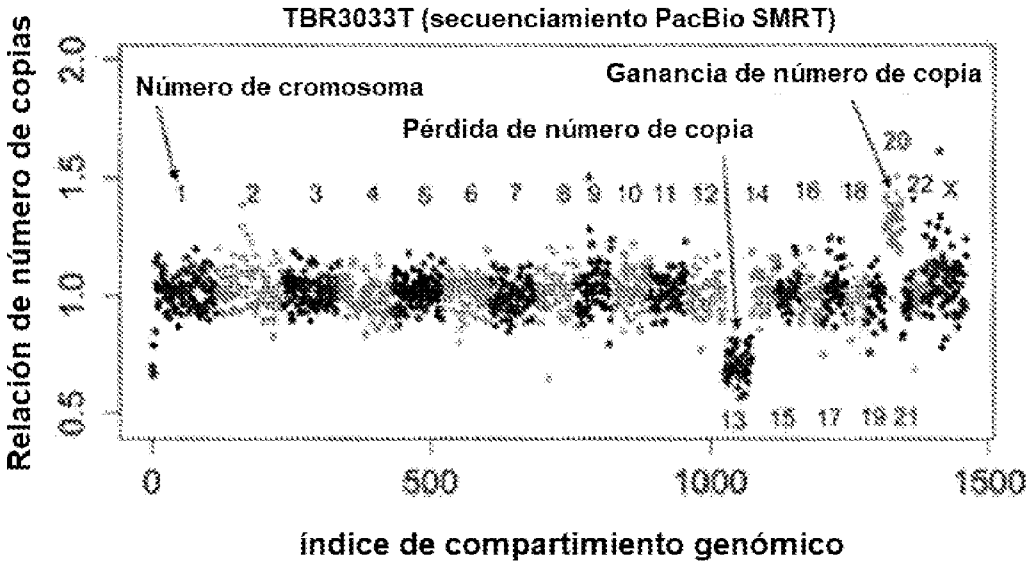


FIG. 56A

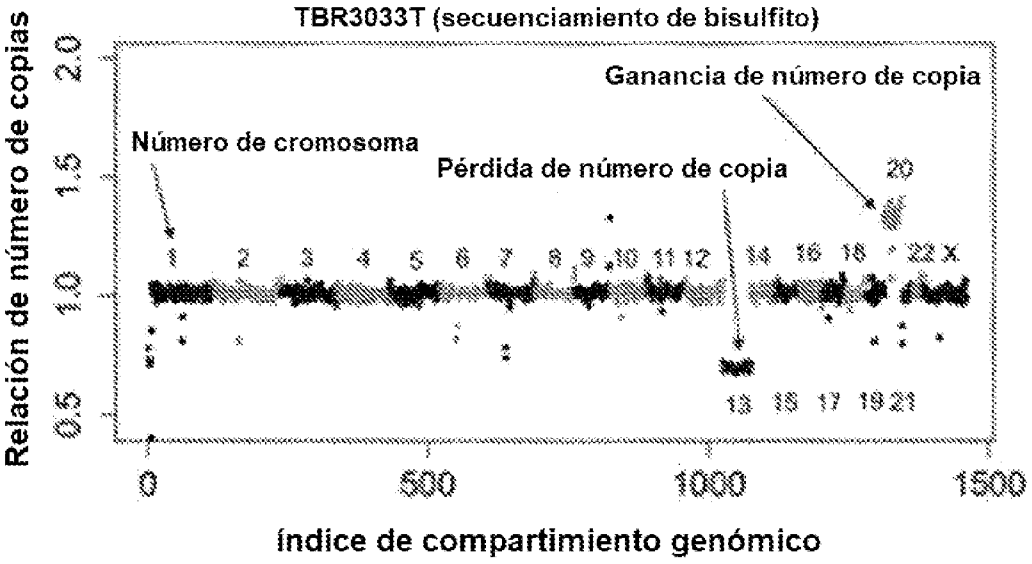


FIG. 56B

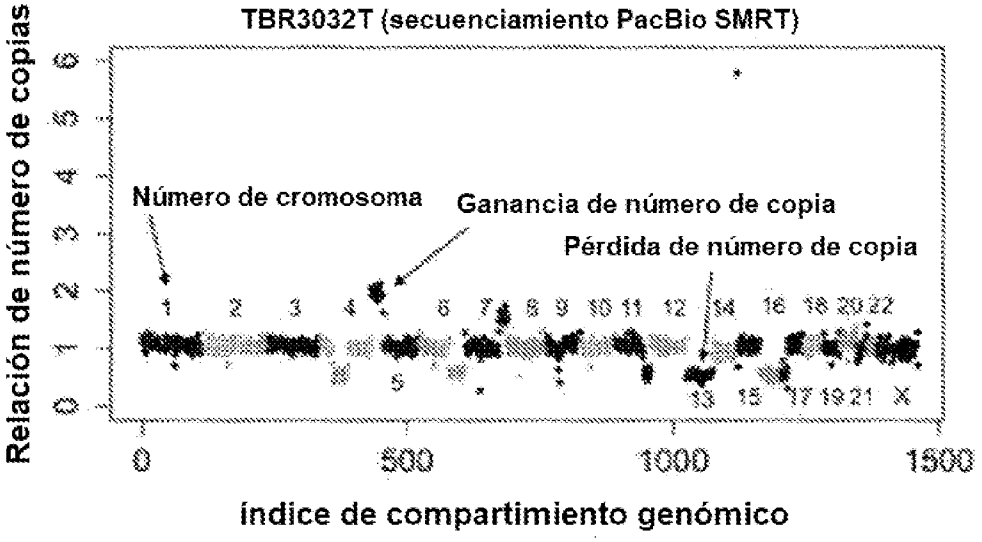


FIG. 57A

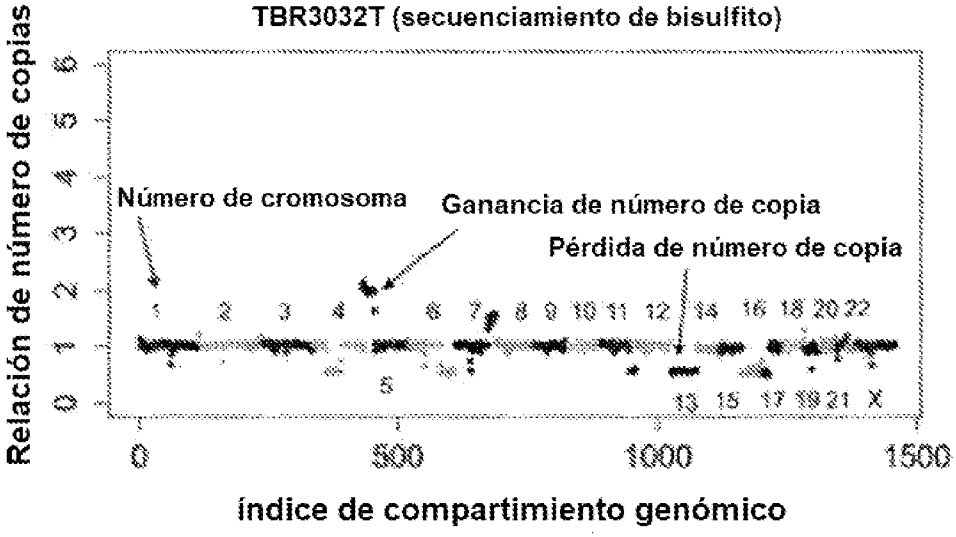


FIG. 57B

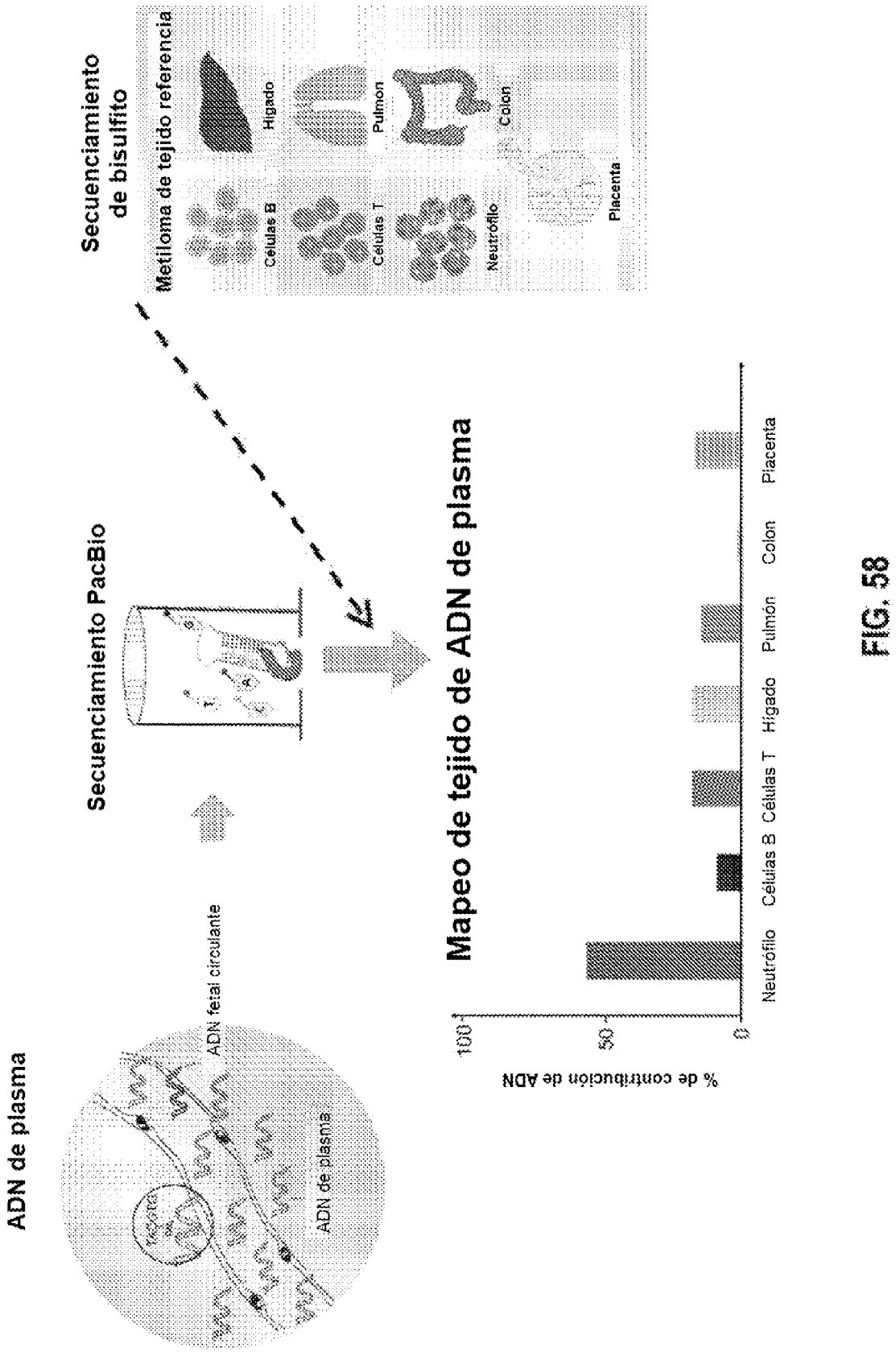

ADN de plasma
ADN fetal circulante
ADN de plasma
Secuenciamiento PacBio
Secuenciamiento de bisulfito
Metiloma de tejido referencia
Células B
Células T
Neutrófilo
Hígado
Pulmon
Colon
Placenta
Mapeo de tejido de ADN de plasma
% de contribución de ADN
100
50
0
Neutrófilo
Células B
Células T
Hígado
Pulmón
Colon
Placenta

FIG. 58

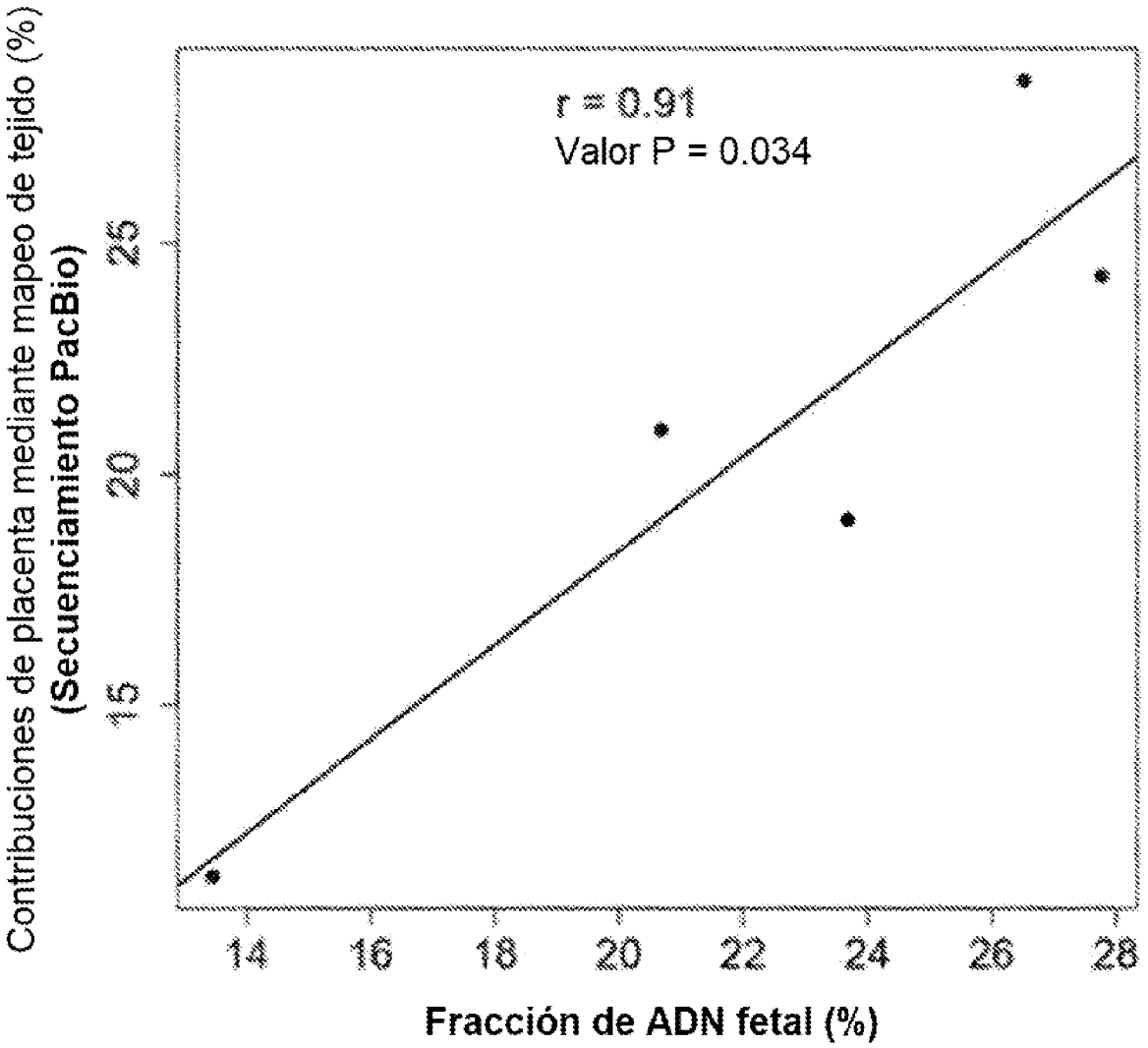

r = 0.91
Valor P = 0.034
Contribuciones de placenta mediante mapeo de tejido (%)
(Secuenciamiento PacBio)
15
20
25
14
16
18
20
22
24
26
28
Fracción de ADN fetal (%)

FIG. 59

Grupos	Muestras	Total sublecturas	Sublecturas mapeadas	Mapeabilidad de sublecturas (%)	Profundidad media de sublectura por pocillo SMRT (x)	Número de pocillos SMRT	Pocillos mapeables	Tasa de pocillos mapeables (%)
Capa leucocitaria materna	[illegible]	38,006,480	30,673,526	78.6	13.4	3,157,310	2,266,002	72.7
Placenta	N13153	23,013,428	16,374,758	71.2	10.4	2,363,400	1,573,546	65.7
Tejidos HCC	TBR30827	20,164,513	15,232,744	75.5	13.1	1,742,999	1,147,985	64.8
	TBR30831	23,639,692	17,479,024	77.2	8.1	2,932,627	2,157,196	78.2
Tejidos normales adyacentes	TBR30830N	23,118,110	20,446,282	77.2	12.6	6,881,142	4,471,370	65.0
	TBR30832N	76,852,680	60,145,452	78.3	13.8	8,000,227	4,702,130	78.4
Capa leucocitaria (sujetos de control saludables)	M1	44,777,423	28,325,587	63.3	7.7	7,316,009	3,659,999	50.0
	F2	49,840,758	32,994,645	66.2	8.6	7,215,112	3,823,329	53.0
	F1	40,012,804	24,717,269	61.8	8.5	7,301,766	3,800,392	52.0
	M2	152,530,411	86,506,520	56.1	7.7	21,794,636	11,563,530	53.1
Estirpe de células HCC	HepG2	47,306,982	34,581,721	73.1	7.3	6,220,059	4,750,381	76.4

FIG. 60

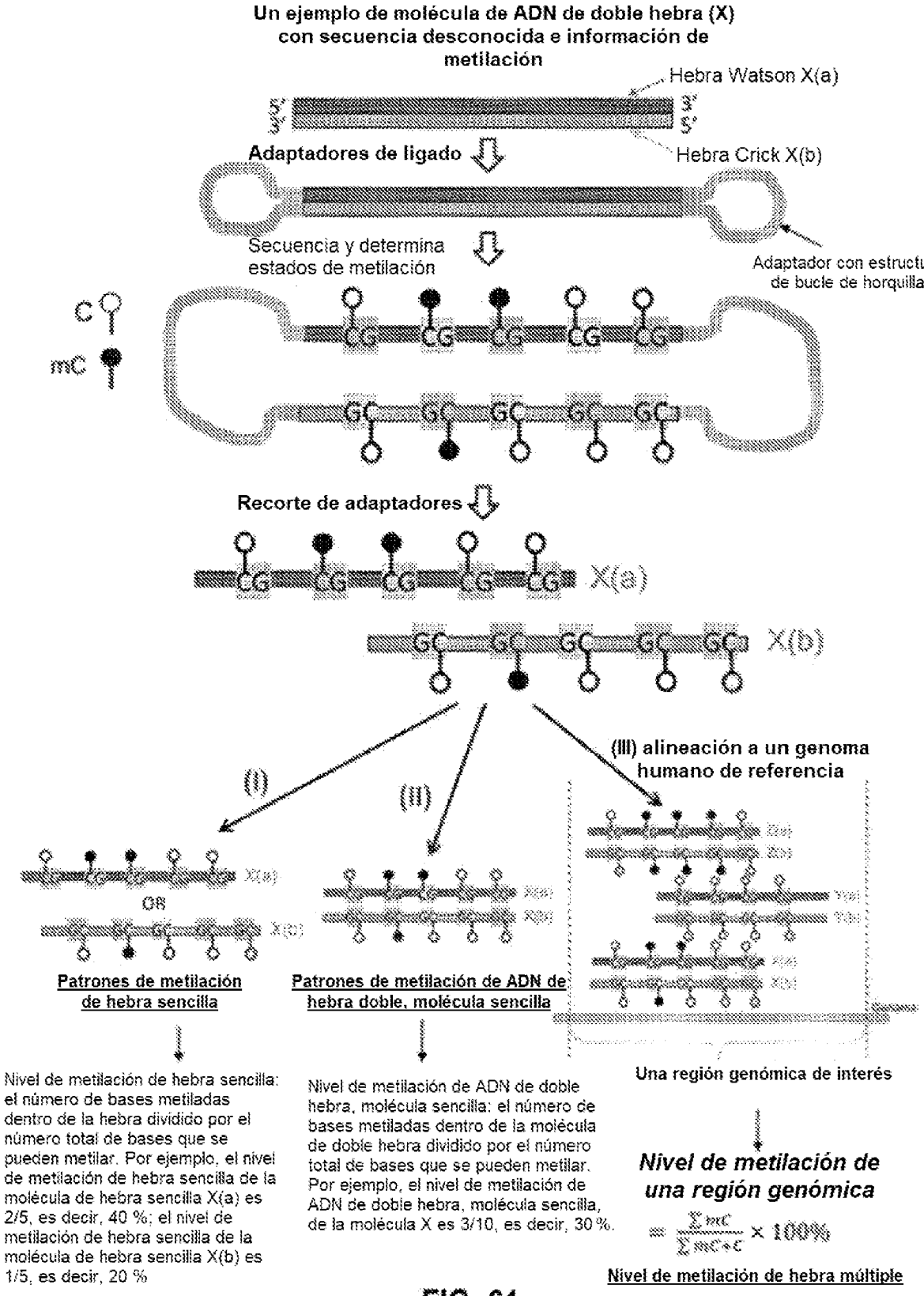

Un ejemplo de molécula de ADN de doble hebra (X) con secuencia desconocida e información de metilación
Hebra Watson X(a)
Hebra Crick X(b)
Adaptadores de ligado
Secuencia y determina estados de metilación
Adaptador con estructura de bucle de horquilla
C
mC
Recorte de adaptadores
X(a)
X(b)
(I)
(II)
(III) alineación a un genoma humano de referencia
OR
Patrones de metilación de hebra sencilla
Patrones de metilación de ADN de hebra doble, molécula sencilla
Una región genómica de interés
Nivel de metilación de hebra sencilla: el número de bases metiladas dentro de la hebra dividido por el número total de bases que se pueden metilar. Por ejemplo, el nivel de metilación de hebra sencilla de la molécula de hebra sencilla X(a) es 2/5, es decir, 40 %; el nivel de metilación de hebra sencilla de la molécula de hebra sencilla X(b) es 1/5, es decir, 20 %
Nivel de metilación de ADN de doble hebra, molécula sencilla: el número de bases metiladas dentro de la molécula de doble hebra dividido por el número total de bases que se pueden metilar. Por ejemplo, el nivel de metilación de ADN de doble hebra, molécula sencilla, de la molécula X es 3/10, es decir, 30 %.
Nivel de metilación de una región genómica
$= \frac{\sum mC}{\sum mC+C} \times 100\%$
Nivel de metilación de hebra múltiple

FIG. 61

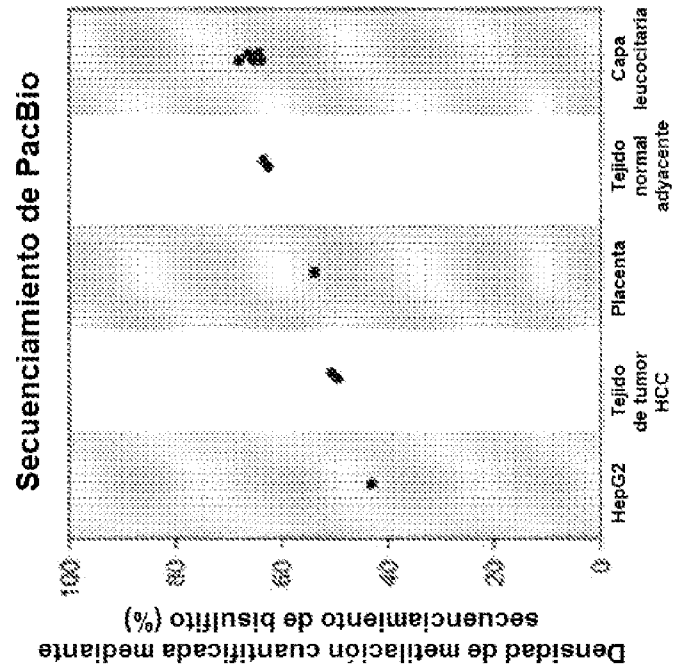


FIG. 62B

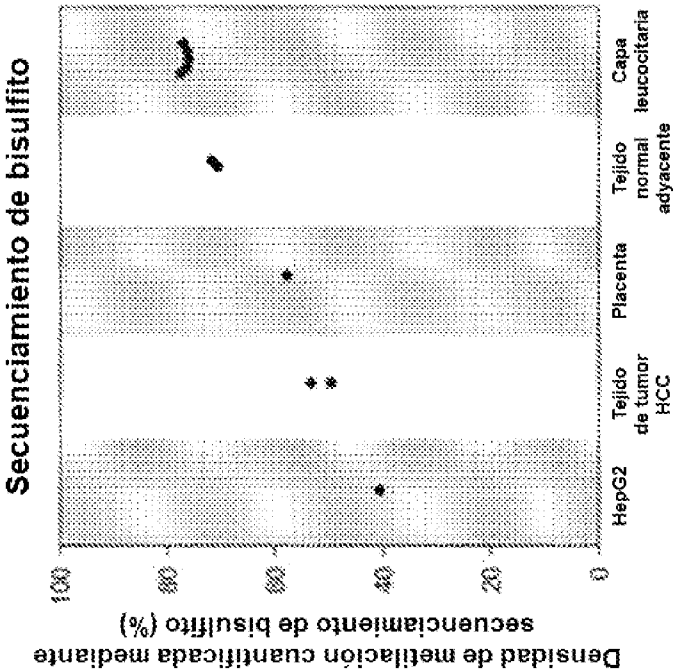


FIG. 62A

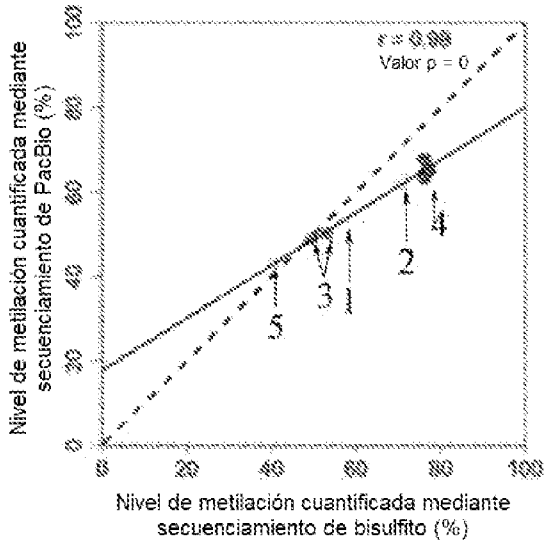


FIG. 63A

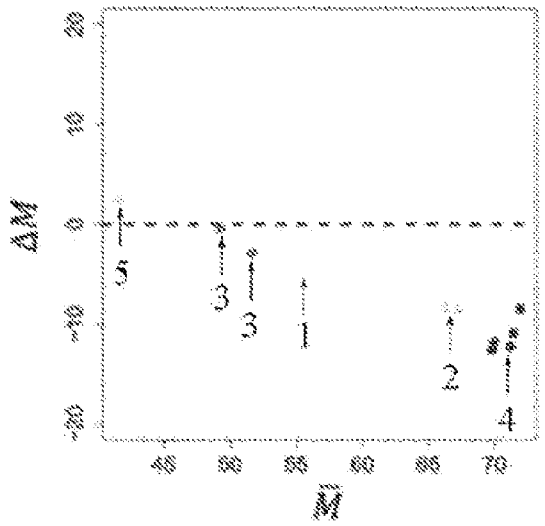
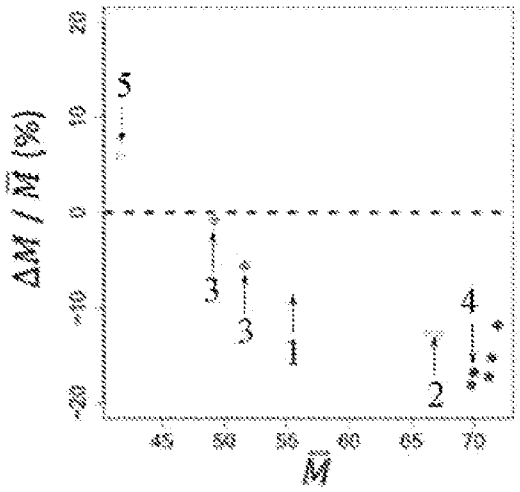


FIG. 63B



1 Placenta
2 Tejido normal adyacente
3 Tejido de tumor HCC
4 Capa leucocitaria
5 HepG2 (estirpe de células HCC)

FIG. 63C

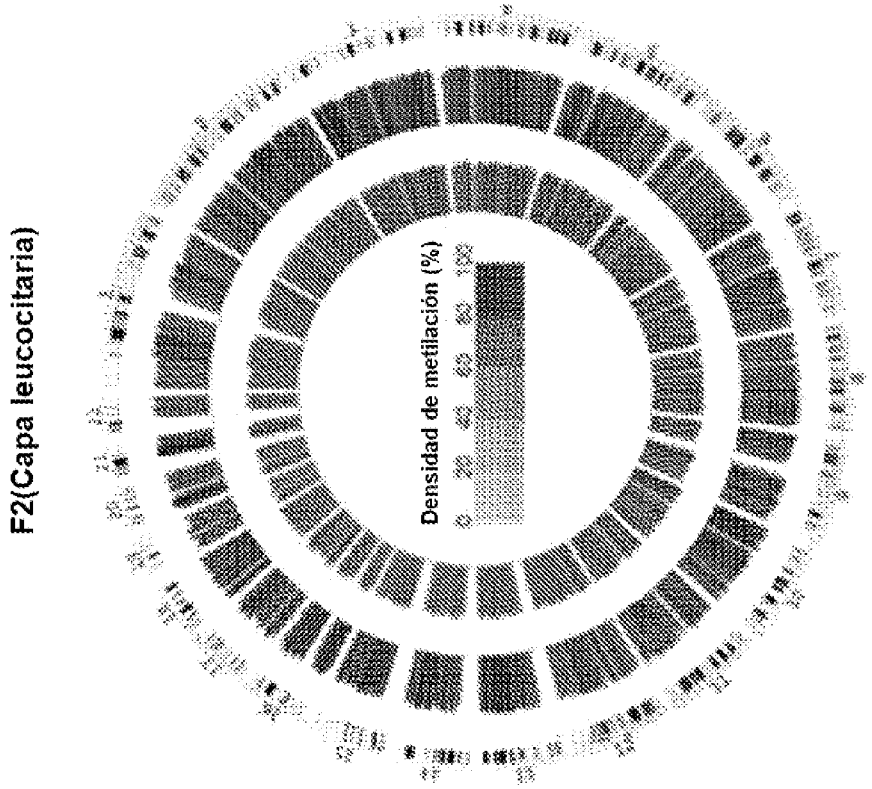


FIG. 64B

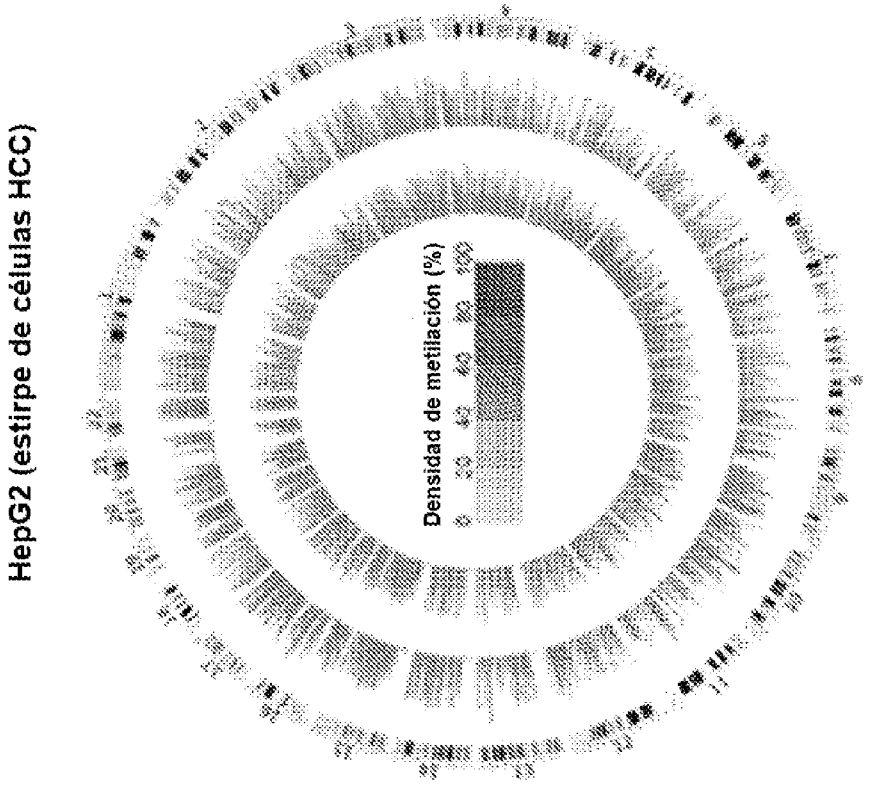


FIG. 64A

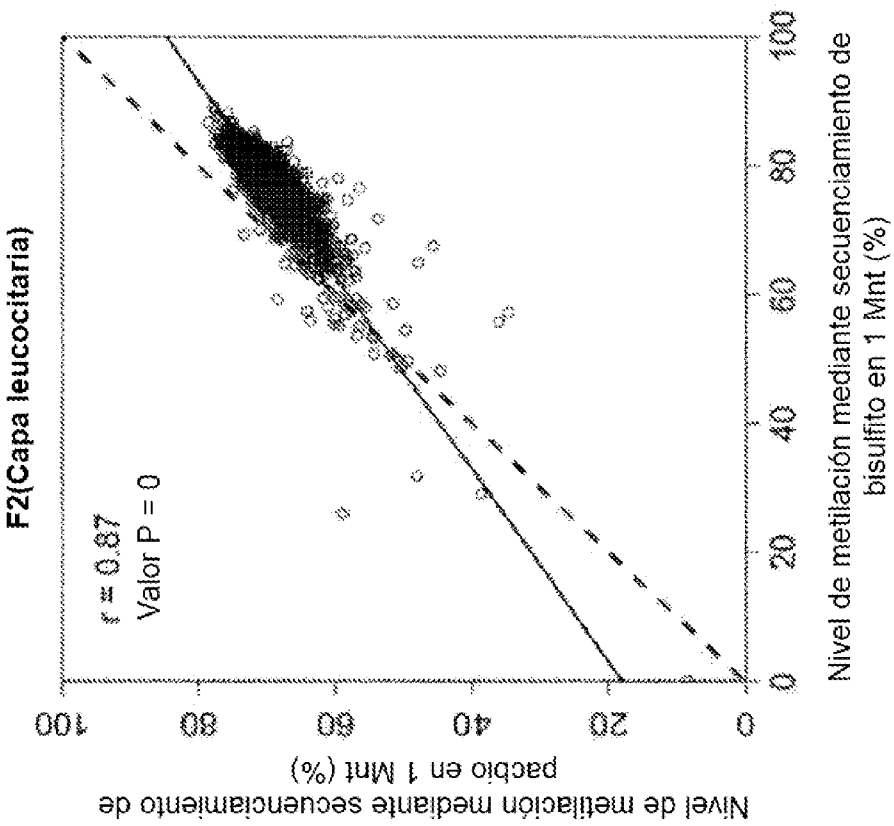


FIG. 65B

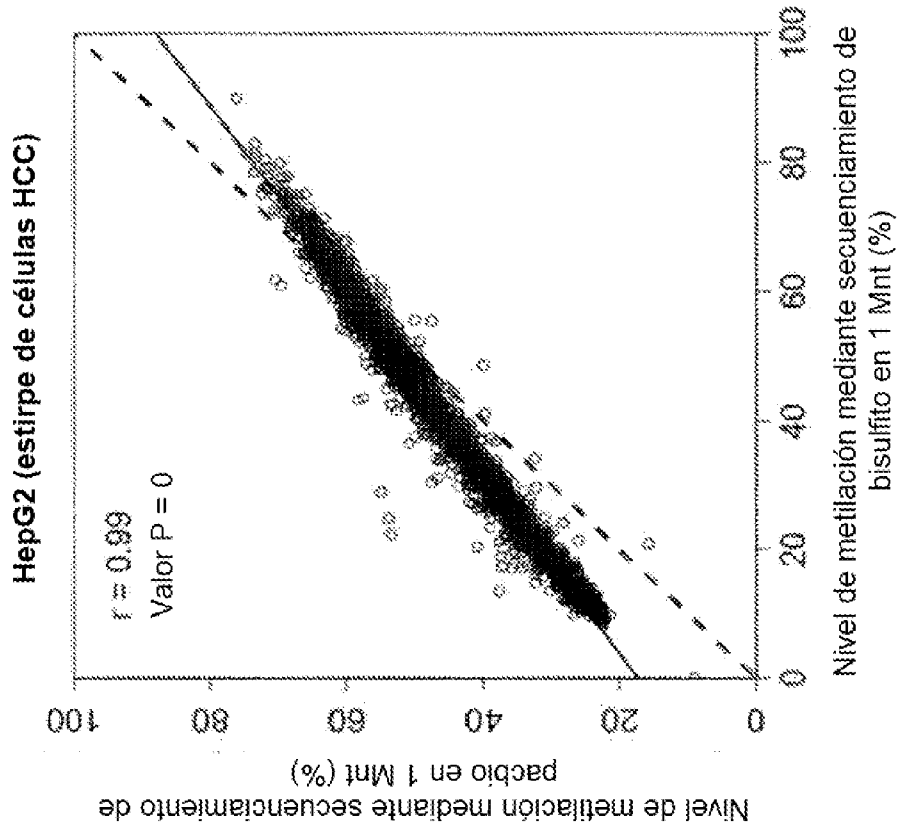


FIG. 65A

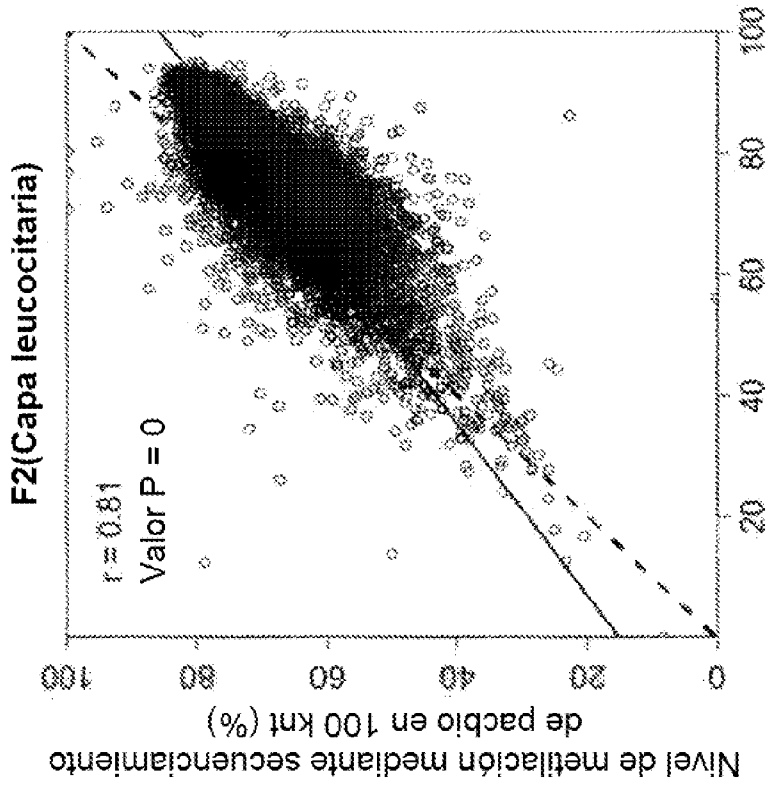


FIG. 66B

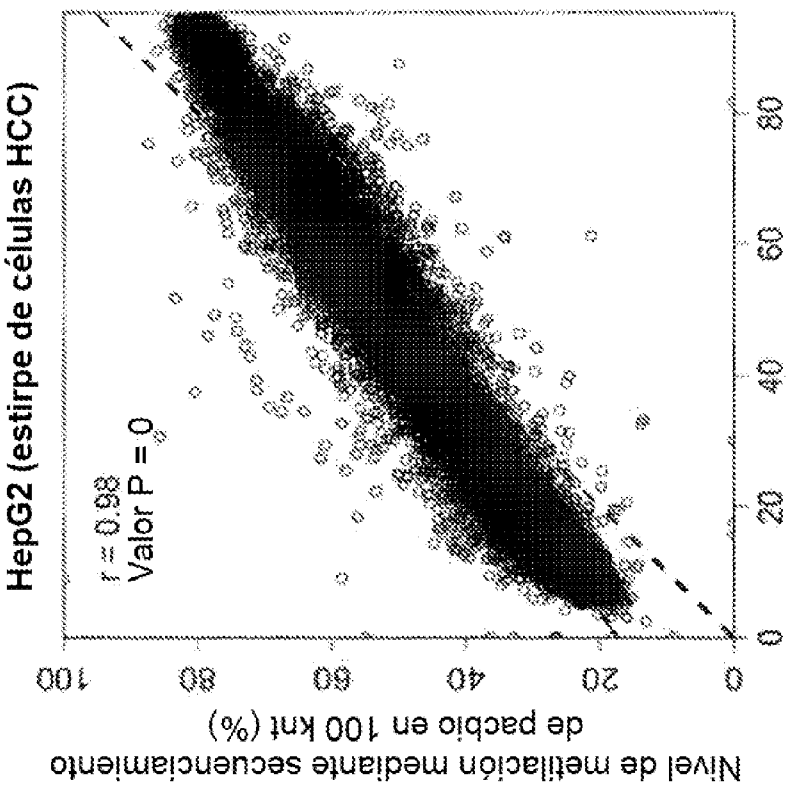


FIG. 66A

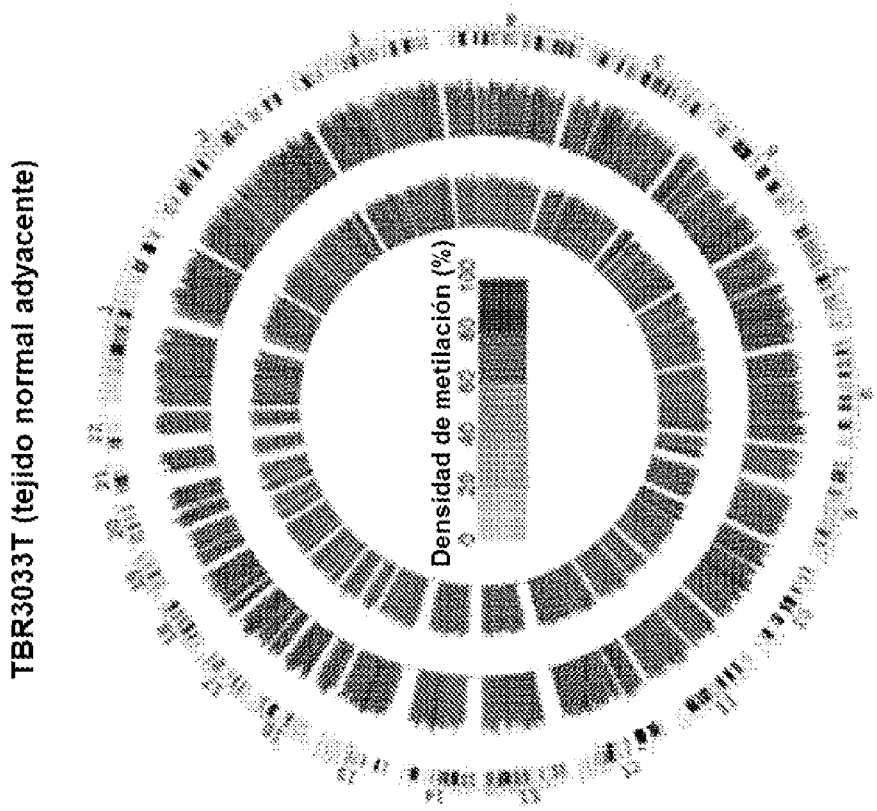


FIG. 67B

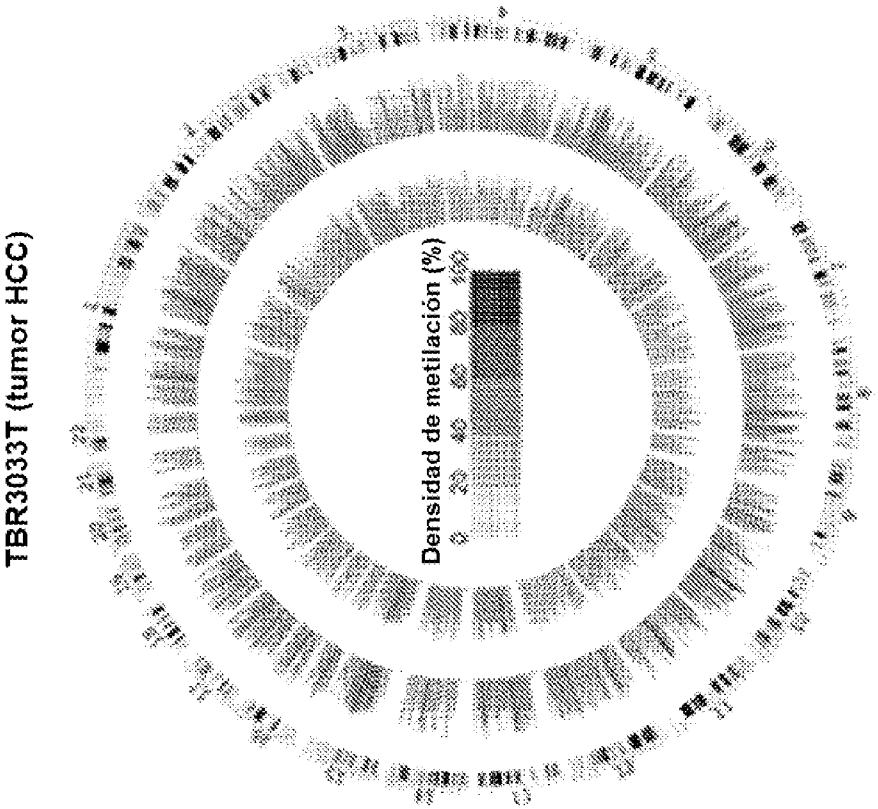


FIG. 67A

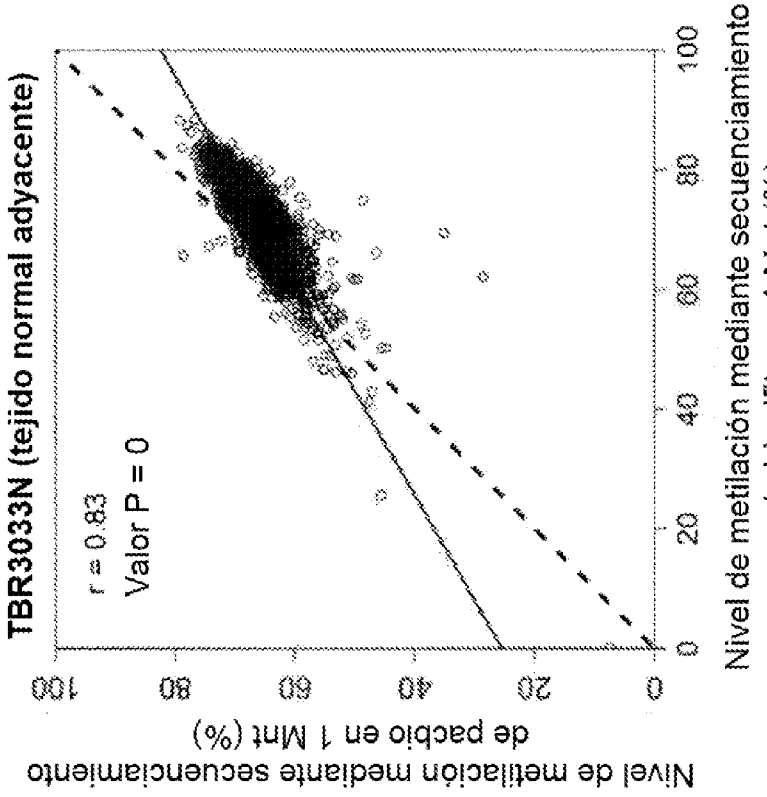


FIG. 68B

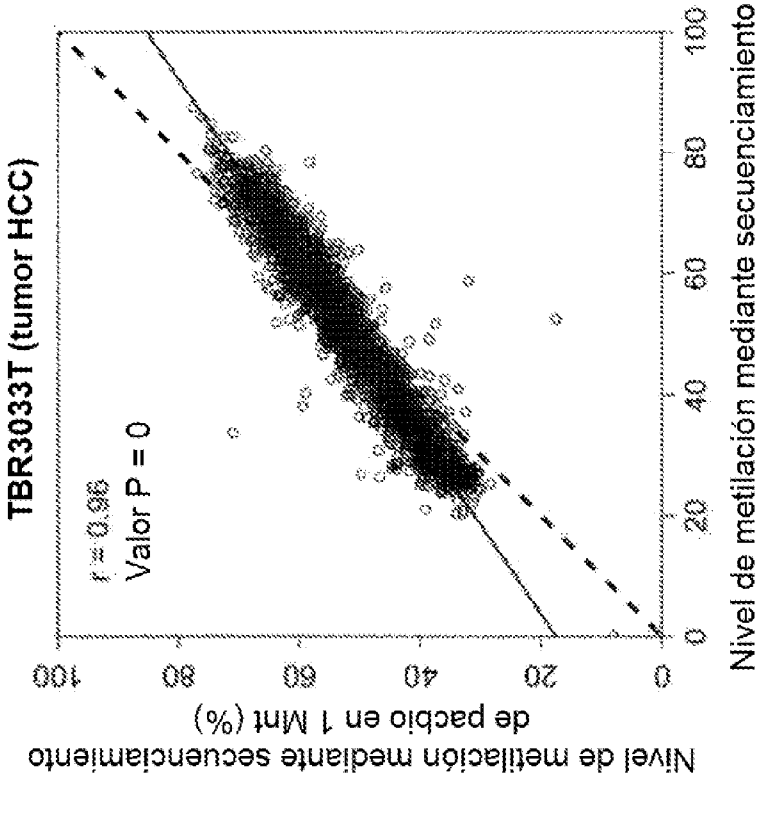


FIG. 68A

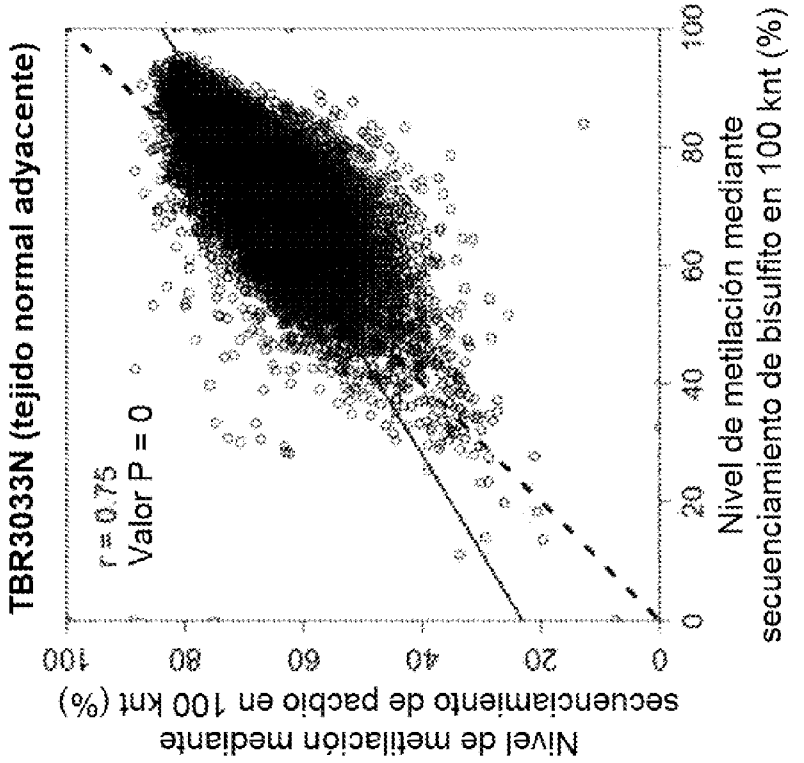


FIG. 69B

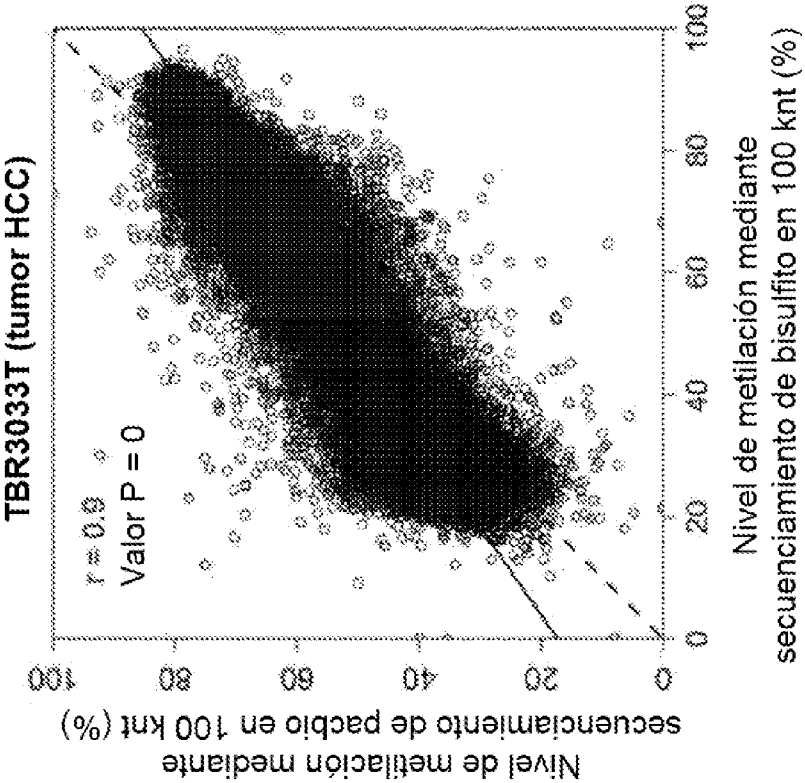


FIG. 69A

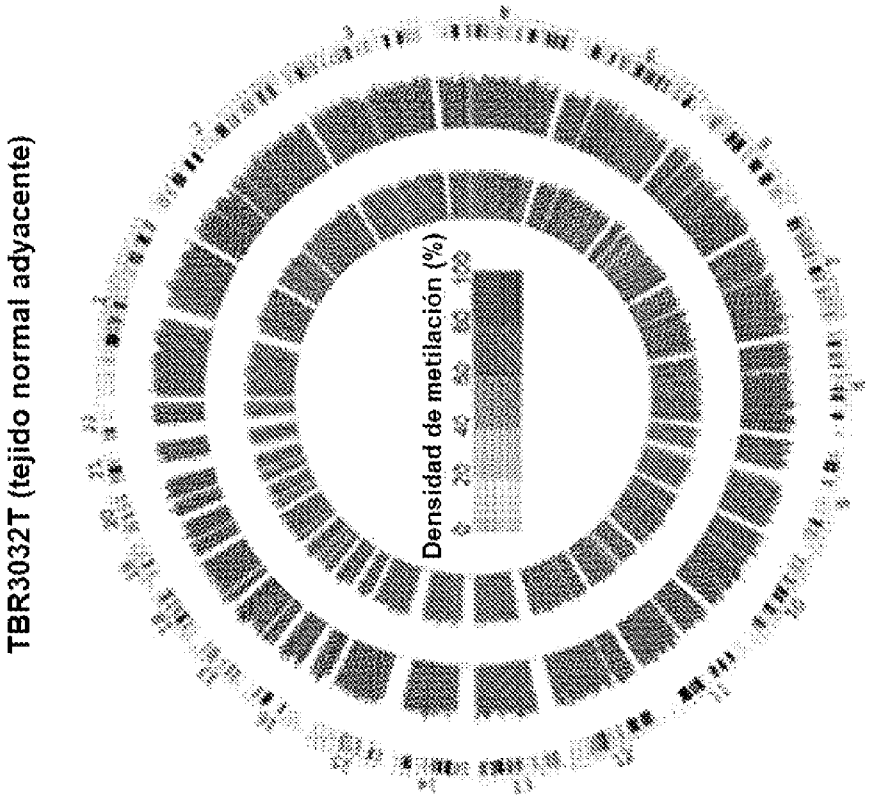


FIG. 70B

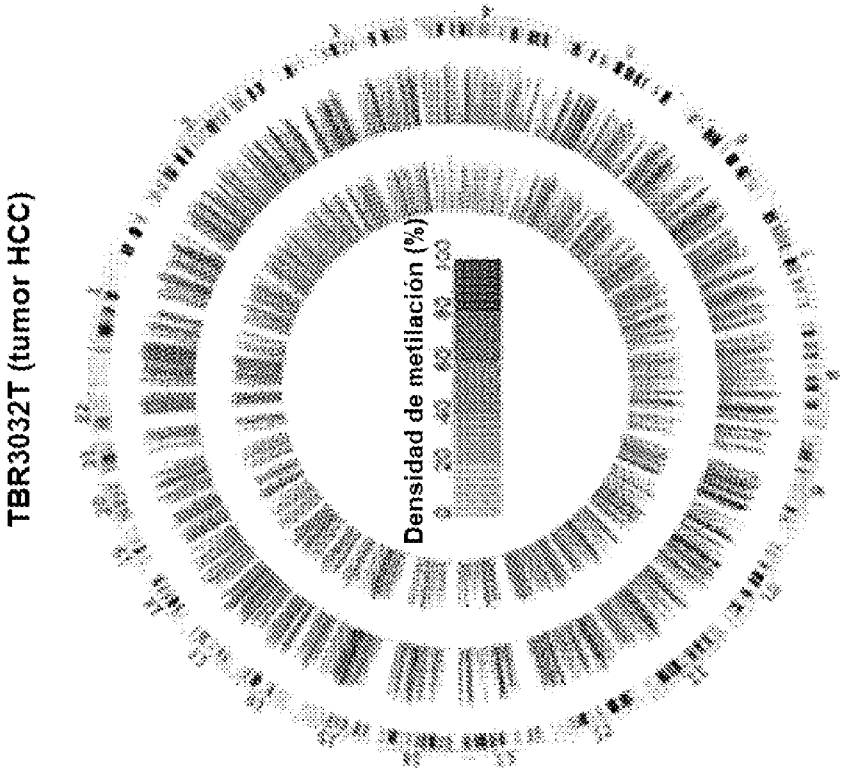


FIG. 70A

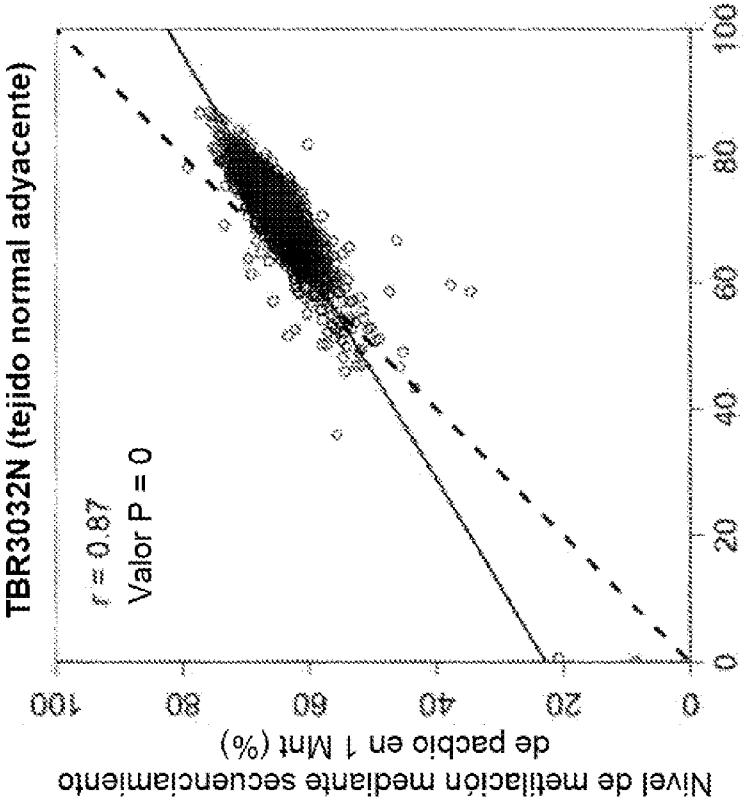


FIG. 71B

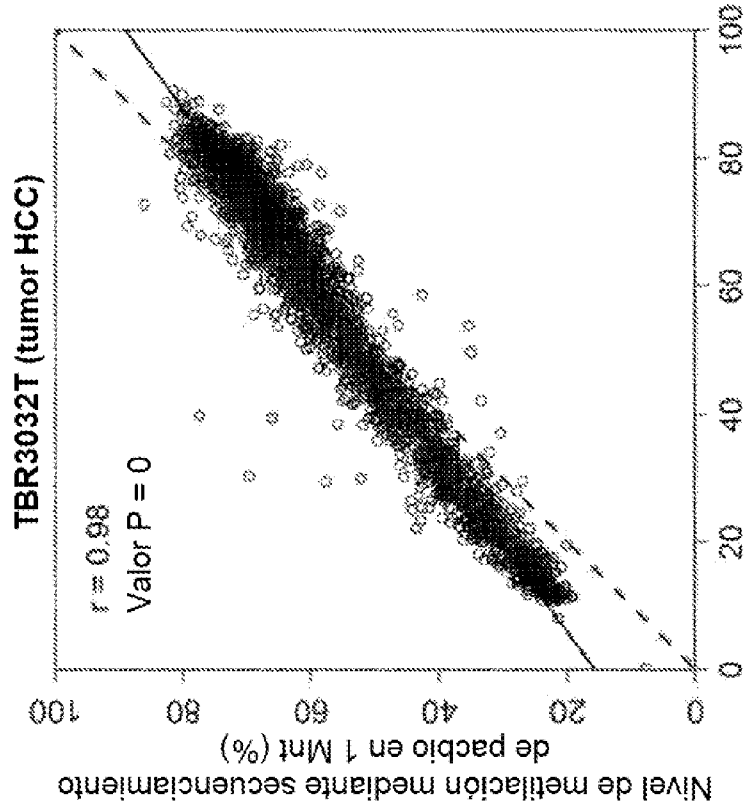


FIG. 71A

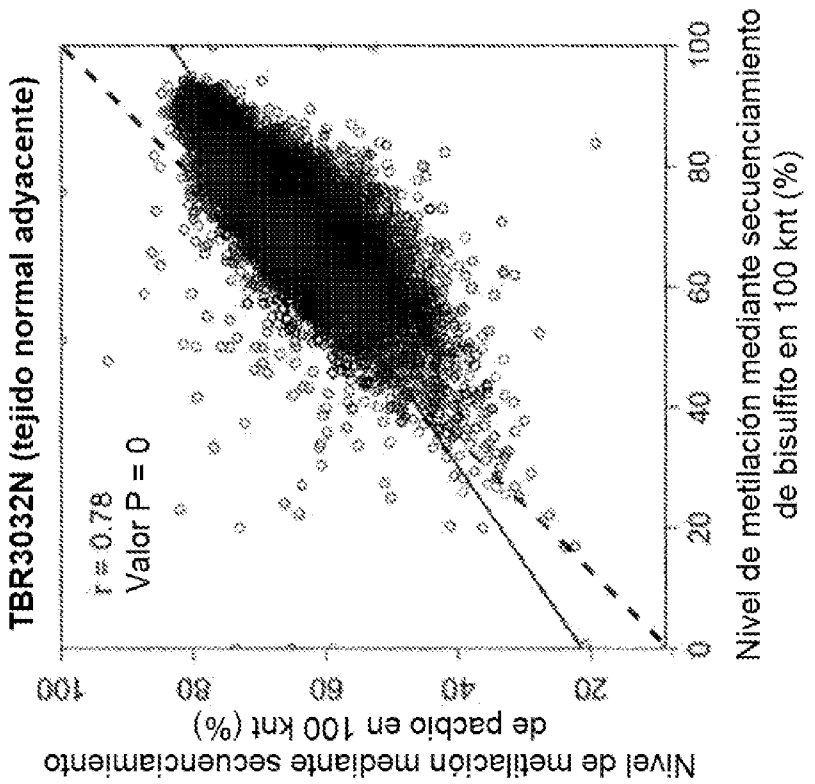


FIG. 72B

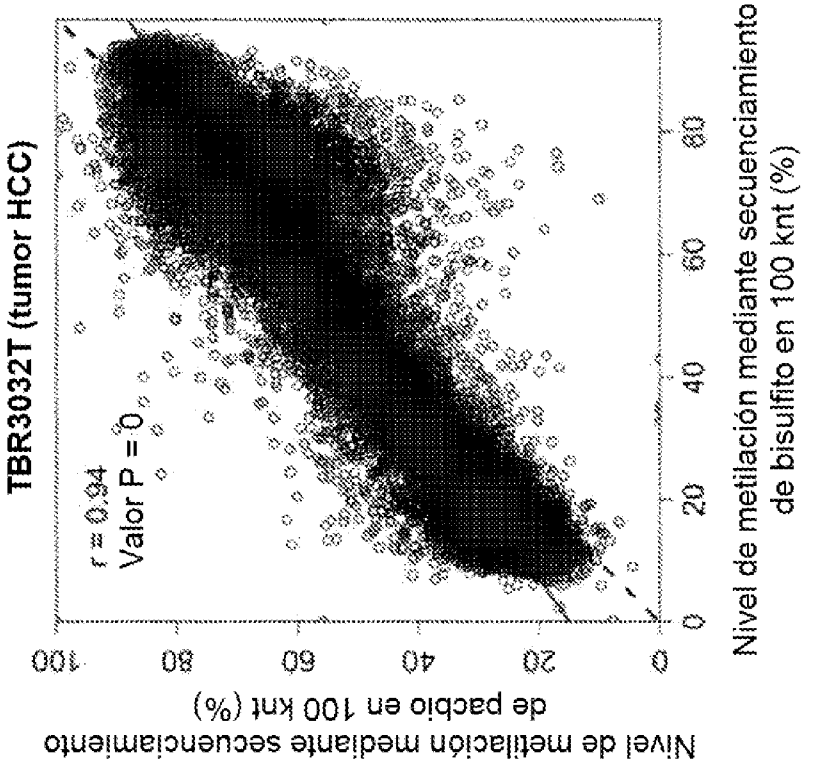


FIG. 72A

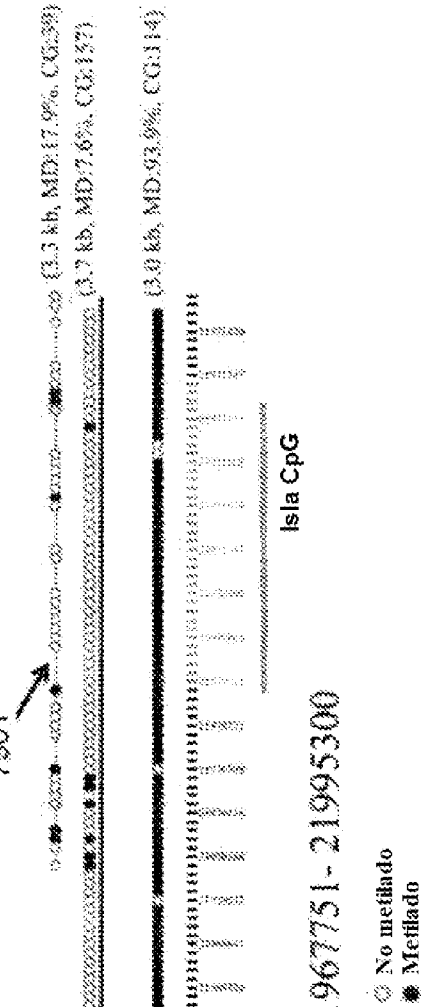

TBR3032
7301
7302
7303
Normal
Tumor
(3.3 kb, MD:17.9%, CG:39)
(3.7 kb, MD:7.6%, CG:137)
(3.0 kb, MD:93.9%, CG:114)
Isla CpG
Isla CpG
CDKN2A chr9:21967751 - 21995300
No metilado
Metilado

FIG. 73

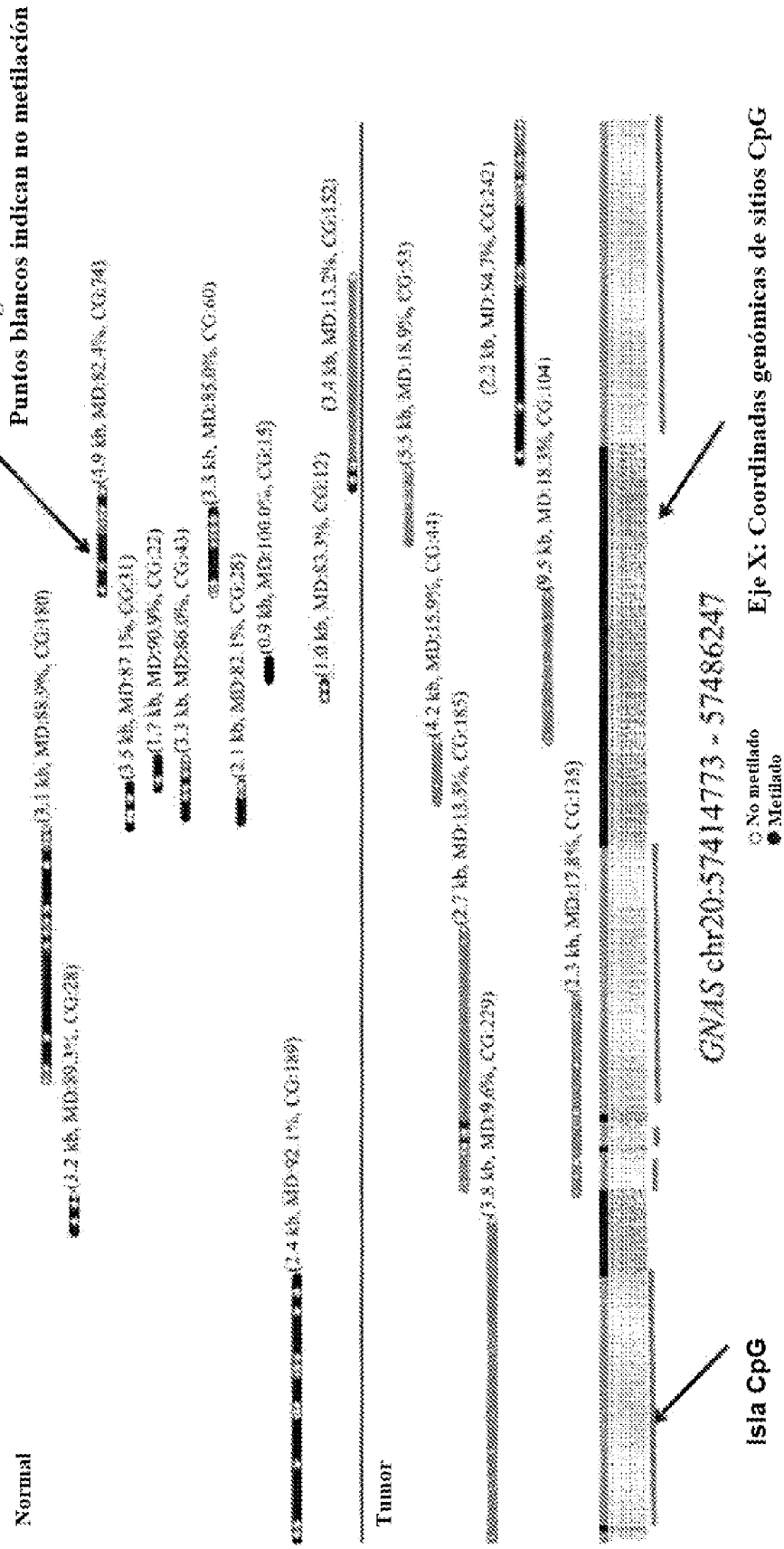

Fragmentos de ADN
Puntos negros indican metilación
Puntos blancos indican no metilación
Normal
Tumor
Isla CpG
GNAS chr20:57414773 - 57486247
No metilado
Metilado
Eje X: Coordinadas genómicas de sitios CpG

FIG. 74A


Fragmentos de ADN
Puntos negros indican metilación
Puntos blancos indican no metilación
Normal
Tumor
Isla CpG
ESR1 chr6:151977826 - 152450754
No metilado
Metilado
Eje X: Coordinadas genómicas de sitios CpG

FIG. 74B

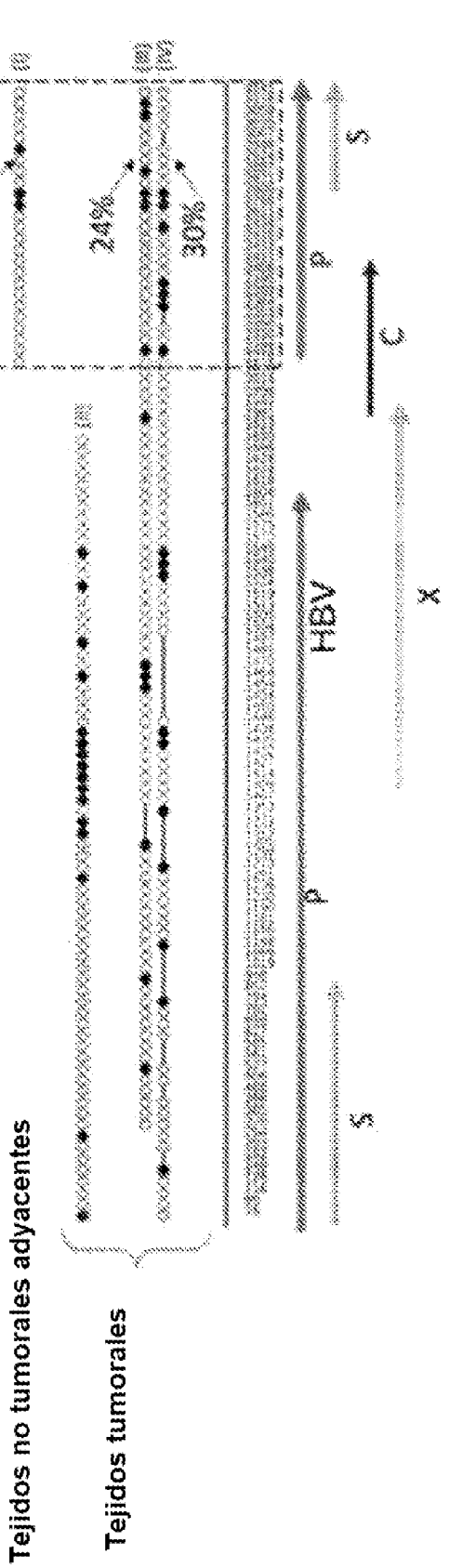
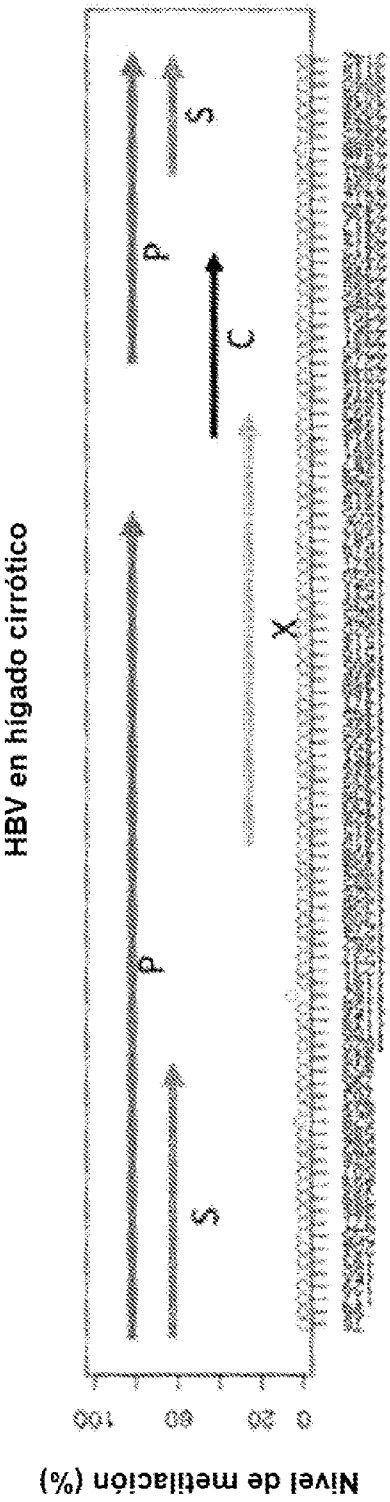
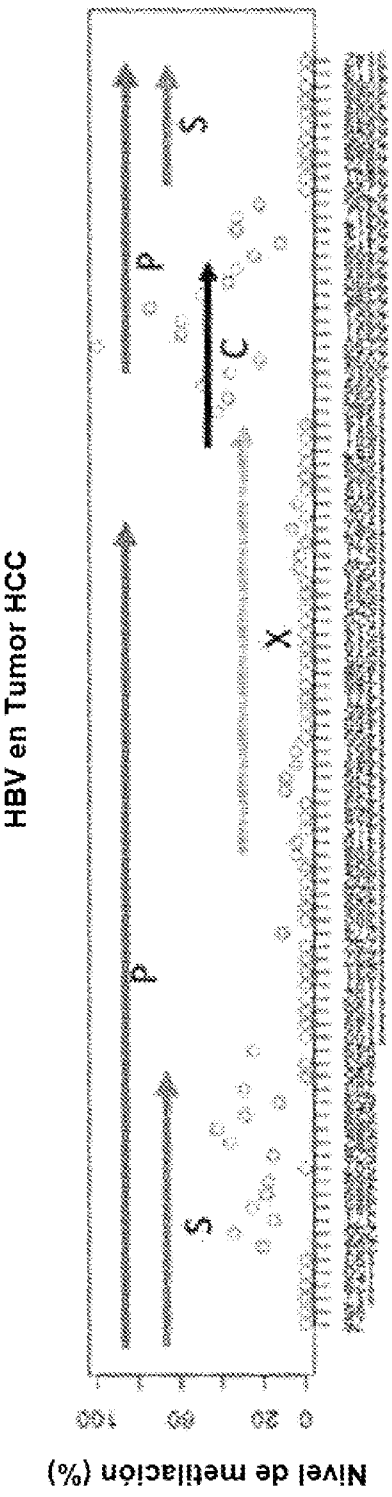

Región metilada diferencialmente
12%
24%
30%
Tejidos no tumorales adyacentes
Tejidos tumorales
HBV
S
P
P
C
S
X

FIG. 75



HBV en 4 tejidos de hígado cirrótico: una mediana de 25 X de profundidad

FIG. 767A



HBV en 15 tejidos de tumor: una mediana de 14 X de profundidad

FIG. 76B

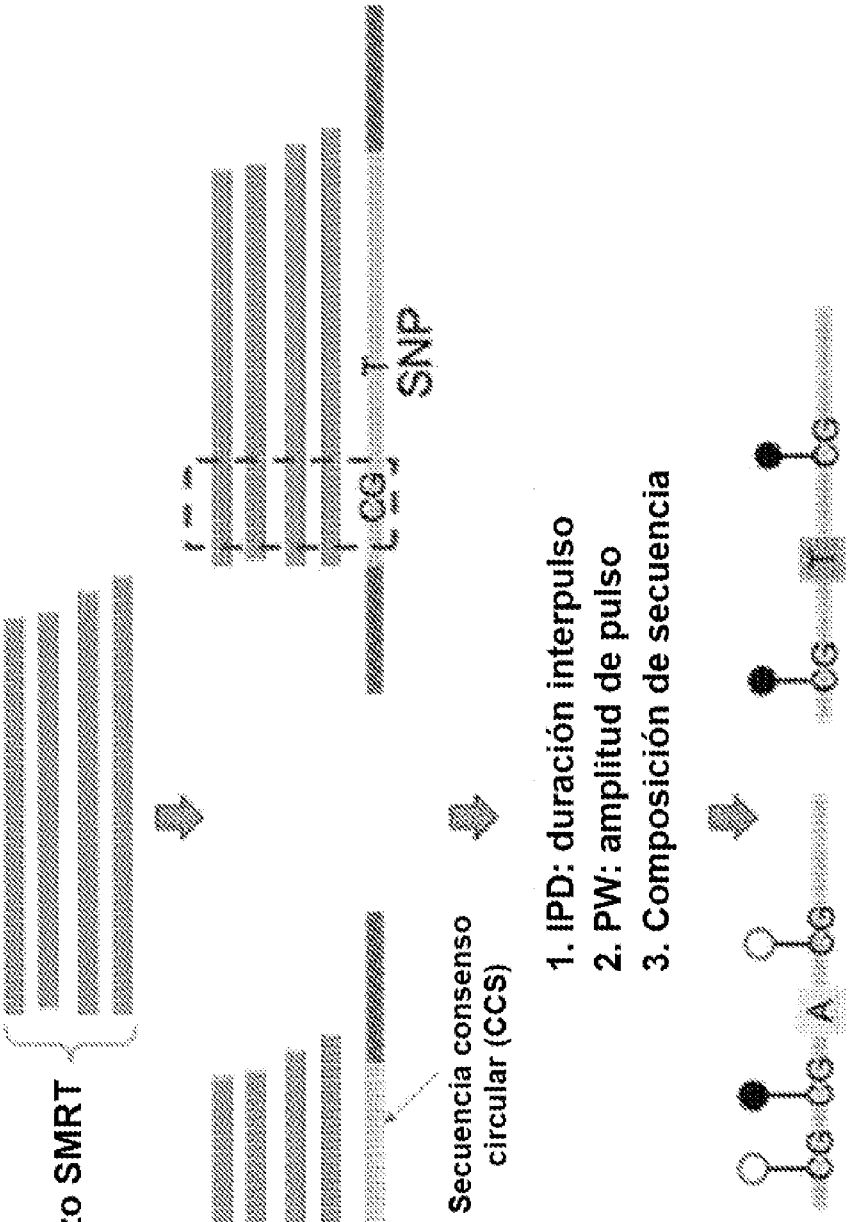

Sublecturas
Desde un pozo SMRT
Genoma
CG
A
SNP
Ventana de medición cinética
Secuencia consenso circular (CCS)
1. IPD: duración interpulso
2. PW: amplitud de pulso
3. Composición de secuencia
CG
T
SNP
Análisis de metilación

FIG. 77

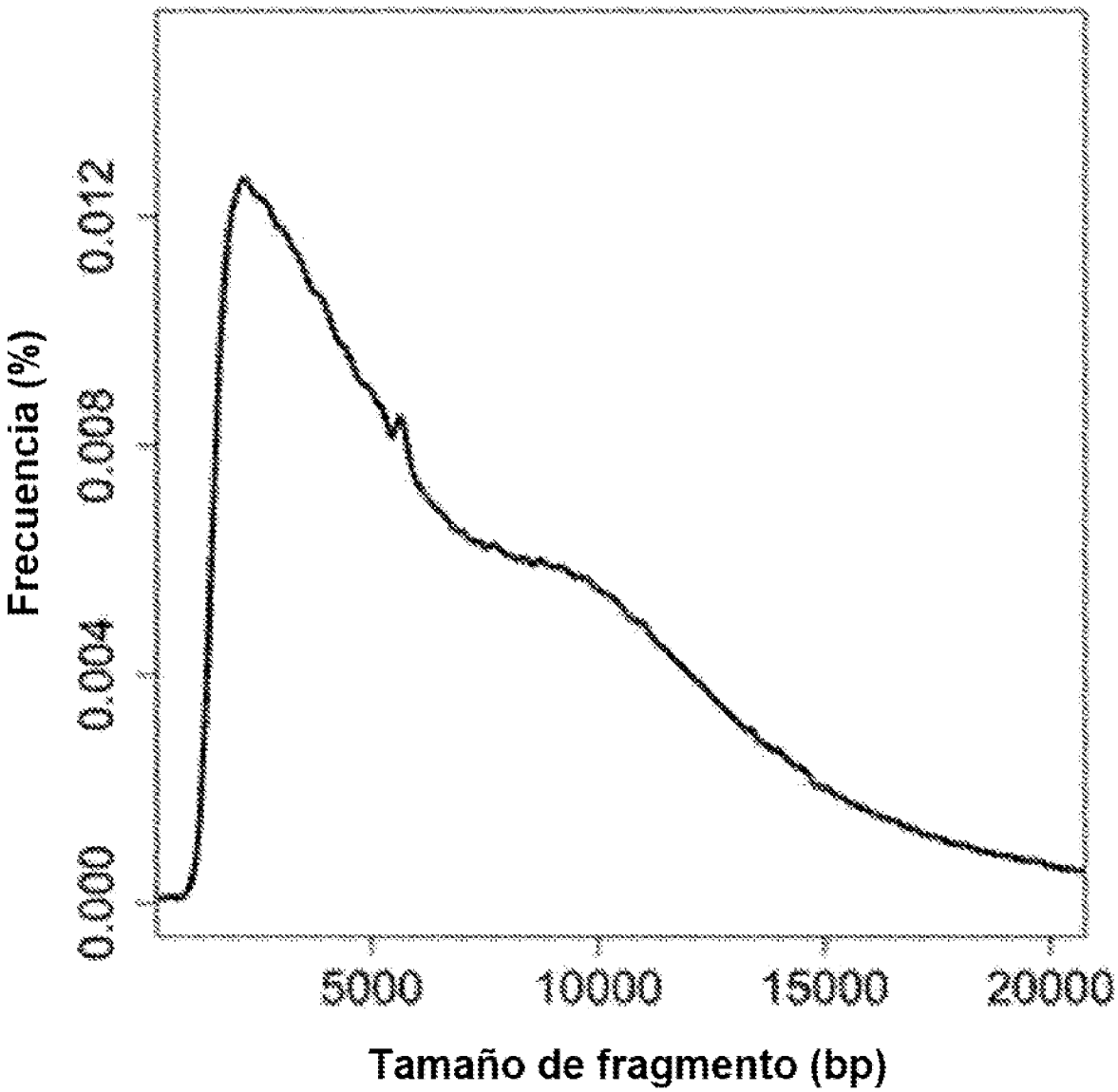

Frecuencia (%)
0.000
0.004
0.008
0.012
5000
10000
15000
20000
Tamaño de fragmento (bp)

FIG. 78

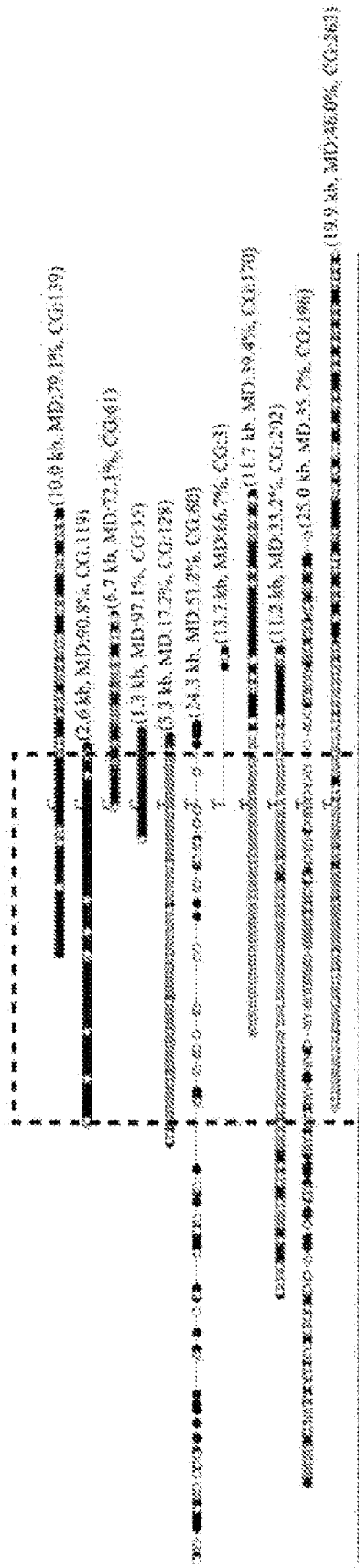


FIG. 79A

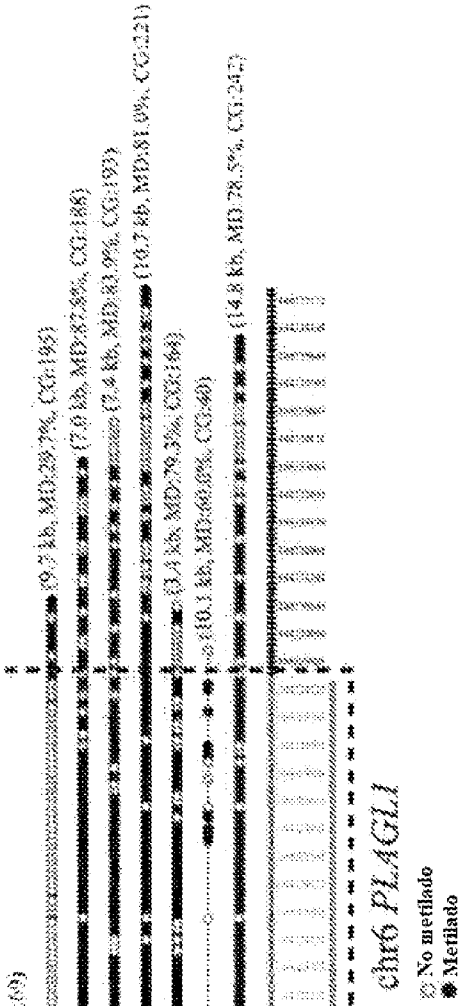


FIG. 79B

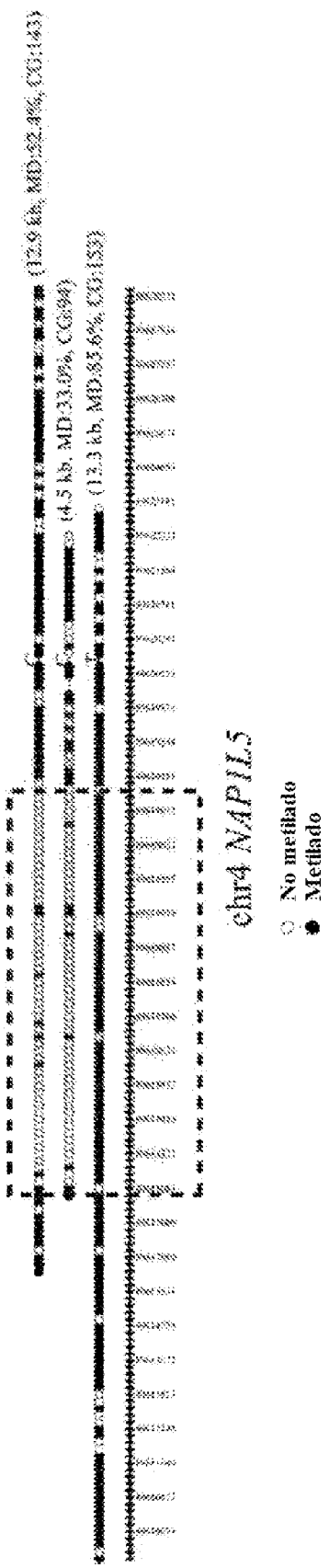

(12.9 kb, MD:92.4%, CG:143)
(4.5 kb, MD:33.0%, CG:94)
(12.3 kb, MD:85.6%, CG:153)
chr4 NAPIL5
No metilado
Metilado

FIG. 79C

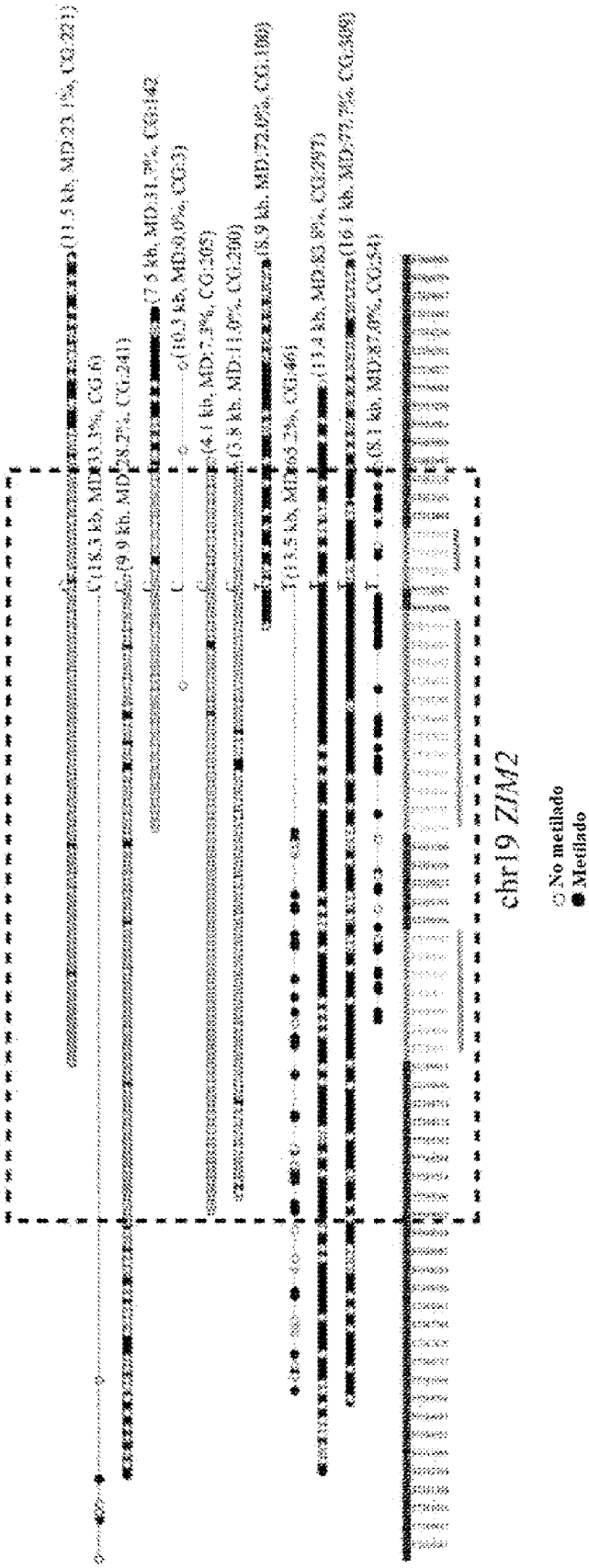


FIG. 79D

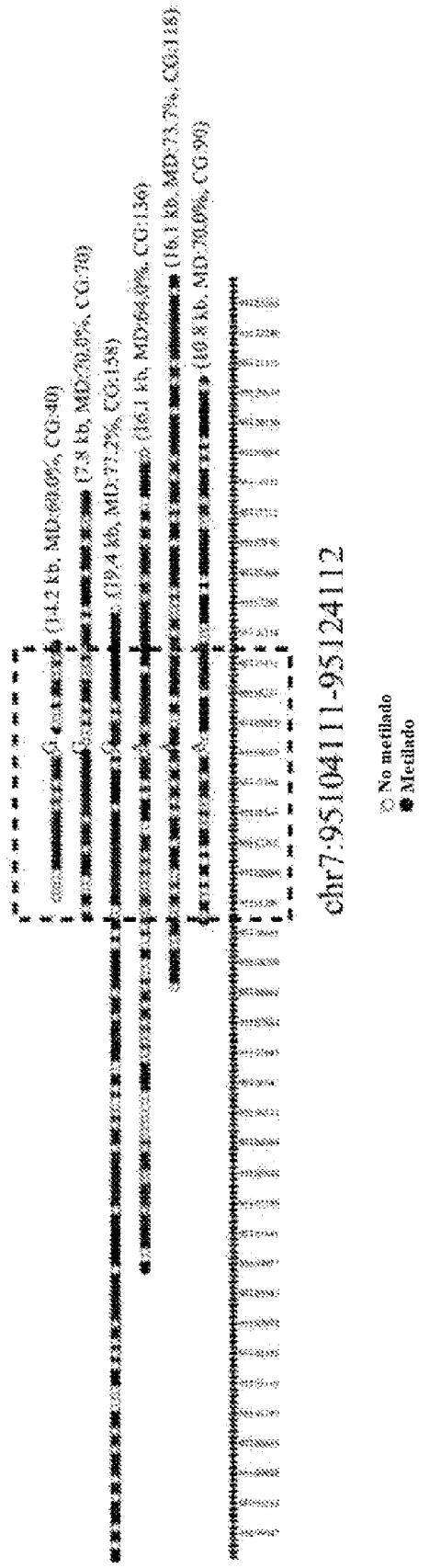

(14.2 kb, MD:99.0%, CG:40)
(7.8 kb, MD:70.8%, CG:70)
(19.4 kb, MD:77.2%, CG:158)
(16.1 kb, MD:64.0%, CG:136)
(16.1 kb, MD:73.7%, CG:118)
(40.8 kb, MD:78.6%, CG:90)
chr7:95104111-95124112
No metilado
Metilado

FIG. 80A

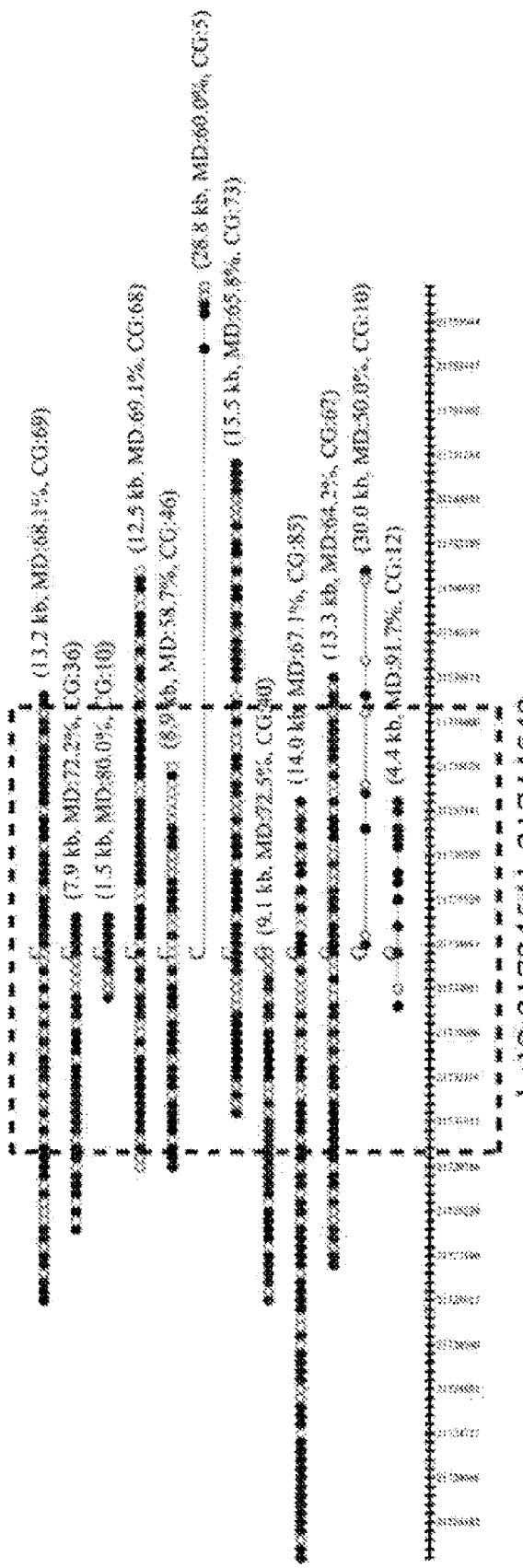

(13.2 kb, MD:68.3%, CG:69)
(7.9 kb, MD:72.2%, CG:36)
(1.5 kb, MD:80.0%, CG:10)
(12.9 kb, MD:69.1%, CG:68)
(8.9 kb, MD:58.7%, CG:46)
(28.8 kb, MD:69.0%, CG:5)
(15.5 kb, MD:65.8%, CG:73)
(9.1 kb, MD:72.5%, CG:30)
(14.0 kb, MD:67.1%, CG:85)
(13.3 kb, MD:64.2%, CG:67)
(30.0 kb, MD:30.0%, CG:10)
(4.4 kb, MD:91.7%, CG:12)
chr12:21724541-21744542
No metilado
Metilado

FIG. 80B

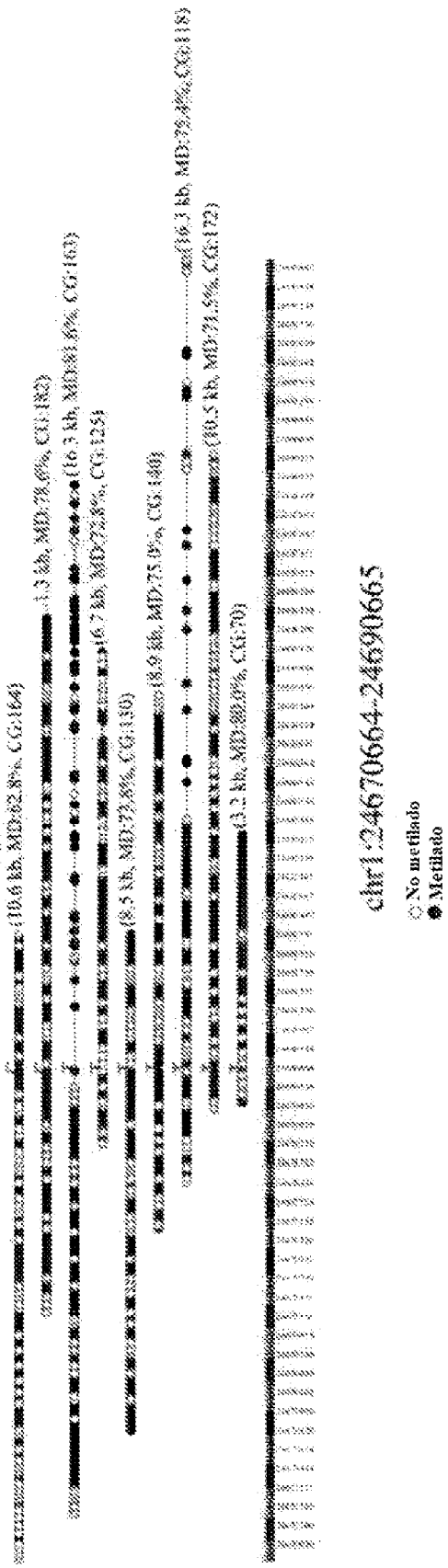


FIG. 80C

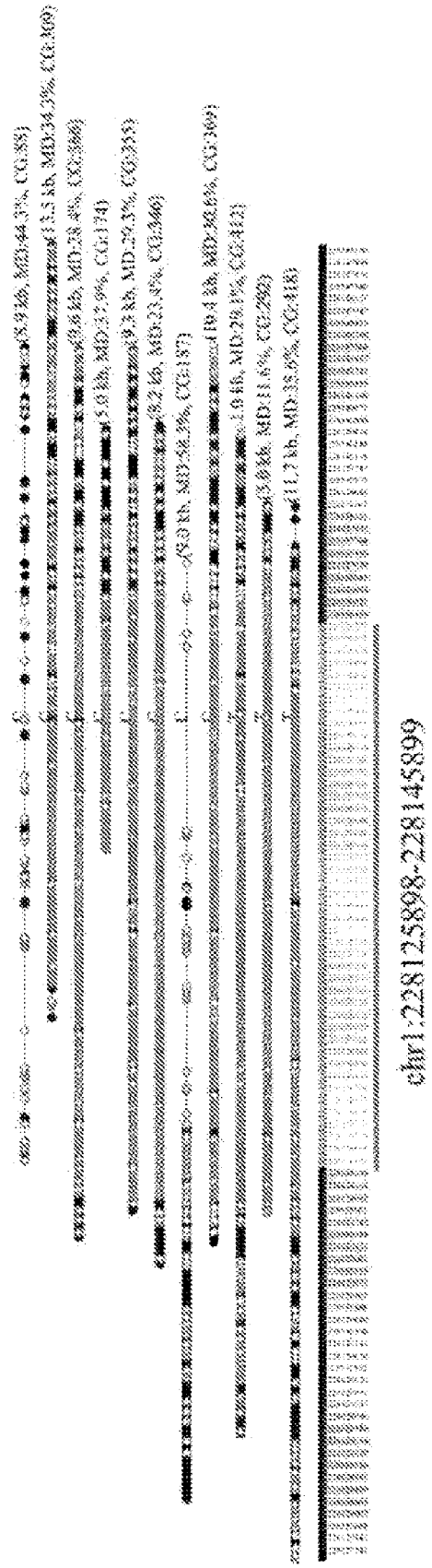


FIG. 80D

	Gen	Alelo 1	Alelo 2	Nivel de metilación (%)	
				Alelo 1	Alelo 2
Genes impresos	*SNURF*	T	C	15.73	89.37
	PLAGL1	T	C	7.56	89.41
	NAP1L5	C	T	12.5	91.07
	ZIM2	C	T	13	84.64
Regiones seleccionadas aleatoriamente	Región 01	G	A	71.79	69.17
	Región 02	T	G	63.22	65.22
	Región 03	C	T	73.33	74.9
	Región 04	C	T	10.83	8.51

FIG. 81

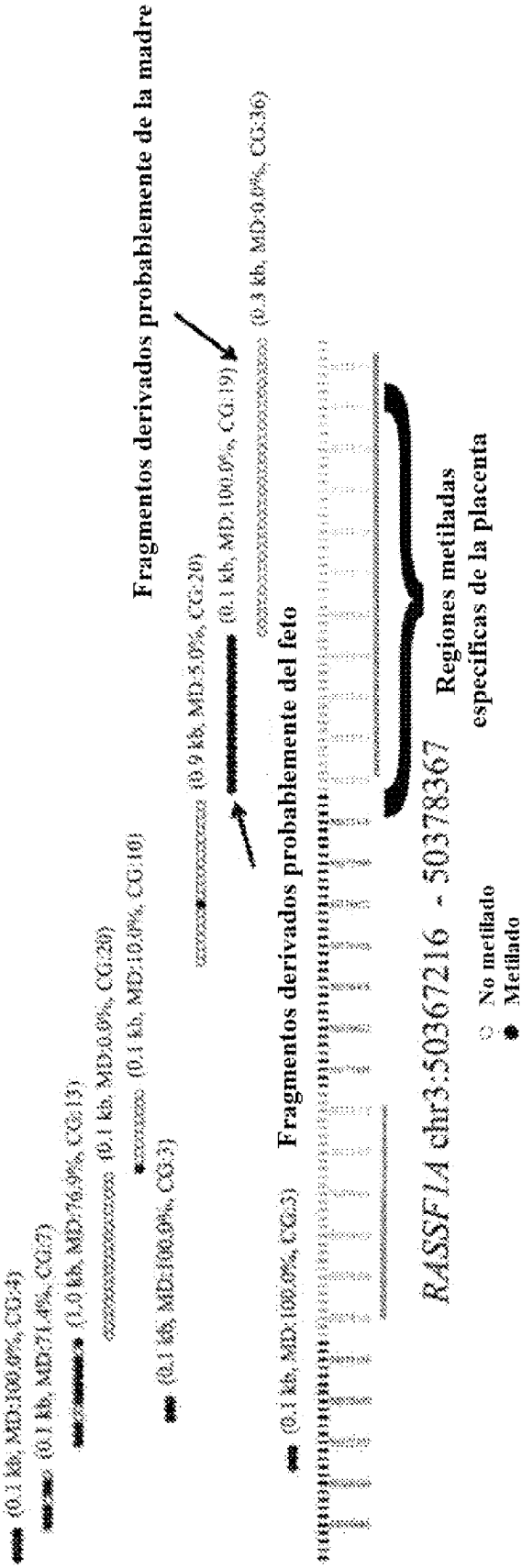

(0.7 kb, MD:20.0%, CG:15)
(0.1 kb, MD:0.0%, CG:2)
(0.1 kb, MD:100.0%, CG:4)
(0.1 kb, MD:71.4%, CG:7)
(1.0 kb, MD:76.9%, CG:13)
(0.1 kb, MD:0.0%, CG:20)
(0.1 kb, MD:10.0%, CG:10)
(0.1 kb, MD:100.0%, CG:3)
(0.9 kb, MD:5.0%, CG:20)
(0.1 kb, MD:100.0%, CG:19)
(0.3 kb, MD:0.0%, CG:36)
Fragmentos derivados probablemente de la madre
Fragmentos derivados probablemente del feto
(0.1 kb, MD:100.0%, CG:3)
RASSF1A chr3:50367216 - 50378367
Regiones metiladas
específicas de la placenta
No metilado
Metilado

FIG. 82

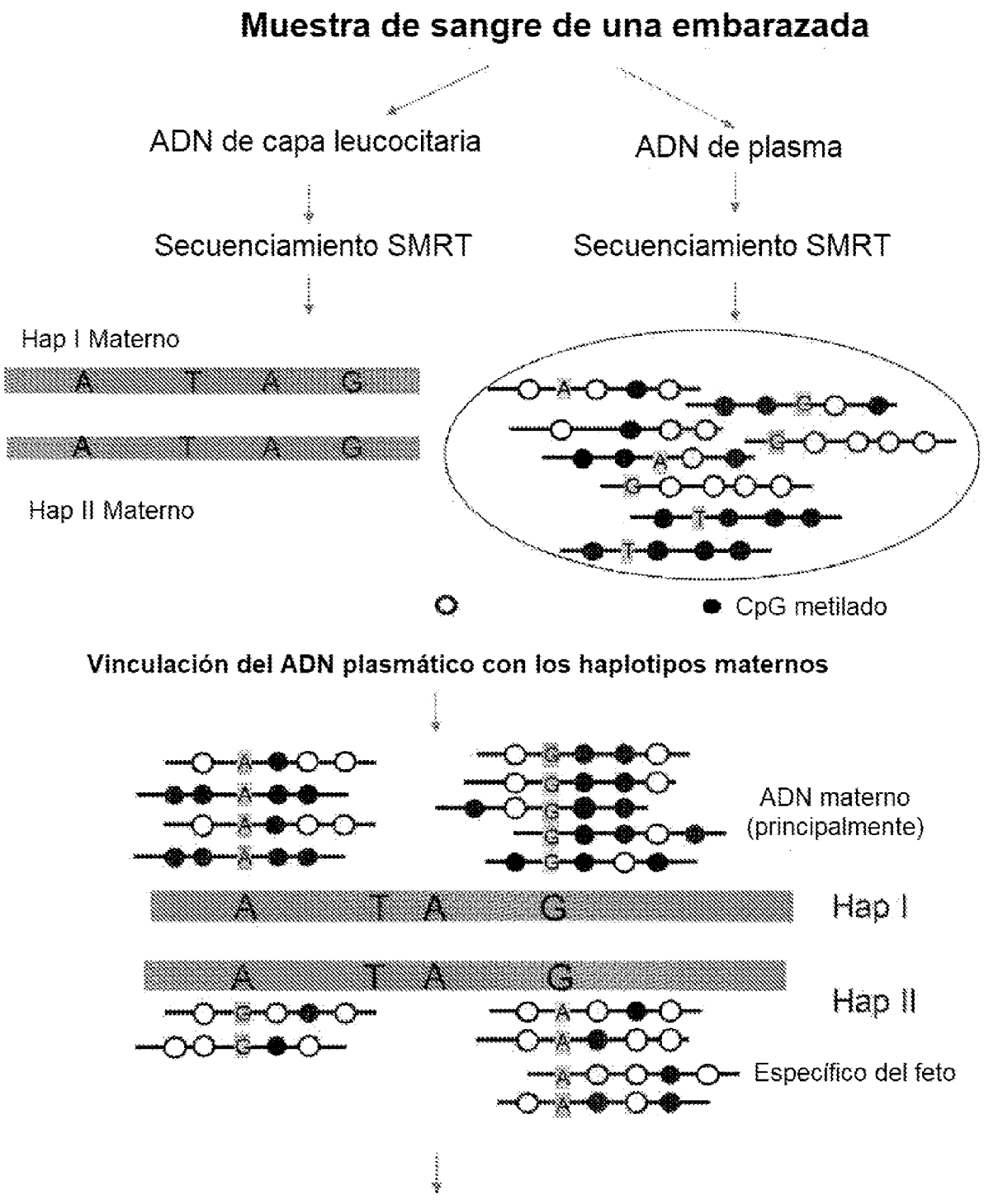

Muestra de sangre de una embarazada
ADN de capa leucocitaria
ADN de plasma
Secuenciamiento SMRT
Secuenciamiento SMRT
Hap I Materno
A T A G
A T A G
Hap II Materno
CpG metilado
Vinculación del ADN plasmático con los haplotipos maternos
ADN materno
(principalmente)
A T A G
Hap I
A T A G
Hap II
Específico del feto
El nivel de metilación de ADN específico de feto sería más bajo en comparación con aquellos de los fragmentos maternos

FIG. 83

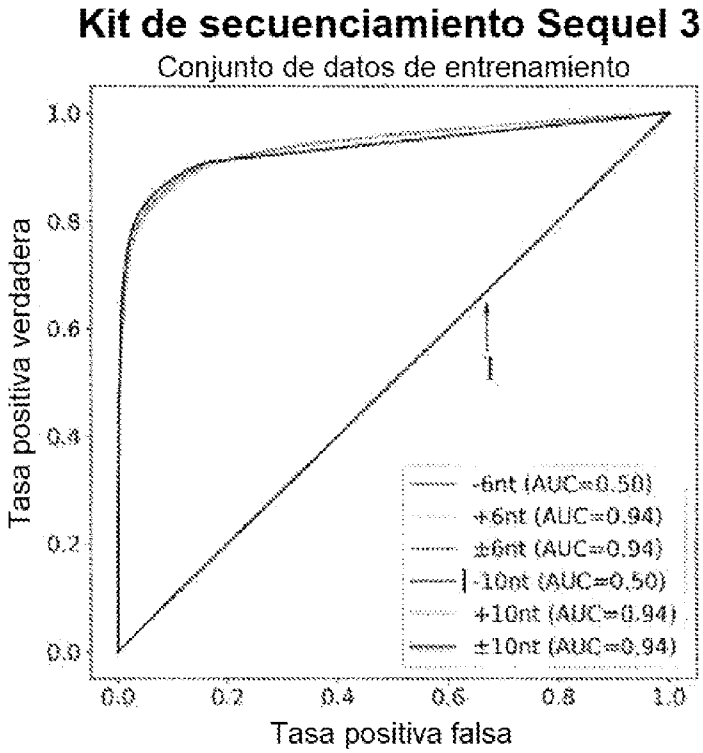


FIG. 84A

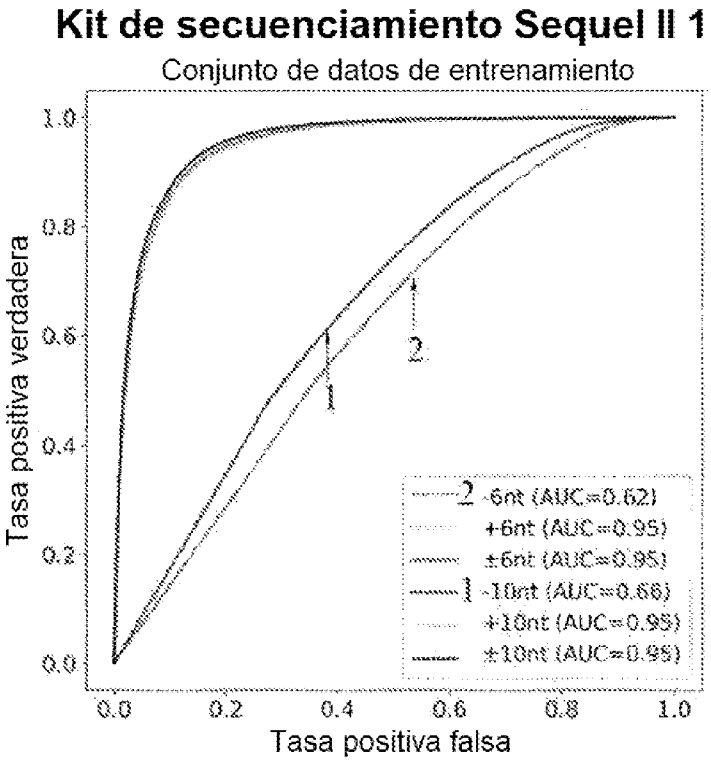


FIG. 84B

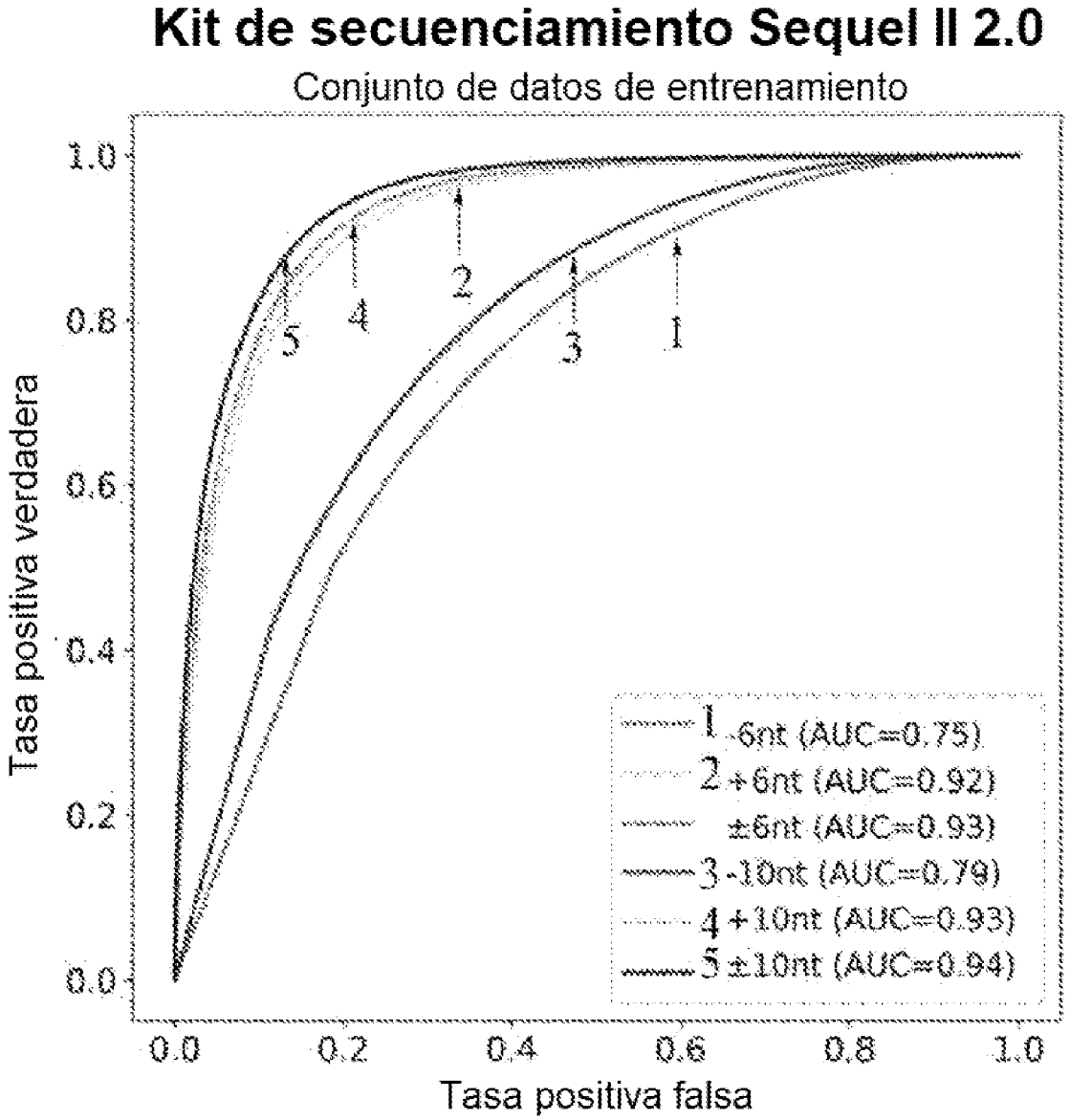

Kit de secuenciamiento Sequel II 2.0
Conjunto de datos de entrenamiento
Tasa positiva verdadera
Tasa positiva falsa
0.0
0.2
0.4
0.6
0.8
1.0
1 -6nt (AUC=0.75)
2 +6nt (AUC=0.92)
±6nt (AUC=0.93)
3 -10nt (AUC=0.79)
4 +10nt (AUC=0.93)
5 ±10nt (AUC=0.94)

FIG. 84C

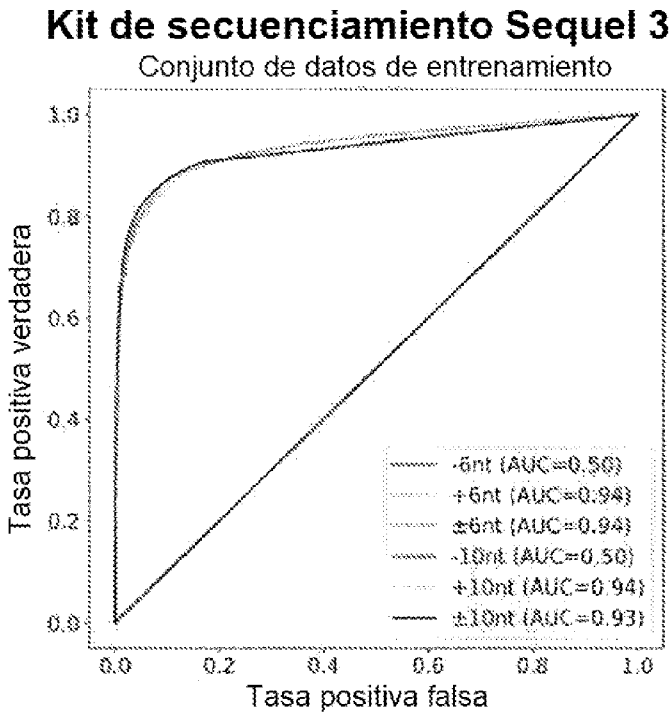


FIG. 85A

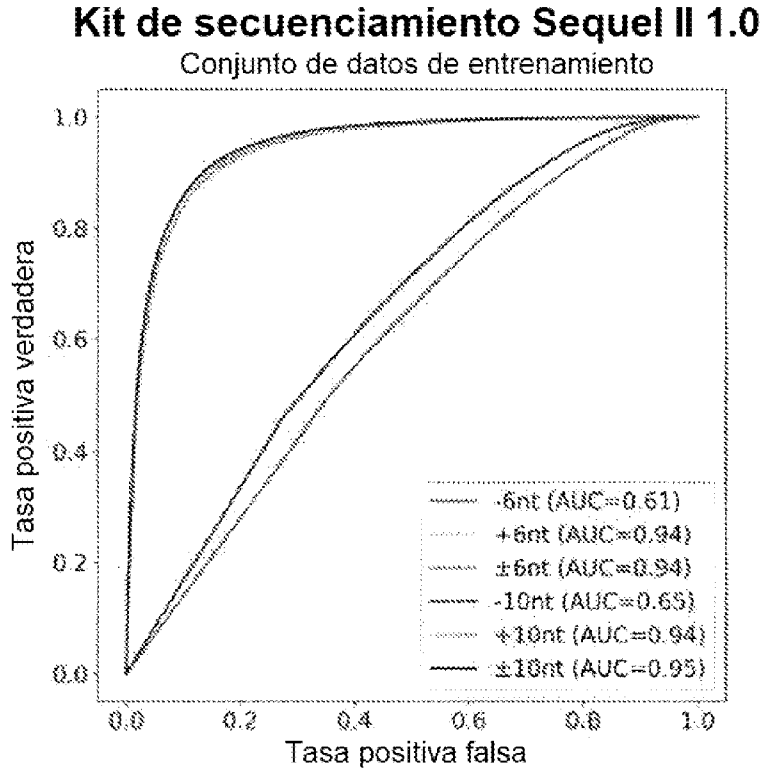


FIG. 85B

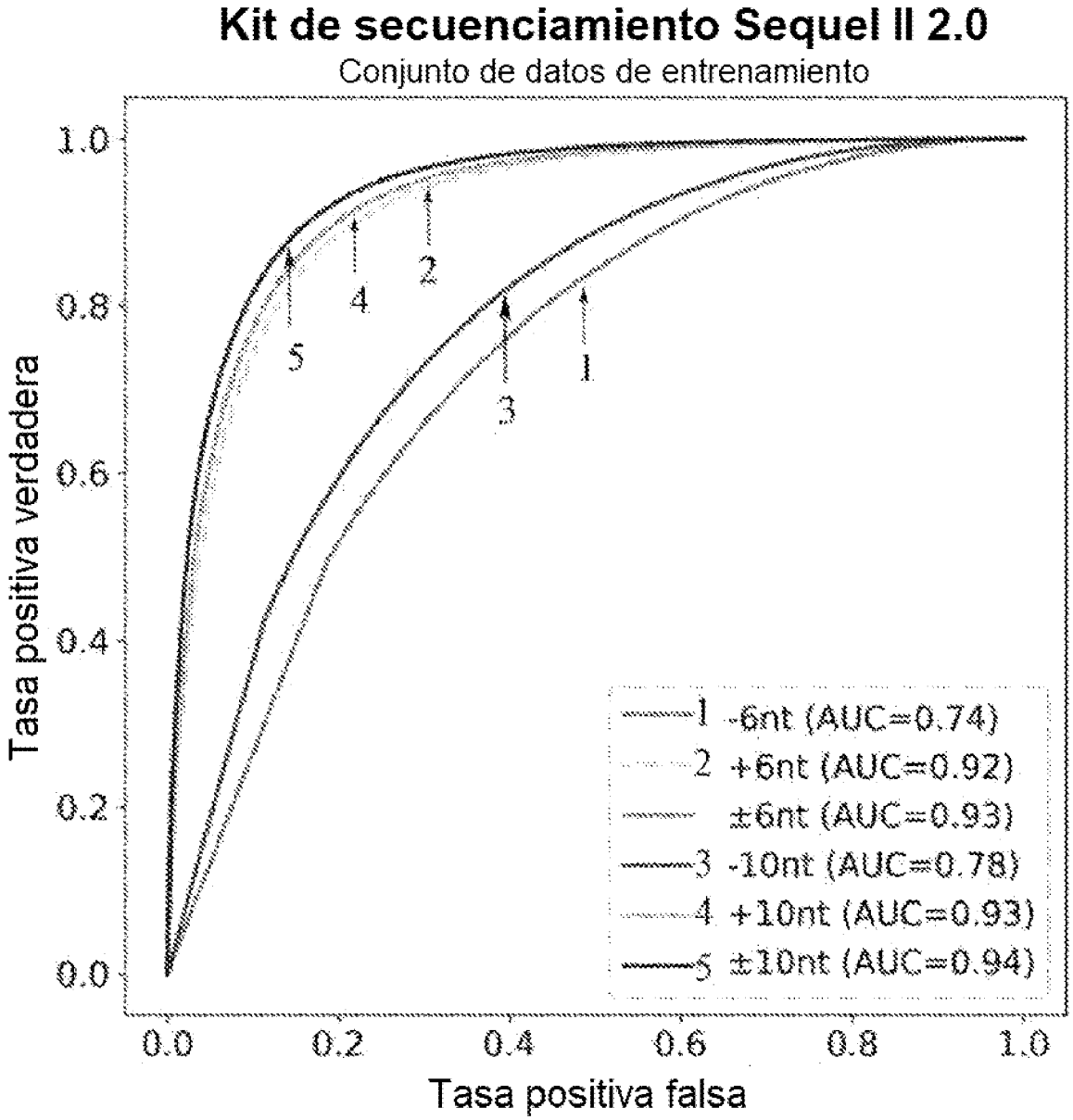

Kit de secuenciamiento Sequel II 2.0
Conjunto de datos de entrenamiento
Tasa positiva verdadera
Tasa positiva falsa
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
1 -6nt (AUC=0.74)
2 +6nt (AUC=0.92)
±6nt (AUC=0.93)
3 -10nt (AUC=0.78)
4 +10nt (AUC=0.93)
5 ±10nt (AUC=0.94)

FIG. 85C

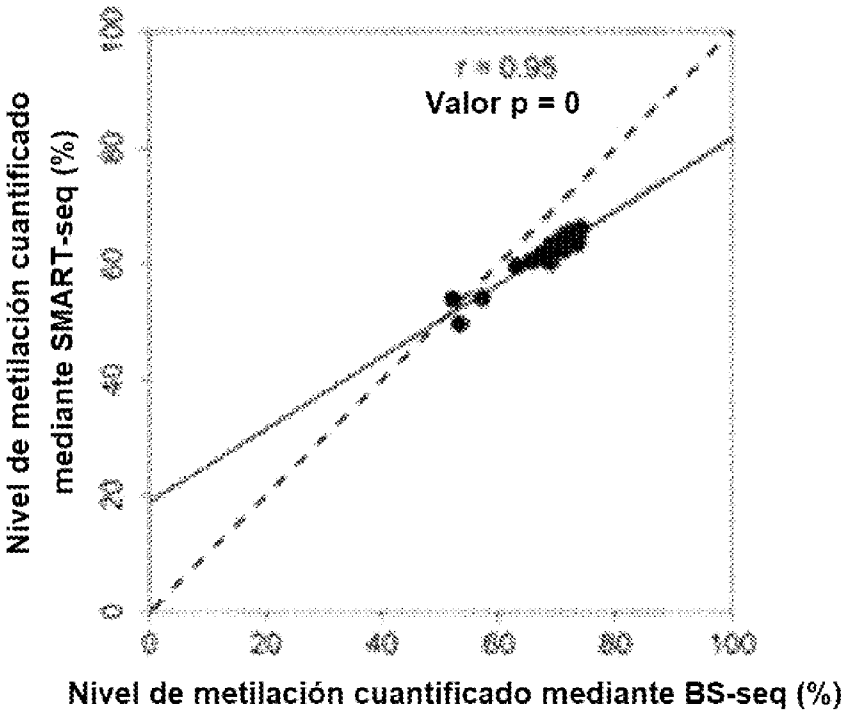


FIG. 86A

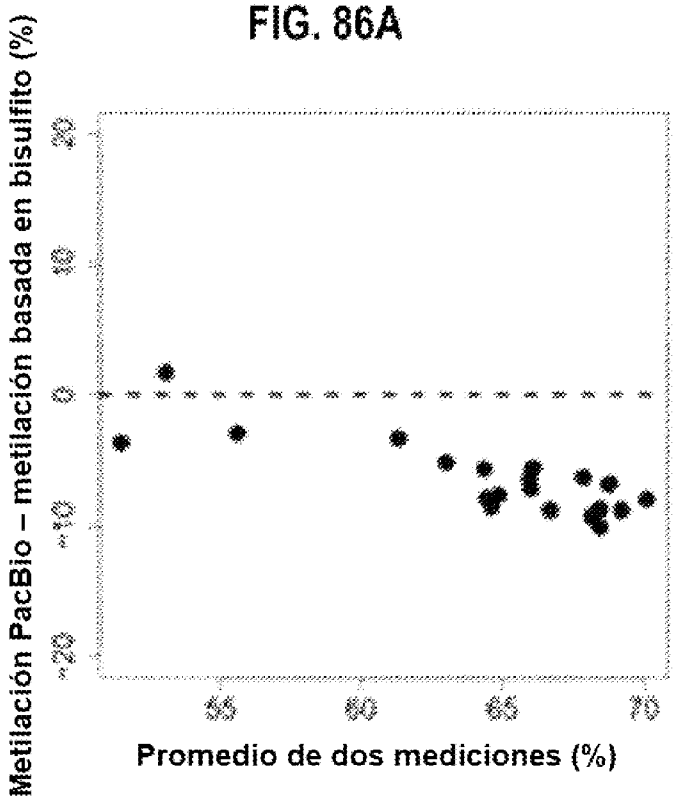


FIG. 86B

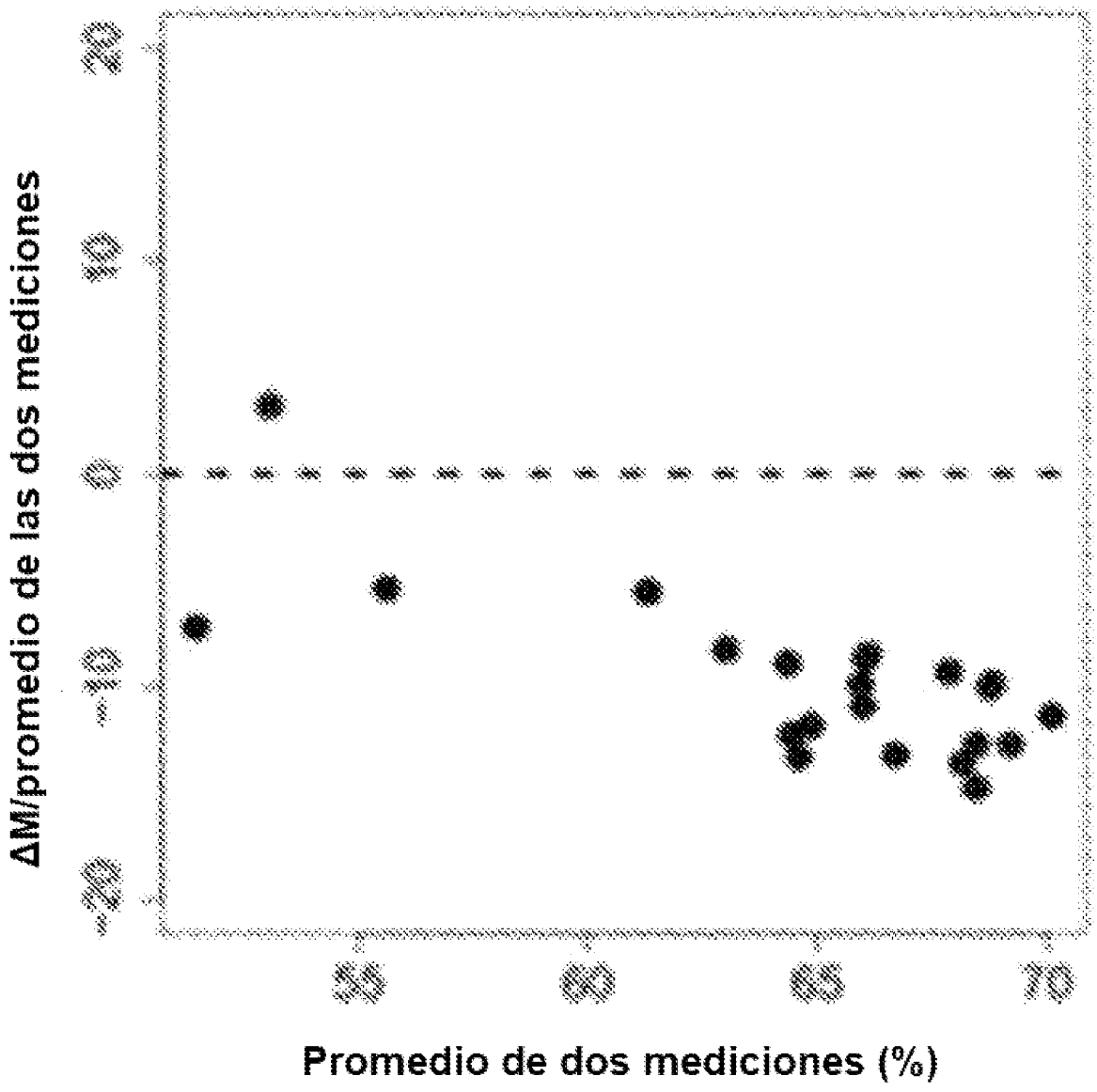

ΔM/promedio de las dos mediciones
Promedio de dos mediciones (%)

FIG. 86C

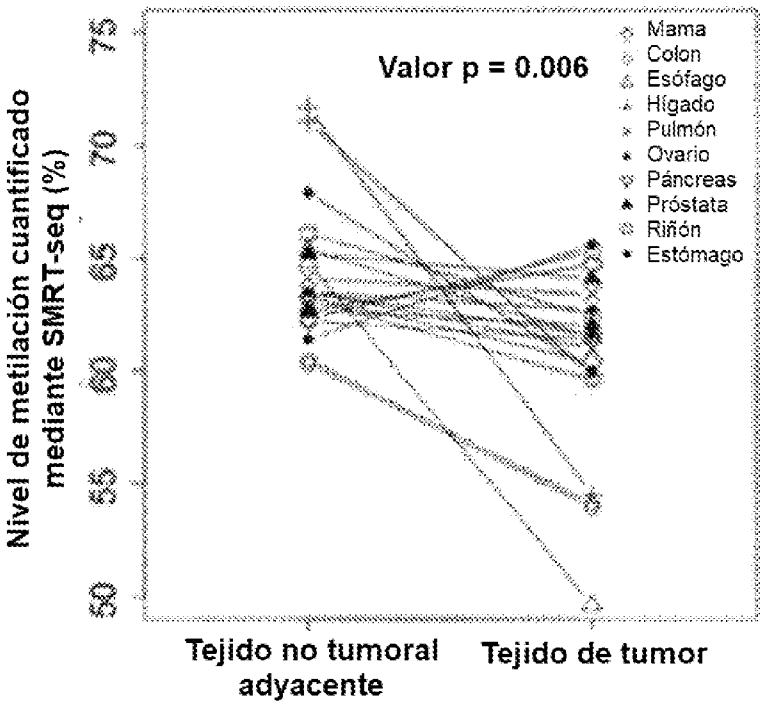


FIG. 87A

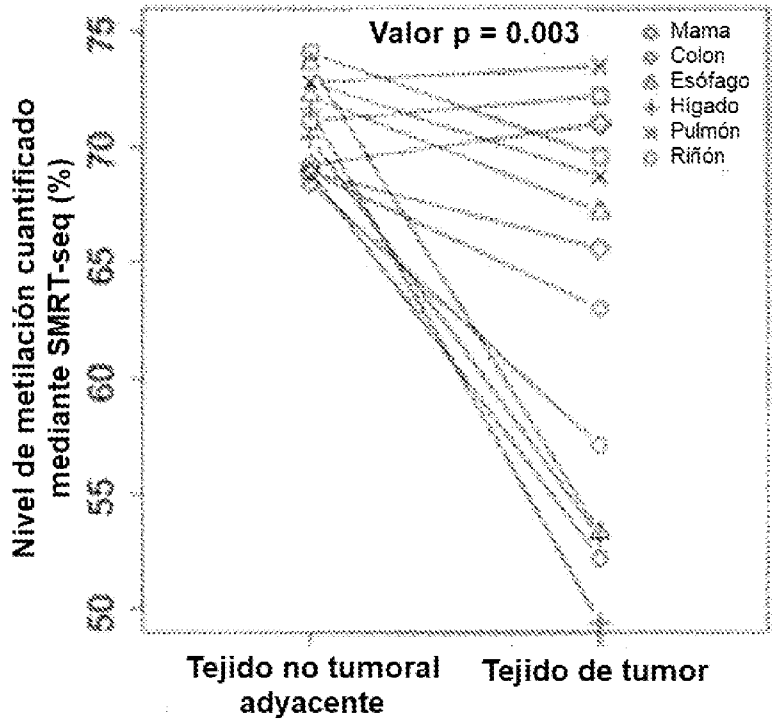


FIG. 87B

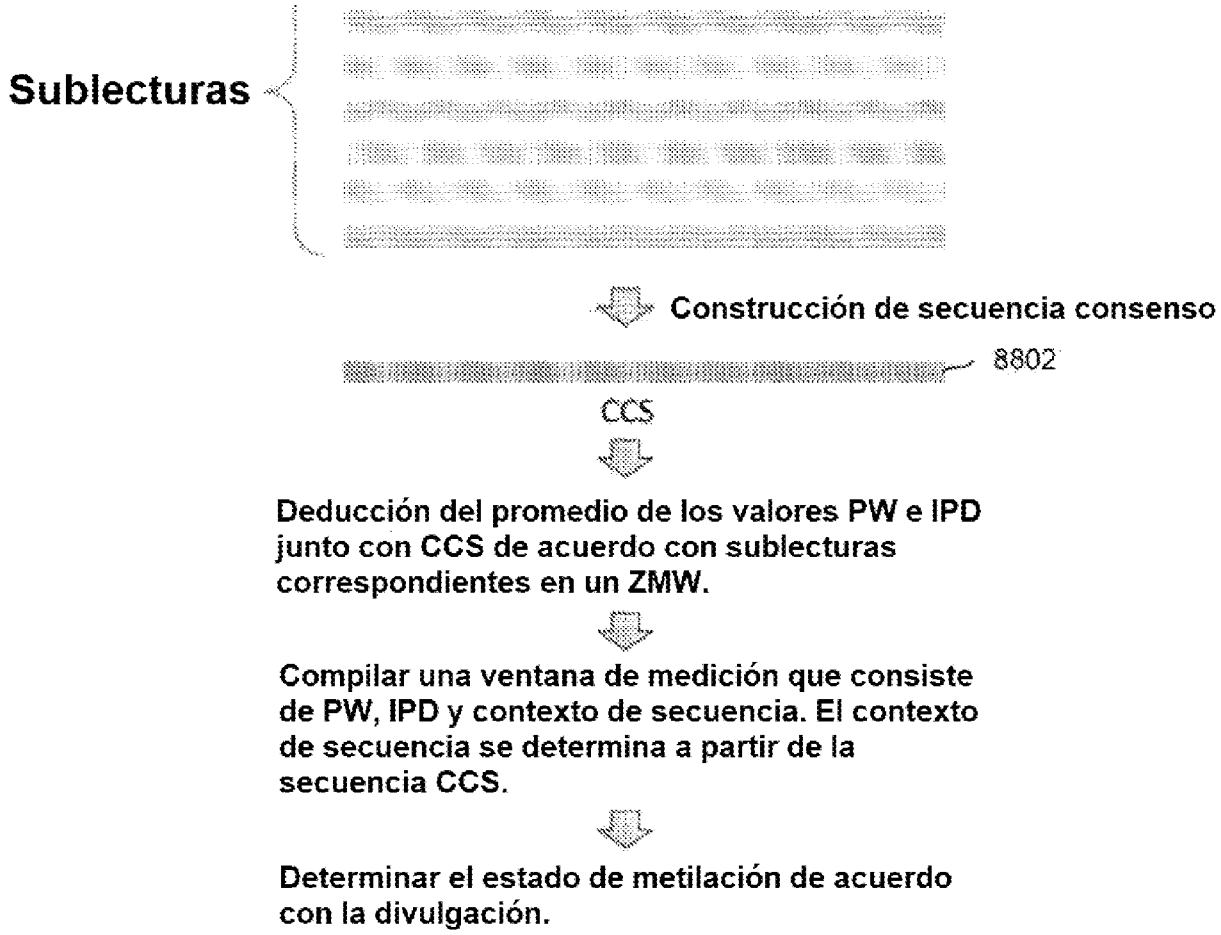

Sublecturas
Construcción de secuencia consenso
8802
CCS
Deducción del promedio de los valores PW e IPD junto con CCS de acuerdo con sublecturas correspondientes en un ZMW.
Compilar una ventana de medición que consiste de PW, IPD y contexto de secuencia. El contexto de secuencia se determina a partir de la secuencia CCS.
Determinar el estado de metilación de acuerdo con la divulgación.

FIG. 88

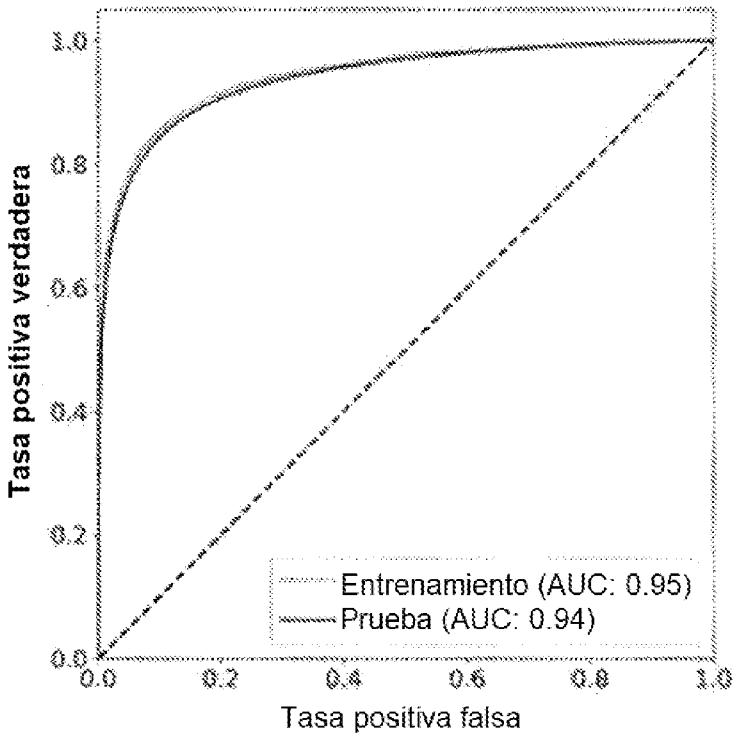

Tasa positiva verdadera
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Tasa positiva falsa
Entrenamiento (AUC: 0.95)
Prueba (AUC: 0.94)

FIG. 89

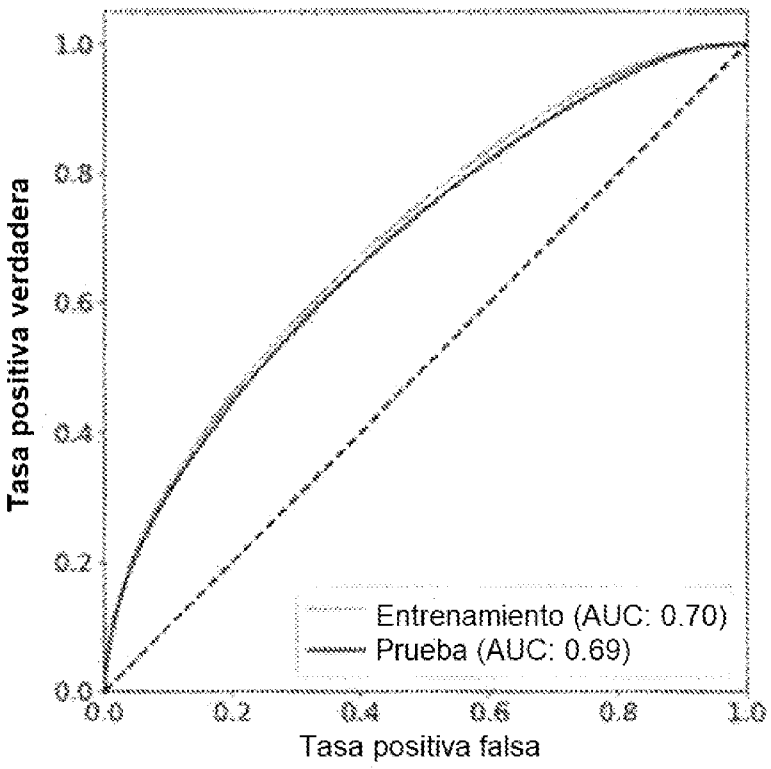

Tasa positiva verdadera
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Tasa positiva falsa
Entrenamiento (AUC: 0.70)
Prueba (AUC: 0.69)

FIG. 90

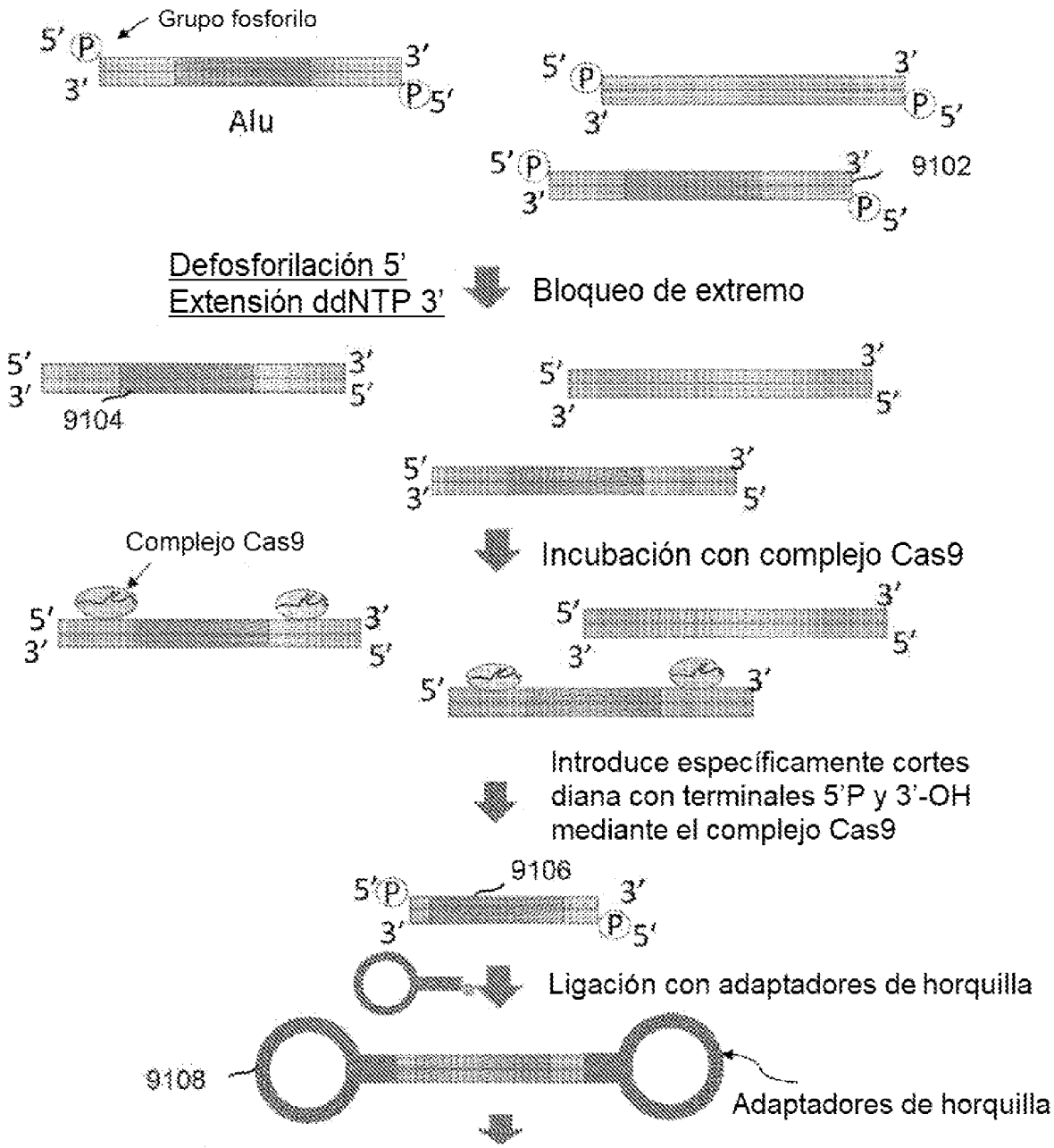

Grupo fosforilo
5'
P
3'
Alu
9102
Defosforilación 5'
Extensión ddNTP 3'
Bloqueo de extremo
9104
Complejo Cas9
Incubación con complejo Cas9
Introduce específicamente cortes
diana con terminales 5'P y 3'-OH
mediante el complejo Cas9
9106
Ligación con adaptadores de horquilla
9108
Adaptadores de horquilla
Elimina los adaptadores, ADN lineal y ADN fuera de diana
utilizando exonucleasa (por ejemplo exonucleasa III y VII)
Molécula sencilla, secuenciamiento en tiempo real
Análisis de modificación base utilizando IPD, PW y
contexto de secuencia de acuerdo con la divulgación

FIG. 91

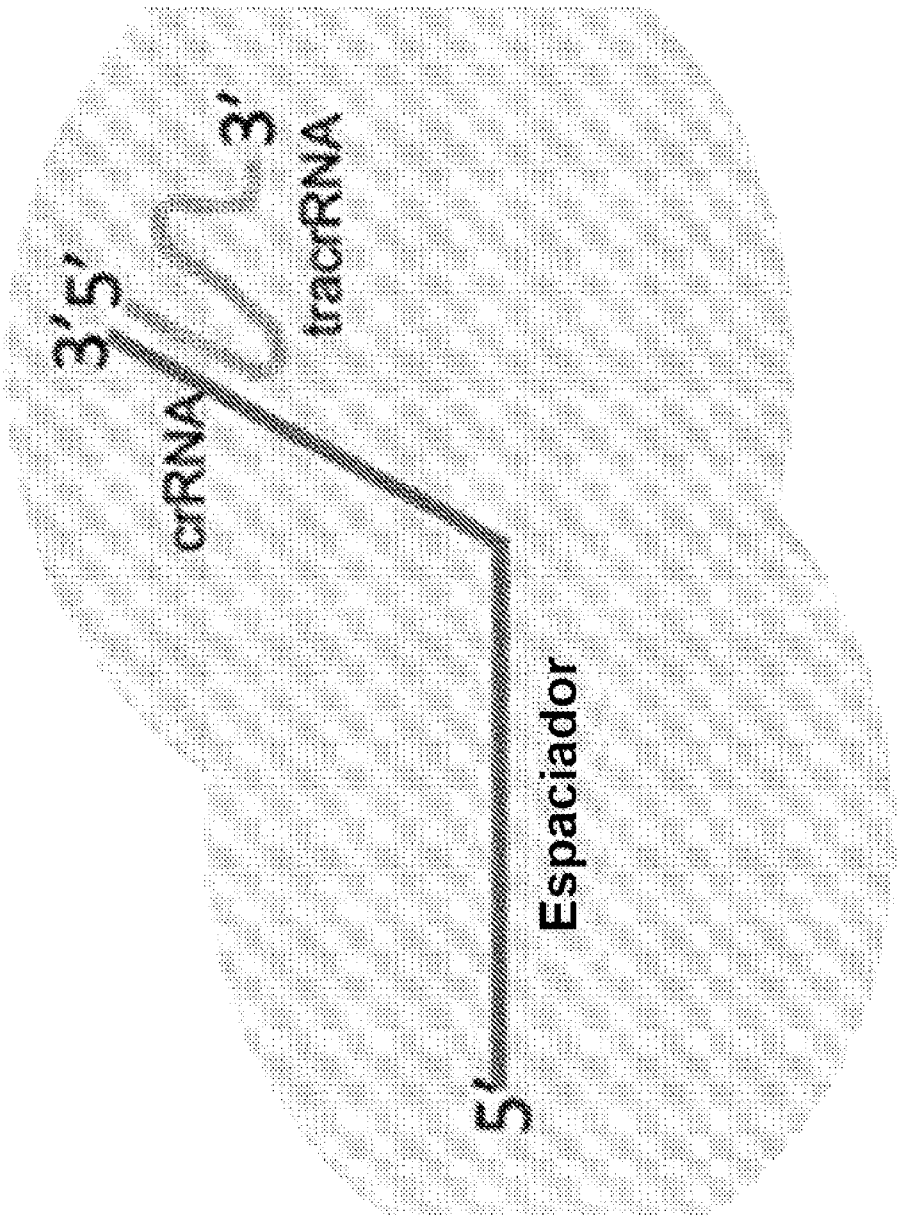


FIG. 92

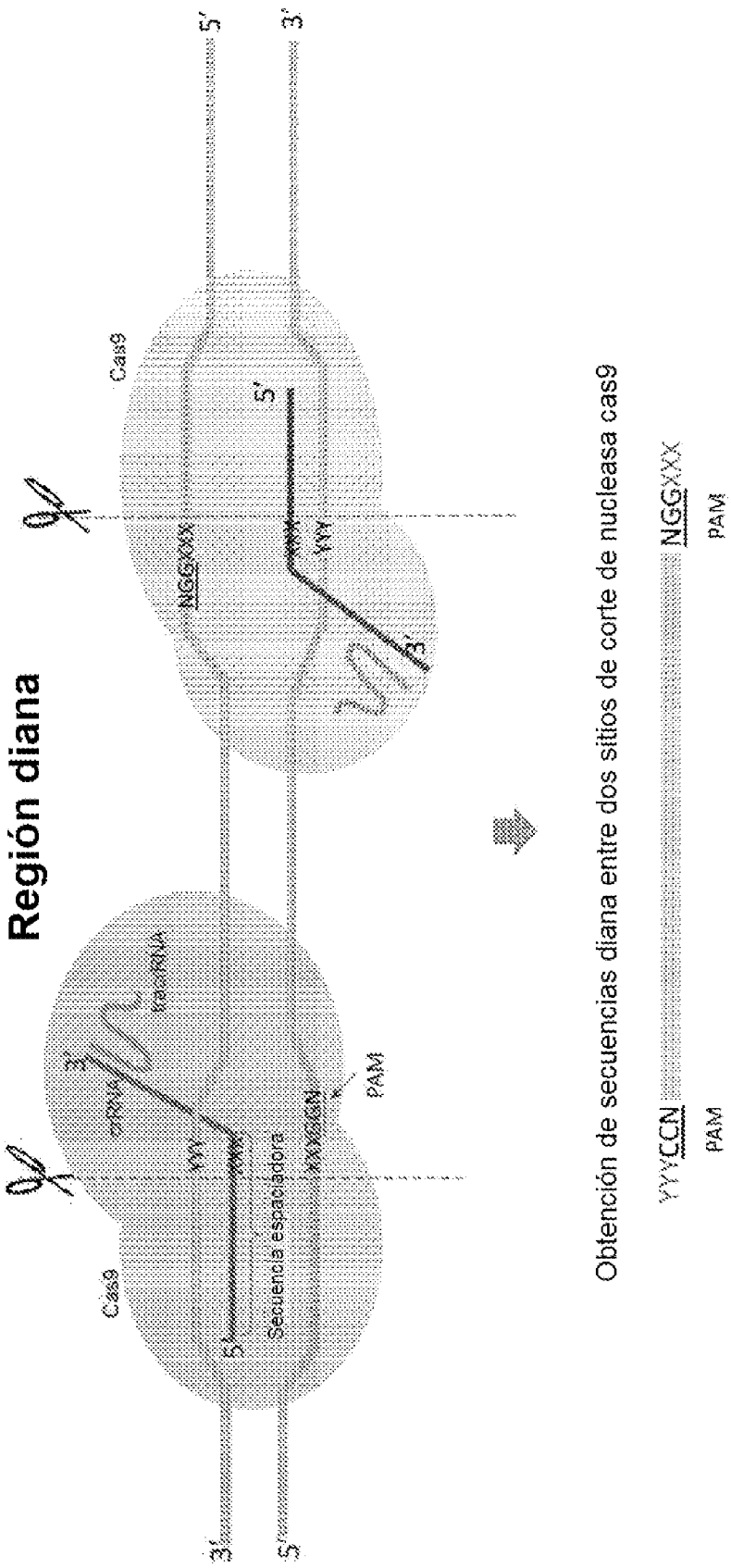

Región diana
Cas9
Cas9
ARNcr
ARNtracr
Secuencia espaciadora
PAM
PAM
Obtención de secuencias diana entre dos sitios de corte de nucleasa cas9
YYYCCN
NGGXXX
PAM

FIG. 93

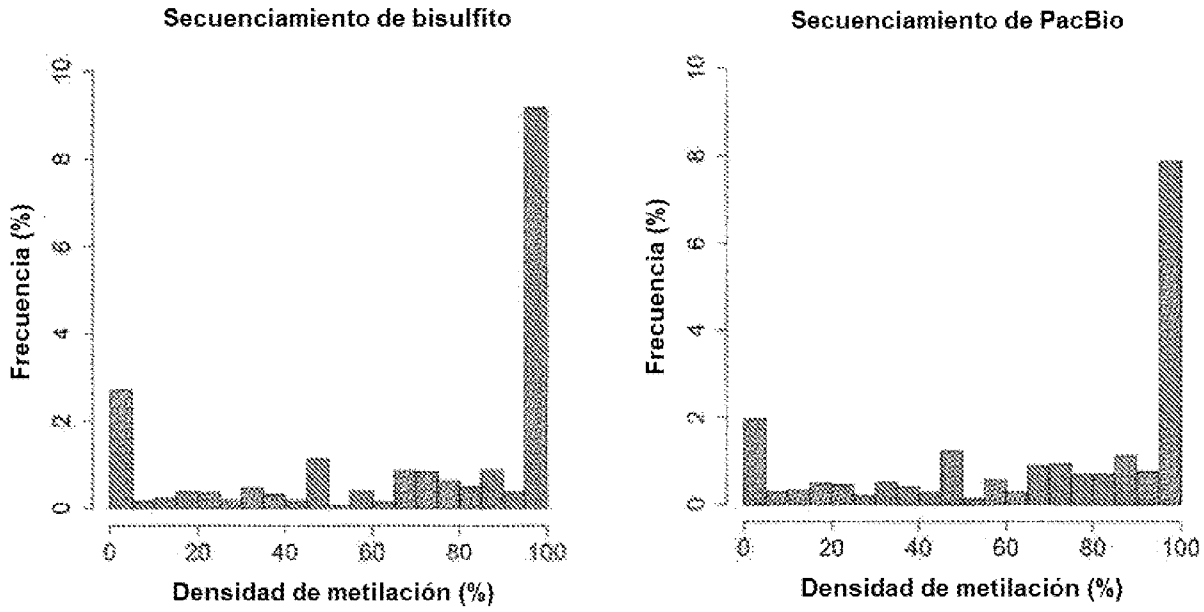

Secuenciamiento de bisulfito
Frecuencia (%)
0
2
4
6
8
10
0
20
40
60
80
100
Densidad de metilación (%)
Secuenciamiento de PacBio
Frecuencia (%)
0
2
4
6
8
10
0
20
40
60
80
100
Densidad de metilación (%)

FIG. 94

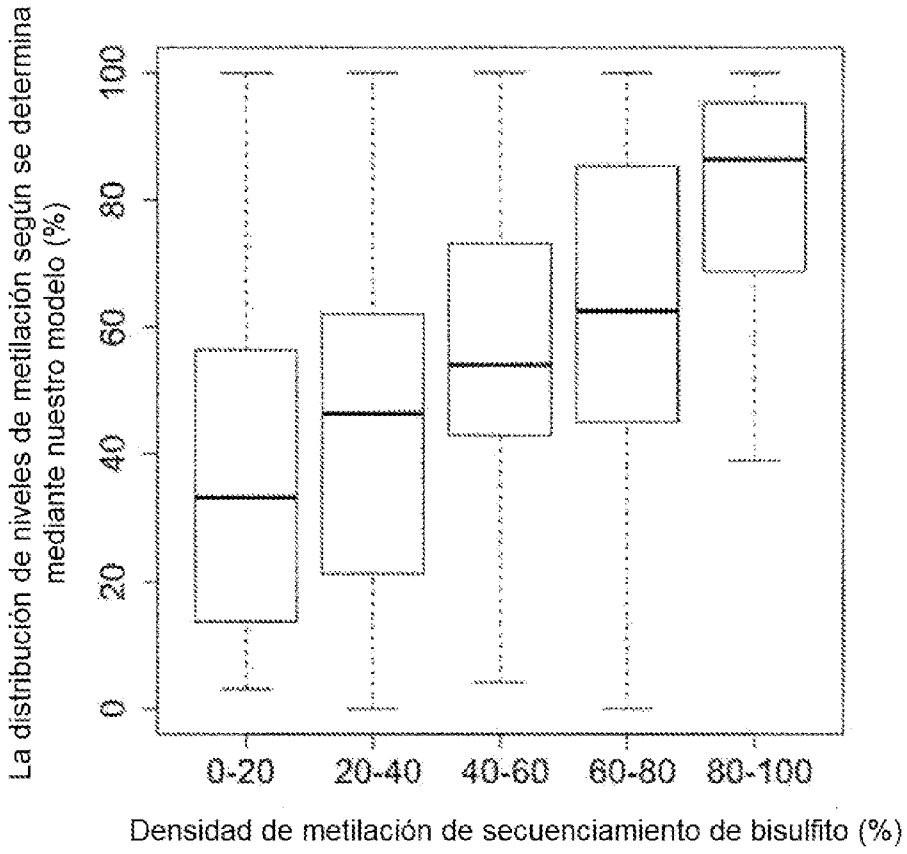

La distribución de niveles de metilación según se determina mediante nuestro modelo (%)
0
20
40
60
80
100
0-20
20-40
40-60
60-80
80-100
Densidad de metilación de secuenciamiento de bisulfito (%)

FIG. 95

Tejidos	Nivel de metilación de Alu (%)
Capa leucocitaria	89.54
Hígado	88.18
Colon	89.56
Pulmón	91.52
Intestino delgado	86.56
Glándula suprarrenal	89.07
Tejido adiposo	91.44
Páncreas	85.82
Cerebro	91.79
HCC	76.74
Placenta	73.04

FIG. 96

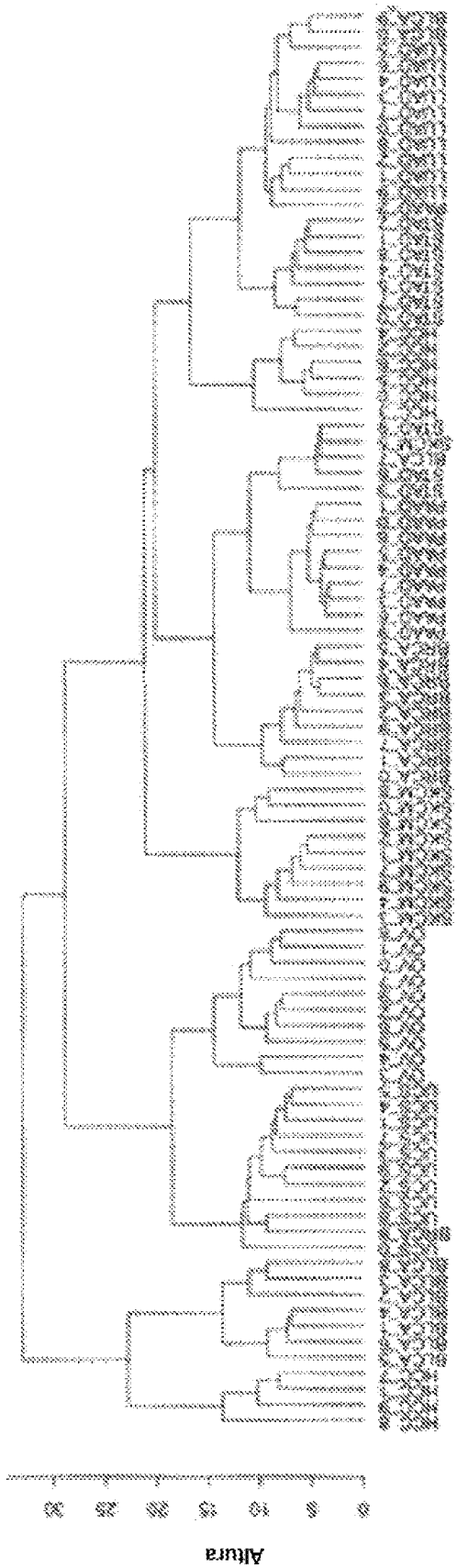

Altura
60
50
40
30
20
10
0
Tipos de cáncer
BLCA: Carcinoma Urotelial de Vejiga
BRCA: Carcinoma Invasivo de Mama
OV: Cistoadenocarcinoma seroso ovárico
PAAD: Adenocarcinoma pancreático
HCC: Carcinoma hepatocelular de hígado
LUAD: Adenocarcinoma de pulmón
STAD: Adenocarcinoma de estómago
SKCM: Melanoma cutáneo de piel
UCS: Carcinosarcoma uterino

FIG. 97

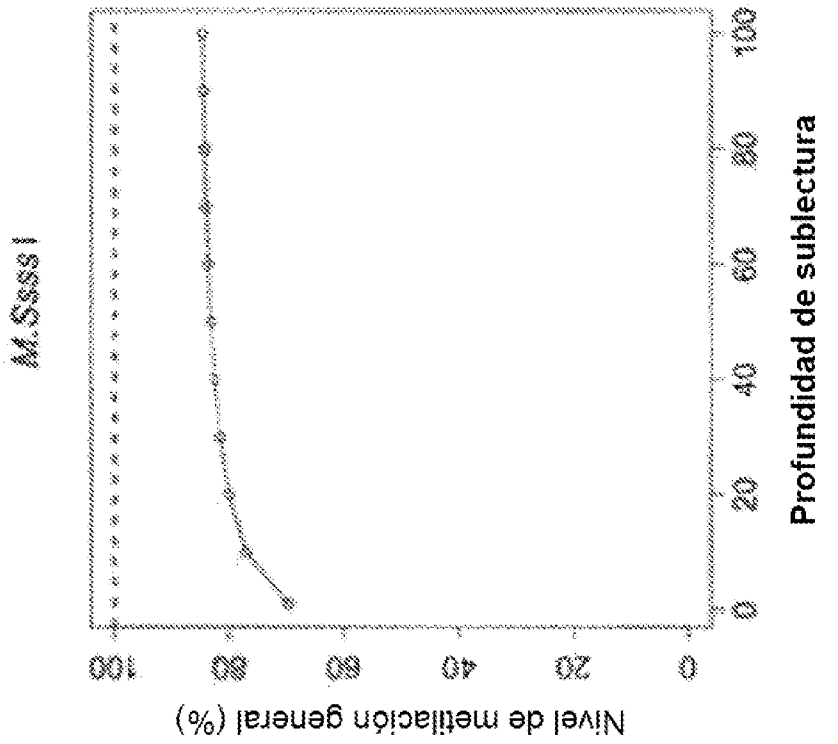


FIG. 98B

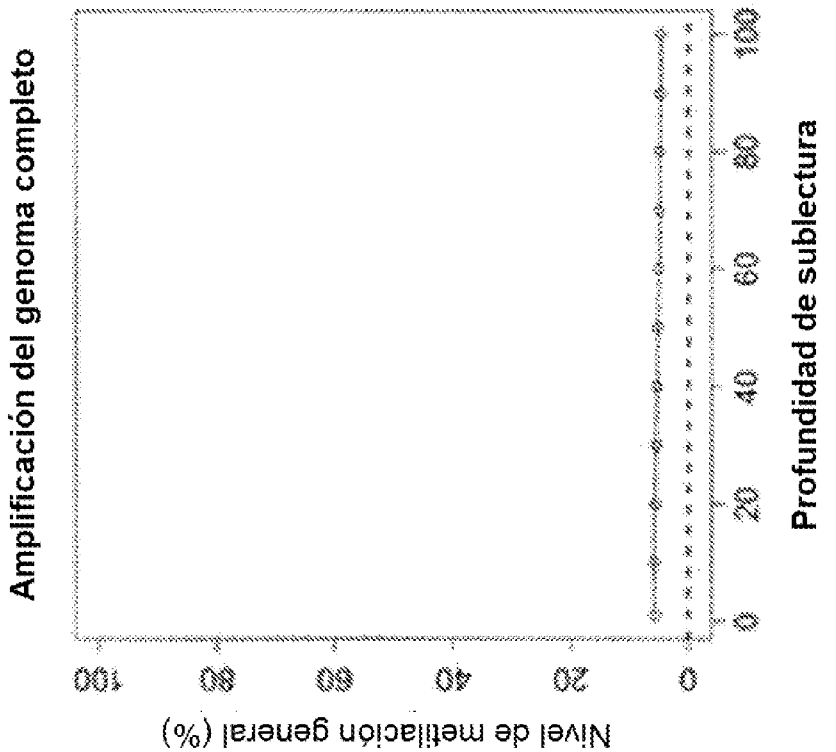


FIG. 98A

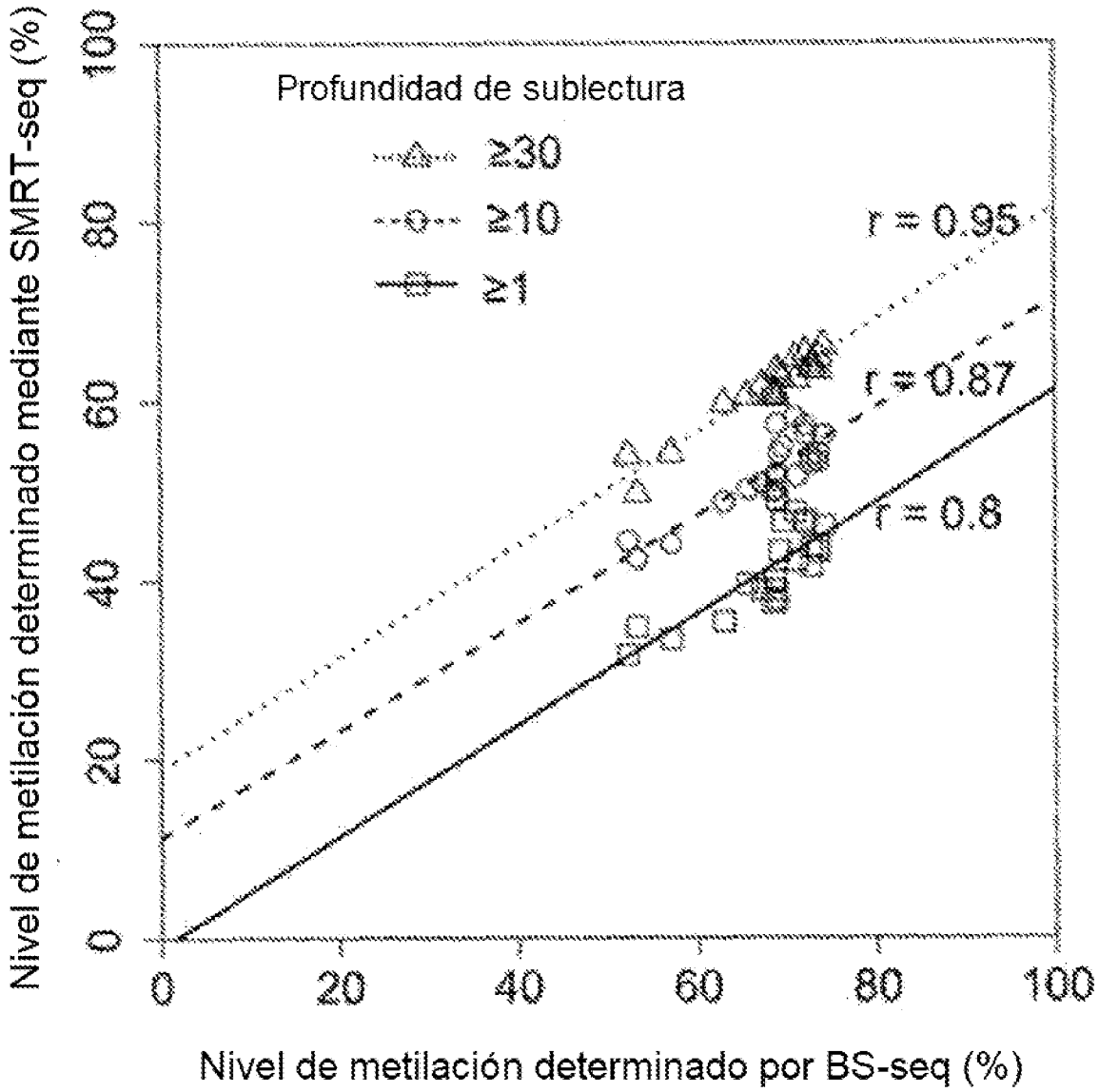

Profundidad de sublectura
≥30
≥10
≥1
r = 0.95
r = 0.87
r = 0.8
Nivel de metilación determinado mediante SMRT-seq (%)
Nivel de metilación determinado por BS-seq (%)
0
20
40
60
80
100

FIG. 99

límites de profundidad de sublectura ≥	r de Pearson (SMRT-seq vs BS-seq)	No. de sitios CpG
1	0.797	25,606,068 (23,949,832-27,008,582)
10	0.873	21,668,418 (18,263,886-23,515,147)
20	0.933	14,276,212 (10,526,406-16,736,887)
30	0.952	6,736,890 (4,255,452-10,449,814)
40	0.948	3,420,790 (2,232,511-5,792,825)
50	0.941	1,684,871 (1,278,475-3,055,876)
60	0.929	911,961 (707,295-1,581,313)
70	0.917	532,422 (350,001-866,045)
80	0.907	284,375 (177,698-534,540)
90	0.906	150,974 (98,000-333,933)
100	0.875	89,788 (58,552-182,861)

FIG. 100

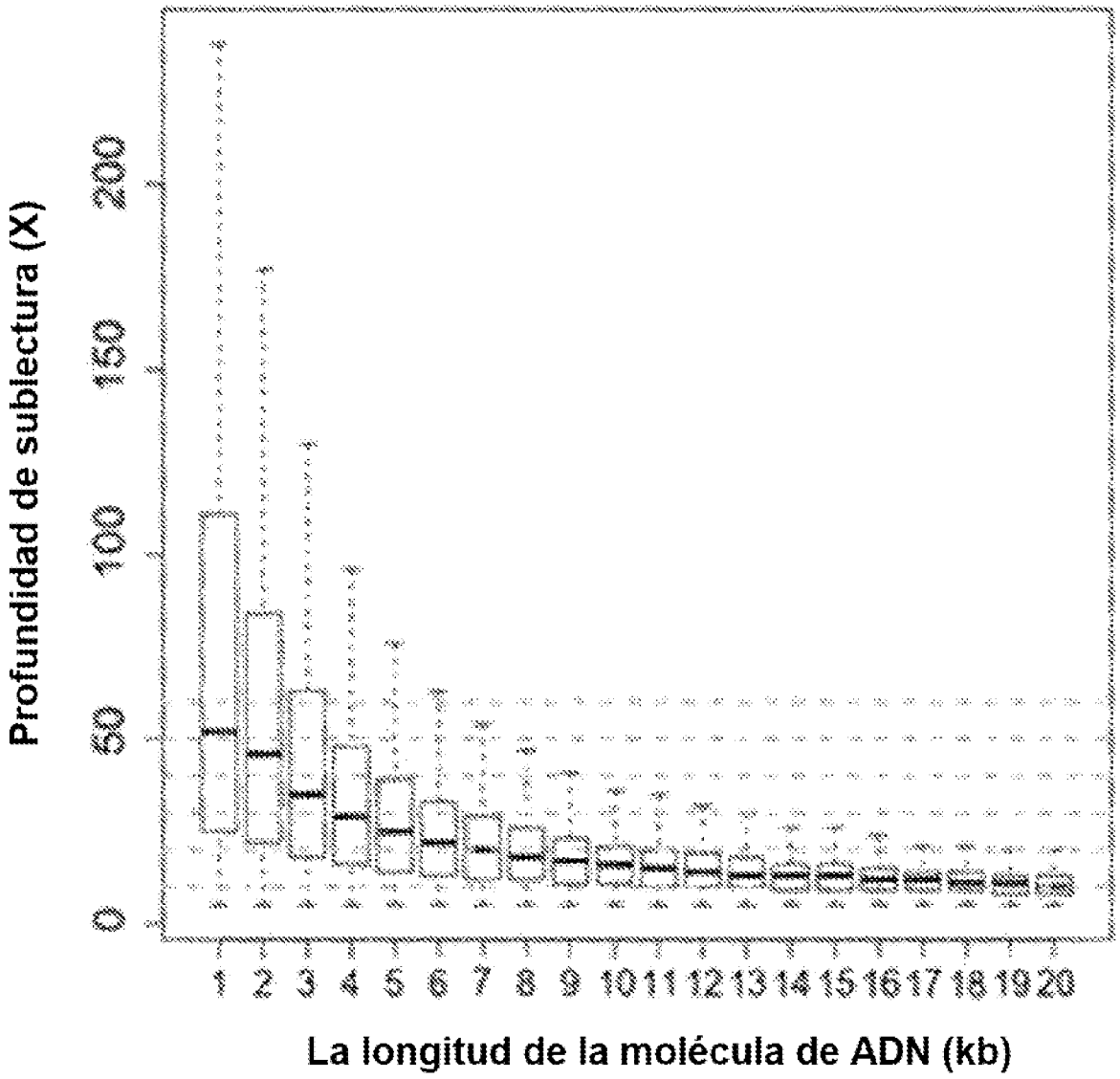

Profundidad de sublectura (X)
0
50
100
150
200
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
La longitud de la molécula de ADN (kb)

FIG. 101

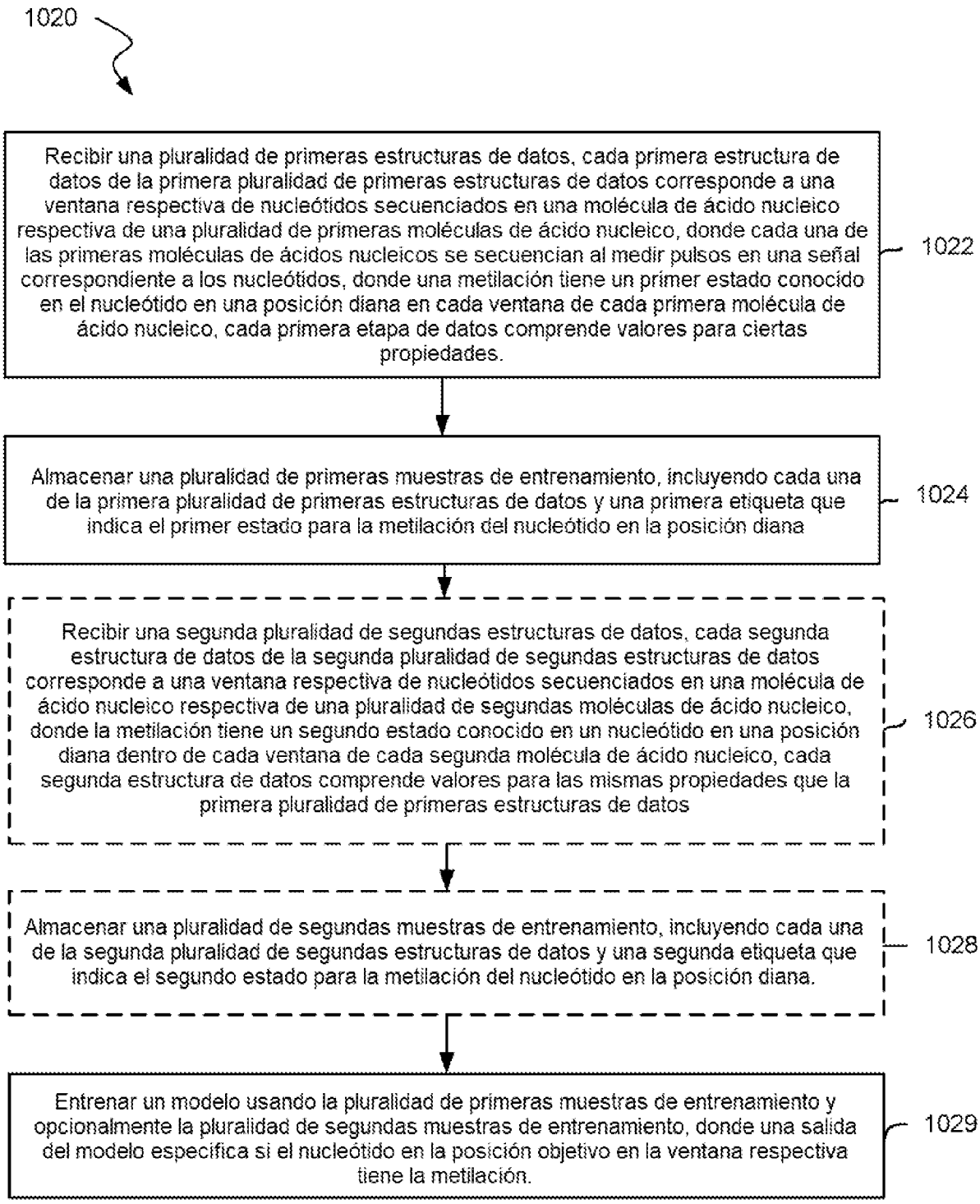

1020
Recibir una pluralidad de primeras estructuras de datos, cada primera estructura de datos de la primera pluralidad de primeras estructuras de datos corresponde a una ventana respectiva de nucleótidos secuenciados en una molécula de ácido nucleico respectiva de una pluralidad de primeras moléculas de ácido nucleico, donde cada una de las primeras moléculas de ácidos nucleicos se secuencian al medir pulsos en una señal correspondiente a los nucleótidos, donde una metilación tiene un primer estado conocido en el nucleótido en una posición diana en cada ventana de cada primera molécula de ácido nucleico, cada primera etapa de datos comprende valores para ciertas propiedades.
1022
Almacenar una pluralidad de primeras muestras de entrenamiento, incluyendo cada una de la primera pluralidad de primeras estructuras de datos y una primera etiqueta que indica el primer estado para la metilación del nucleótido en la posición diana
1024
Recibir una segunda pluralidad de segundas estructuras de datos, cada segunda estructura de datos de la segunda pluralidad de segundas estructuras de datos corresponde a una ventana respectiva de nucleótidos secuenciados en una molécula de ácido nucleico respectiva de una pluralidad de segundas moléculas de ácido nucleico, donde la metilación tiene un segundo estado conocido en un nucleótido en una posición diana dentro de cada ventana de cada segunda molécula de ácido nucleico, cada segunda estructura de datos comprende valores para las mismas propiedades que la primera pluralidad de primeras estructuras de datos
1026
Almacenar una pluralidad de segundas muestras de entrenamiento, incluyendo cada una de la segunda pluralidad de segundas estructuras de datos y una segunda etiqueta que indica el segundo estado para la metilación del nucleótido en la posición diana.
1028
Entrenar un modelo usando la pluralidad de primeras muestras de entrenamiento y opcionalmente la pluralidad de segundas muestras de entrenamiento, donde una salida del modelo especifica si el nucleótido en la posición objetivo en la ventana respectiva tiene la metilación.
1029

FIG. 102

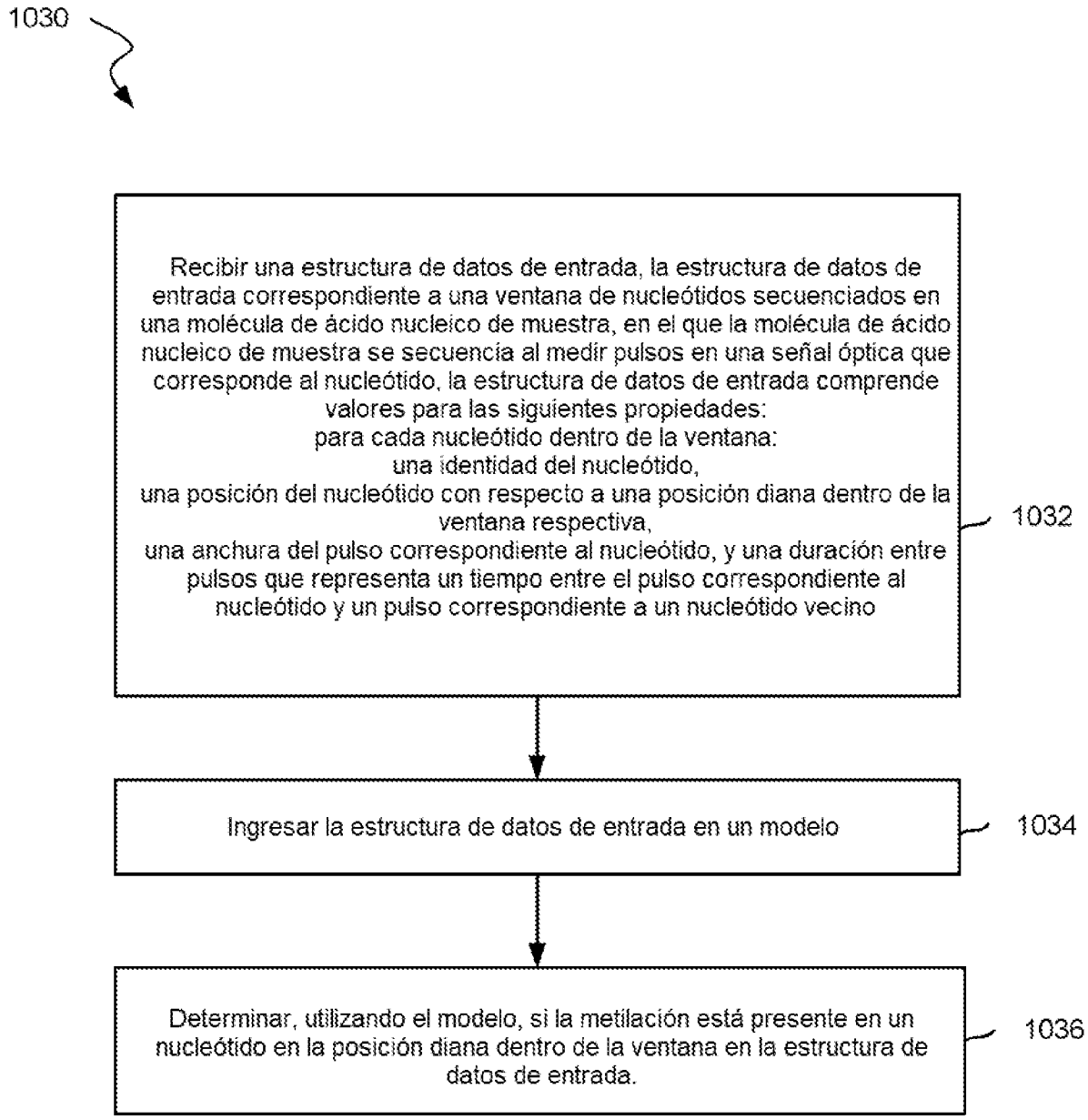
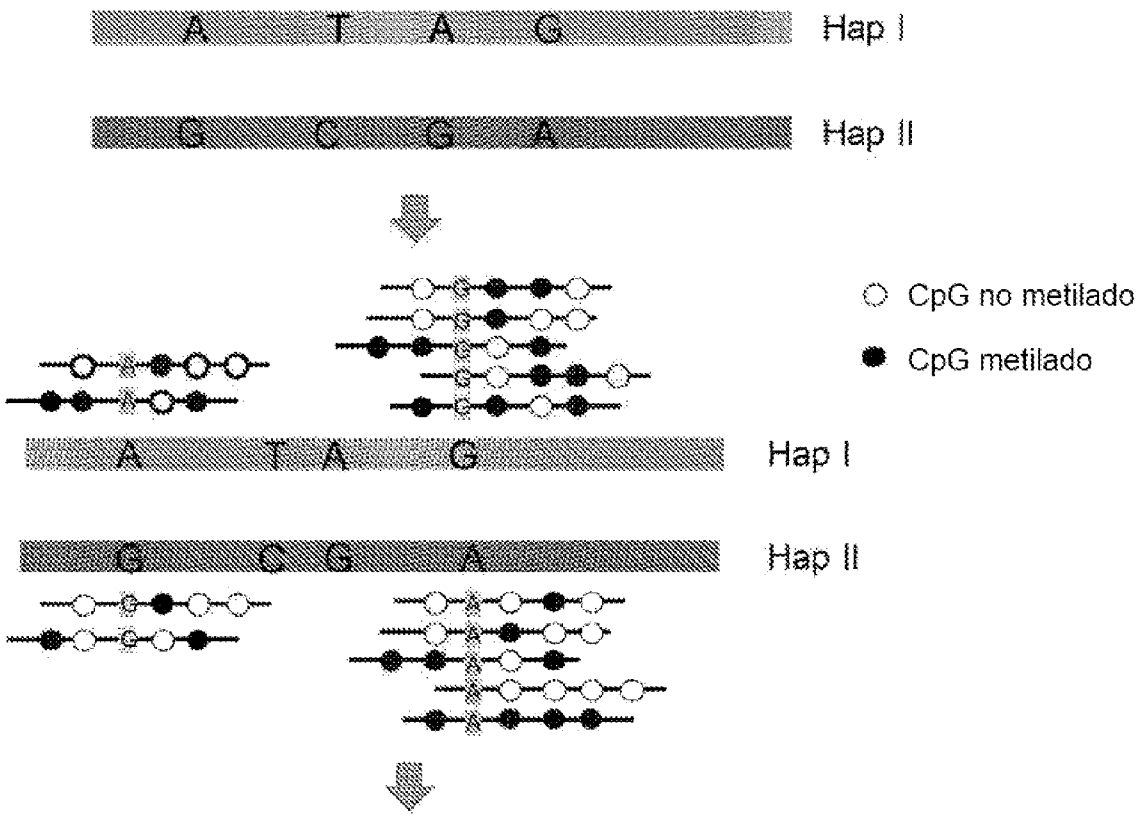

1030
Recibir una estructura de datos de entrada, la estructura de datos de entrada correspondiente a una ventana de nucleótidos secuenciados en una molécula de ácido nucleico de muestra, en el que la molécula de ácido nucleico de muestra se secuencia al medir pulsos en una señal óptica que corresponde al nucleótido, la estructura de datos de entrada comprende valores para las siguientes propiedades:
para cada nucleótido dentro de la ventana:
una identidad del nucleótido,
una posición del nucleótido con respecto a una posición diana dentro de la ventana respectiva,
una anchura del pulso correspondiente al nucleótido, y una duración entre pulsos que representa un tiempo entre el pulso correspondiente al nucleótido y un pulso correspondiente a un nucleótido vecino
1032
Ingresar la estructura de datos de entrada en un modelo
1034
Determinar, utilizando el modelo, si la metilación está presente en un nucleótido en la posición diana dentro de la ventana en la estructura de datos de entrada.
1036

FIG. 103



Índice de desequilibrio de metilación basado en haplotipo relativo

por ejemplo nivel de metilación de Hap I – nivel de metilación de Hap II

FIG. 104

Chr	Inicio	Fin	Longitud	Id de bloque de haplotipo	Secuenciamiento PacBio			
					Nivel de metilación en tejido no tumoral adyacente		Nivel de metilación en tejido tumoral	
					Hap I	Hap II	Hap I	Hap II
chr1	56312395	56347696	35301	hap1927	68.2	67.4	60.3	23.5
chr1	194413819	194424806	10987	hap5953	52.8	49.5	48.8	9.3
chr1	220674478	220699011	24533	hap6863	63.0	64.5	50.4	17.3
chr10	113088792	113124248	35456	hap11838	62.7	63.4	38.1	5.7
chr11	5482746	5498801	16055	hap12904	70.3	75.0	16.3	51.7
chr11	42819351	42852772	33421	hap14385	54.6	54.9	65.3	17.8
chr11	57983961	58051078	67117	hap14930	67.3	66.4	58.2	18.6
chr11	60174708	60204209	29501	hap14990	58.4	59.8	49.6	10.8
chr12	128079419	128114656	35237	hap22249	60.0	58.3	12.1	45.2
chr15	20480575	20533464	52889	hap29631	64.7	69.1	27.7	59.3
chr15	94902853	94946231	43378	hap32161	74.1	74.5	74.9	15.8
chr15	96526684	96549225	22541	hap32221	70.8	68.8	28.9	64.4
chr16	31595372	31613277	17905	hap33499	55.9	59.3	46.3	14.4
chr16	80151778	80182097	30319	hap34821	71.1	71.0	11.5	51.8
chr16	82519715	82554191	34476	hap34920	71.3	66.5	47.4	13.0
chr17	21668593	21685572	16979	hap36049	50.3	47.8	67.4	19.6
chr17	44999177	45012087	12910	hap36640	47.1	45.2	81.6	35.1
chr17	69911623	69926625	15002	hap37435	67.3	63.0	37.8	5.2
chr18	11441122	11458521	17399	hap38335	65.5	66.8	65.9	22.4
chr18	23405569	23423387	17818	hap38673	66.3	61.7	3.3	48.1
chr18	68887284	68925031	37747	hap40390	63.0	61.0	22.0	53.4
chr18	69487809	69505470	17661	hap40414	74.5	74.1	33.3	72.2
chr2	41480394	41514135	33741	hap43972	54.0	54.0	14.9	77.8
chr2	114171214	114182880	11666	hap46226	72.4	68.8	79.7	16.7
chr2	123762541	123797629	35088	hap46589	66.7	68.3	24.0	54.5
chr2	125236882	125241950	5068	hap46673	58.9	59.2	10.7	46.4
chr2	130016110	130040331	24221	hap46835	54.6	50.8	5.6	41.6
chr2	137757638	137783716	26078	hap47090	61.8	61.4	13.5	69.2
chr2	144128597	144160845	32248	hap47343	65.8	66.6	9.3	50.3
chr20	15736792	15753459	16667	hap51505	78.9	74.3	45.8	77.3
chr20	26167979	26177235	9256	hap51868	55.0	52.2	38.5	68.6
chr20	44255808	44264190	8382	hap52246	57.4	56.1	9.7	50.6
chr20	59518410	59559273	40863	hap52761	61.0	62.4	30.0	72.8
chr21	21402034	21424129	22095	hap53197	63.5	67.3	25.0	75.5
chr21	24750027	24768793	18766	hap53333	68.2	64.6	3.4	38.9
chr21	26666833	26701575	34742	hap53418	62.1	66.5	47.6	16.7
chr3	2364024	2387896	23872	hap55539	67.4	67.8	54.9	10.9
chr3	21036965	21049451	12486	hap56223	54.8	51.4	53.1	21.1
chr3	56011690	56046642	34952	hap57346	64.2	61.2	71.2	22.6

FIG. 105A

chr3	73330942	73371216	40274	hap57939	60.9	62.9	9.4	42.9
chr3	106372440	106401301	28861	hap59077	67.8	67.9	13.8	53.2
chr3	107772994	107807482	34488	hap59122	69.6	73.5	30.4	66.4
chr3	116742501	116776747	34246	hap59493	64.3	69.1	14.1	51.6
chr3	171076306	171100102	23796	hap61495	68.0	66.0	80.6	48.8
chr3	193058272	193080344	22072	hap62231	65.5	64.7	54.6	20.0
chr4	30411613	30432317	20704	hap63589	59.3	60.6	53.4	14.6
chr4	31304718	31338193	33475	hap63633	60.2	60.0	7.2	55.0
chr4	92003467	92030505	27038	hap65794	65.3	65.1	54.1	21.7
chr4	155224697	155250915	26218	hap68104	60.5	57.5	57.3	25.0
chr5	2281802	2299281	17479	hap69632	71.5	66.9	69.9	6.6
chr5	4624948	4664704	39756	hap69739	62.8	61.0	14.0	52.0
chr5	89593236	89606080	12844	hap72628	76.6	74.0	20.3	78.4
chr5	119214026	119233058	19032	hap73698	62.8	61.2	57.6	13.1
chr5	119940397	119972658	32261	hap73720	59.1	54.7	53.8	12.2
chr5	132859668	132877415	17747	hap74150	62.5	66.6	59.5	28.3
chr6	26914610	26936918	22308	hap76887	41.9	40.9	71.9	32.6
chr6	66879106	66957243	78137	hap78266	61.6	59.6	23.4	62.0
chr6	77349083	77377529	28446	hap78674	64.5	66.4	27.0	62.9
chr6	159738794	159751033	12239	hap81616	79.6	79.0	21.2	59.8
chr7	26585255	26641907	56652	hap83161	66.2	64.7	49.4	13.3
chr7	48214640	48248036	33396	hap84003	76.0	76.7	78.0	32.3
chr7	88558182	88575482	17300	hap85335	63.8	59.6	63.8	22.9
chr7	96588562	96607580	19018	hap85620	60.4	63.1	19.7	50.0
chr7	122942180	122956897	14717	hap86454	42.3	39.0	19.2	50.0
chr7	132321970	132344802	22832	hap86807	61.4	60.7	52.5	11.5
chr7	153296219	153302441	6222	hap87487	48.7	53.7	64.4	19.3
chr7	156356247	156371897	15650	hap87631	74.9	71.6	87.5	56.6
chr7	159091986	159119486	27500	hap87738	54.0	49.1	52.0	13.2
chr8	51530582	51550889	20307	hap89477	66.4	65.7	68.0	19.9
chr8	63513932	63537543	23611	hap89942	62.0	63.3	11.6	48.4
chr8	72373321	72398122	24801	hap90226	58.0	54.9	71.6	32.0
chr8	94100451	94141855	41404	hap90991	65.2	65.7	36.2	68.7
chr8	109300499	109326404	25905	hap91510	63.6	67.7	29.3	65.8

FIG. 105B

Chr	Inicio	Fin	Longitud	Id de bloque de haplotipo	Secuenciamiento PacBio			
					Nivel de metilación en tejido no tumoral adyacente		Nivel de metilación en tejido tumoral	
					Hap I	Hap II	Hap I	Hap II
chr9	27803548	27888202	84654	hap58508	64.2	60.9	20.6	75.4
chr6	242149	386636	144487	hap47880	62.3	63.3	77.4	32.2
chr5	28219159	28302858	83699	hap44666	59.3	58.0	16.8	58.2
chr5	18119943	18153743	33800	hap44475	61.6	65.0	53.2	21.7
chr7	24906307	25046195	139888	hap52069	69.3	68.7	44.0	76.2
chr15	27689897	27752573	62676	hap18337	65.9	61.9	64.8	20.5
chr12	42183870	42212433	28563	hap12045	63.5	68.4	19.4	51.2
chr21	9825597	9935752	110155	hap34175	54.3	53.5	60.9	29.1
chr2	118813055	118893366	80311	hap30060	62.6	62.3	77.0	38.6
chr6	90307702	90344869	37167	hap49779	69.1	66.4	84.7	53.9
chr7	107932914	108049376	116462	hap53838	67.2	62.9	43.8	76.4
chr7	137039327	137160933	121606	hap54447	59.5	60.9	22.9	72.0
chr17	21193754	21254930	61176	hap22633	59.2	54.3	69.7	31.6
chr12	11473697	11644714	171017	hap11451	62.8	66.4	35.5	75.9
chr5	129212299	129353349	141050	hap46632	50.9	54.5	45.5	14.0
chr11	93910738	94028887	118149	hap10288	67.6	63.6	36.6	74.2
chr3	131707434	132003636	296202	hap38642	57.8	55.9	17.9	60.2
chr3	43024004	43161785	137781	hap36769	69.1	66.5	46.1	80.2
chr3	190403156	190606658	203502	hap39947	60.9	61.6	36.9	72.7
chr15	40218970	40279780	60810	hap18606	53.4	57.5	79.1	47.4

FIG. 106

Tipos de tejido	No. de bloques de haplotipo que muestran desequilibrio de metilación entre dos haplotipos en tejidos de tumor	No. de bloques de haplotipo que muestran desequilibrio de metilación entre dos haplotipos en tejidos no tumorales adyacentes pareados
Colon	92	47
Mama	57	13
Riñón	68	18
Pulmón	31	21
Próstata	26	19
Estómago	2	0

200

ES 2 985 191 T3

FIG. 107A

Tipos de tejido	No. de bloques de haplotipo que muestran desequilibrio de metilación entre dos haplotipos en tejidos de tumor	Información sobre estadificación del tumor (TNM) disponible
Mama	18	T2
	57	T3
Riñón	68	T3a
	0	T2

FIG. 107B

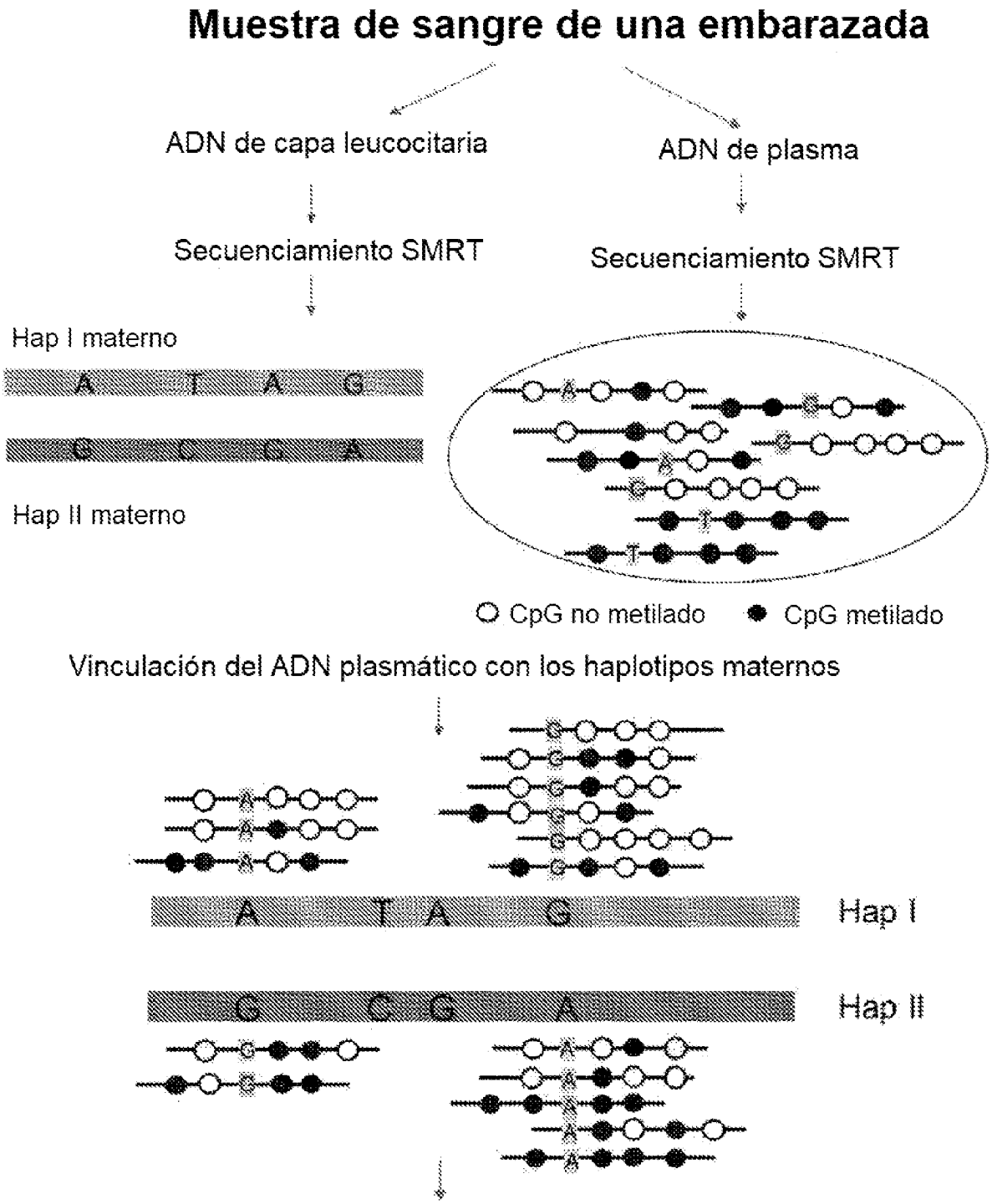

Muestra de sangre de una embarazada
ADN de capa leucocitaria
ADN de plasma
Secuenciamiento SMRT
Secuenciamiento SMRT
Hap I materno
A T A G
G C G A
Hap II materno
CpG no metilado
CpG metilado
Vinculación del ADN plasmático con los haplotipos maternos
A T A G
Hap I
G C G A
Hap II
Nivel de metilación de Hap I < nivel de metilación de Hap II
El feto tiene Hap I heredado

FIG. 108

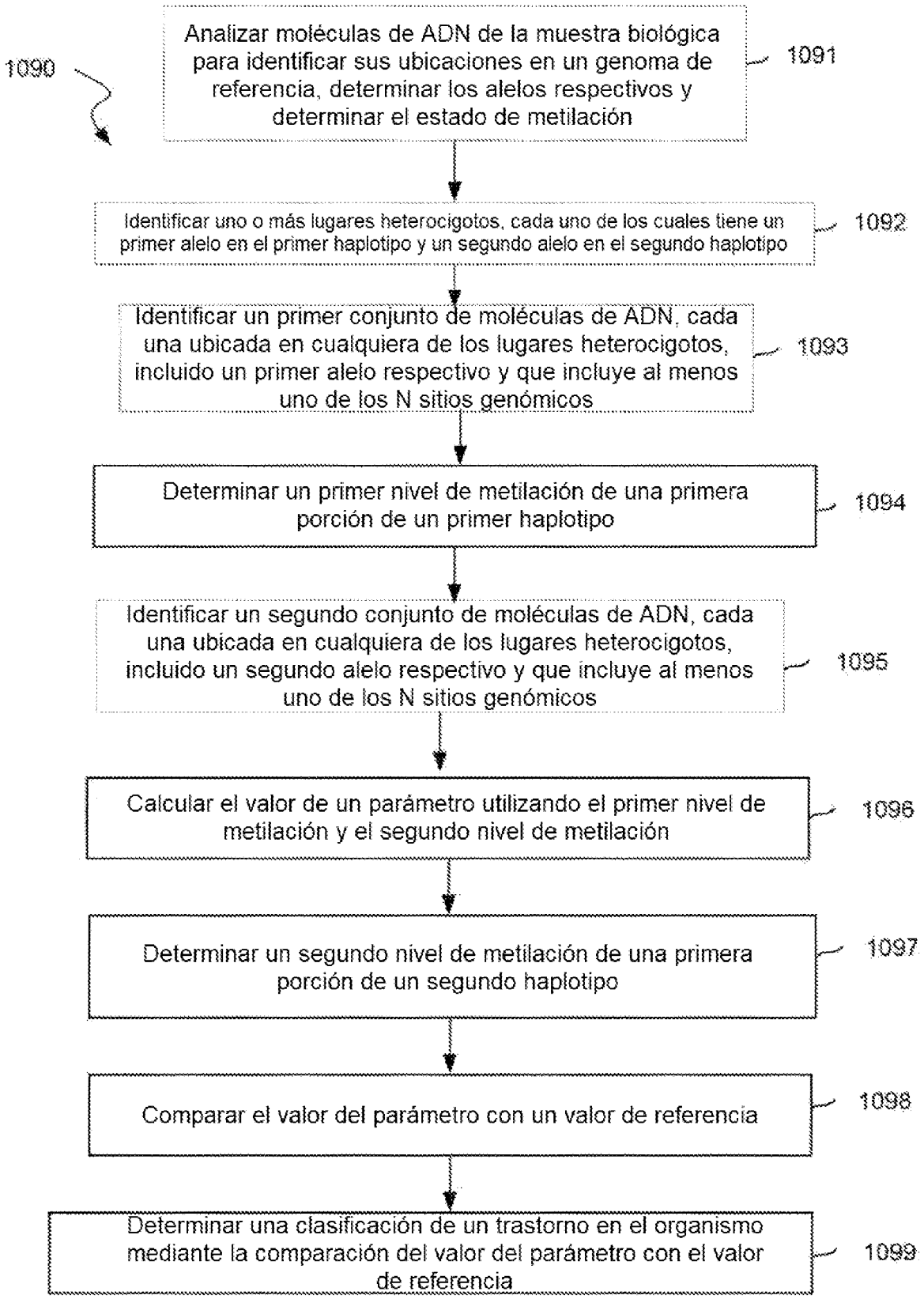

1090
Analizar moléculas de ADN de la muestra biológica para identificar sus ubicaciones en un genoma de referencia, determinar los alelos respectivos y determinar el estado de metilación
1091
Identificar uno o más lugares heterocigotos, cada uno de los cuales tiene un primer alelo en el primer haplotipo y un segundo alelo en el segundo haplotipo
1092
Identificar un primer conjunto de moléculas de ADN, cada una ubicada en cualquiera de los lugares heterocigotos, incluido un primer alelo respectivo y que incluye al menos uno de los N sitios genómicos
1093
Determinar un primer nivel de metilación de una primera porción de un primer haplotipo
1094
Identificar un segundo conjunto de moléculas de ADN, cada una ubicada en cualquiera de los lugares heterocigotos, incluido un segundo alelo respectivo y que incluye al menos uno de los N sitios genómicos
1095
Calcular el valor de un parámetro utilizando el primer nivel de metilación y el segundo nivel de metilación
1096
Determinar un segundo nivel de metilación de una primera porción de un segundo haplotipo
1097
Comparar el valor del parámetro con un valor de referencia
1098
Determinar una clasificación de un trastorno en el organismo mediante la comparación del valor del parámetro con el valor de referencia
1099

FIG. 109

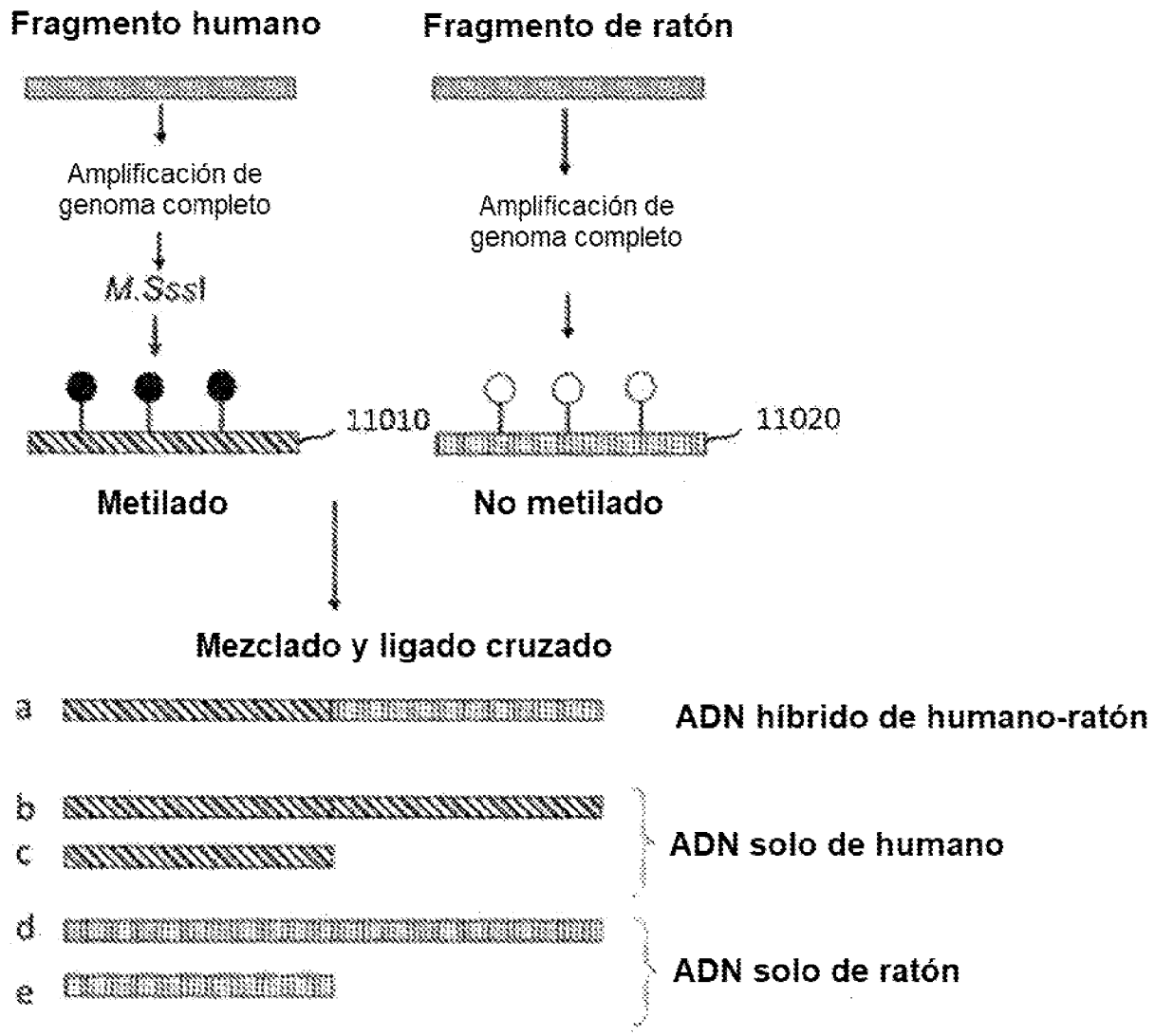

Fragmento humano
Fragmento de ratón
Amplificación de genoma completo
M.SssI
Amplificación de genoma completo
11010
11020
Metilado
No metilado
Mezclado y ligado cruzado
a
ADN híbrido de humano-ratón
b
c
ADN solo de humano
d
e
ADN solo de ratón

FIG. 110

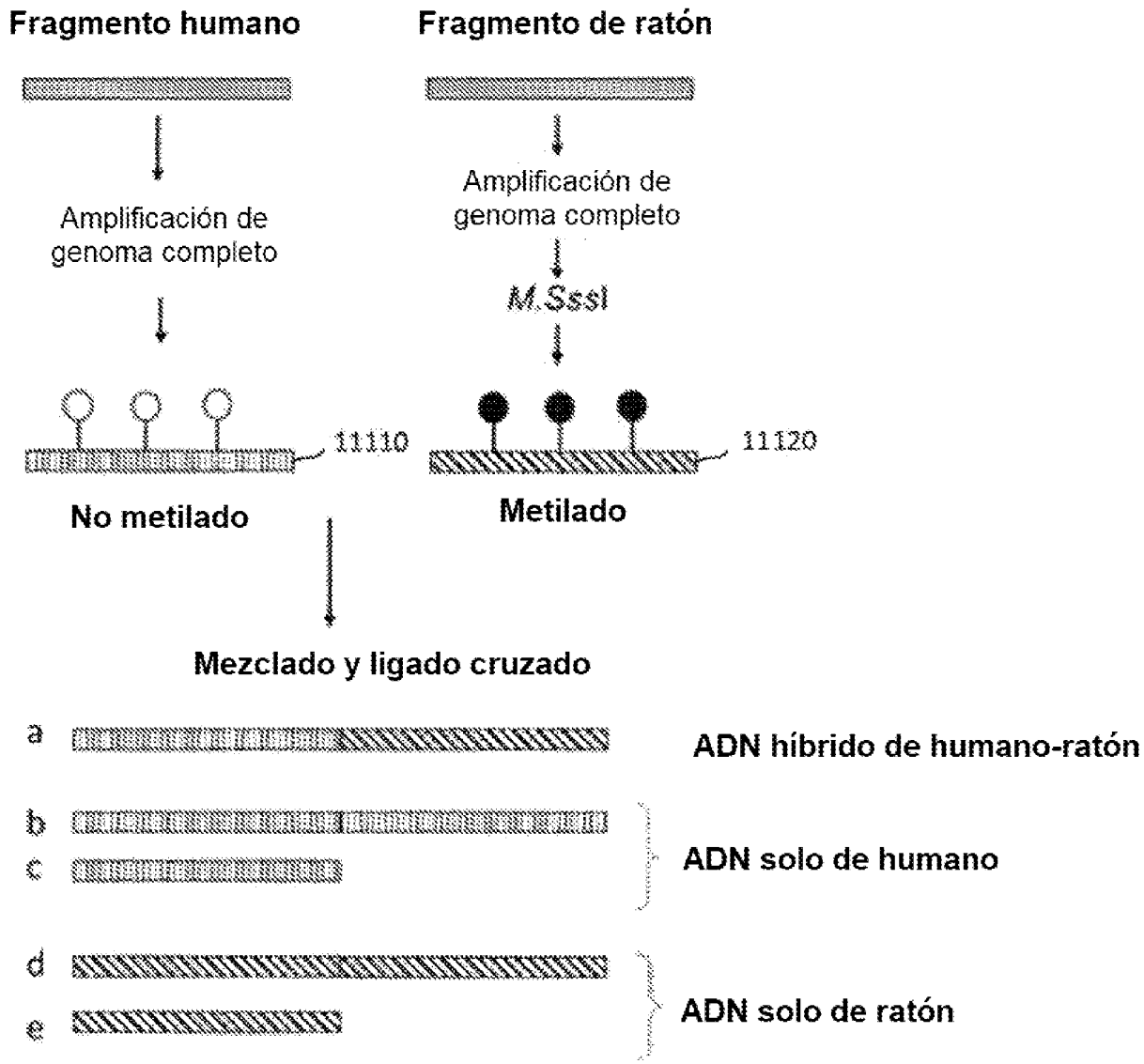

Fragmento humano
Fragmento de ratón
Amplificación de genoma completo
Amplificación de genoma completo
M.SssI
11110
11120
No metilado
Metilado
Mezclado y ligado cruzado
a
b
c
d
e
ADN híbrido de humano-ratón
ADN solo de humano
ADN solo de ratón

FIG. 111

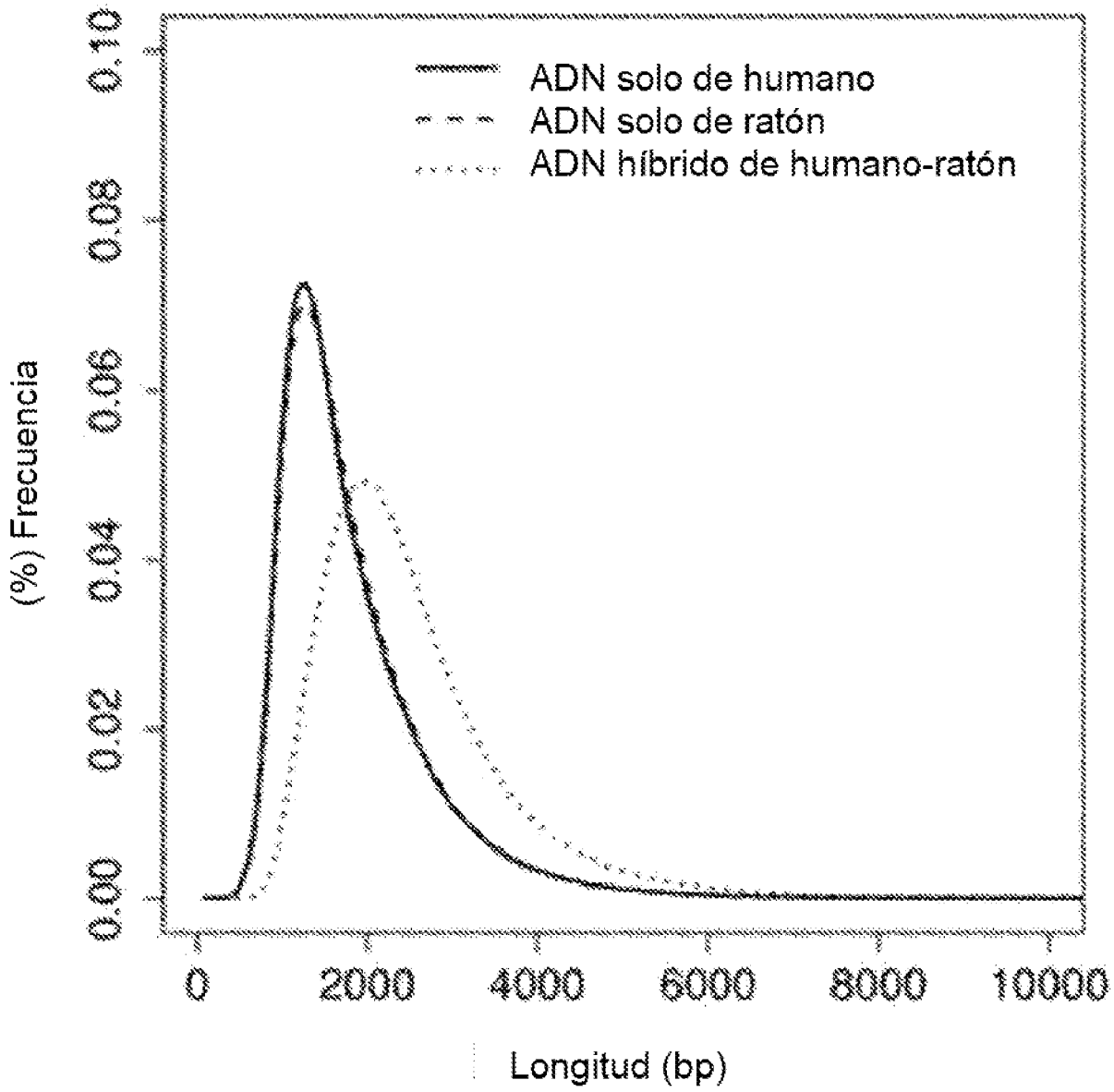

ADN solo de humano
ADN solo de ratón
ADN híbrido de humano-ratón
(%) Frecuencia
0.00
0.02
0.04
0.06
0.08
0.10
0
2000
4000
6000
8000
10000
Longitud (bp)

FIG. 112

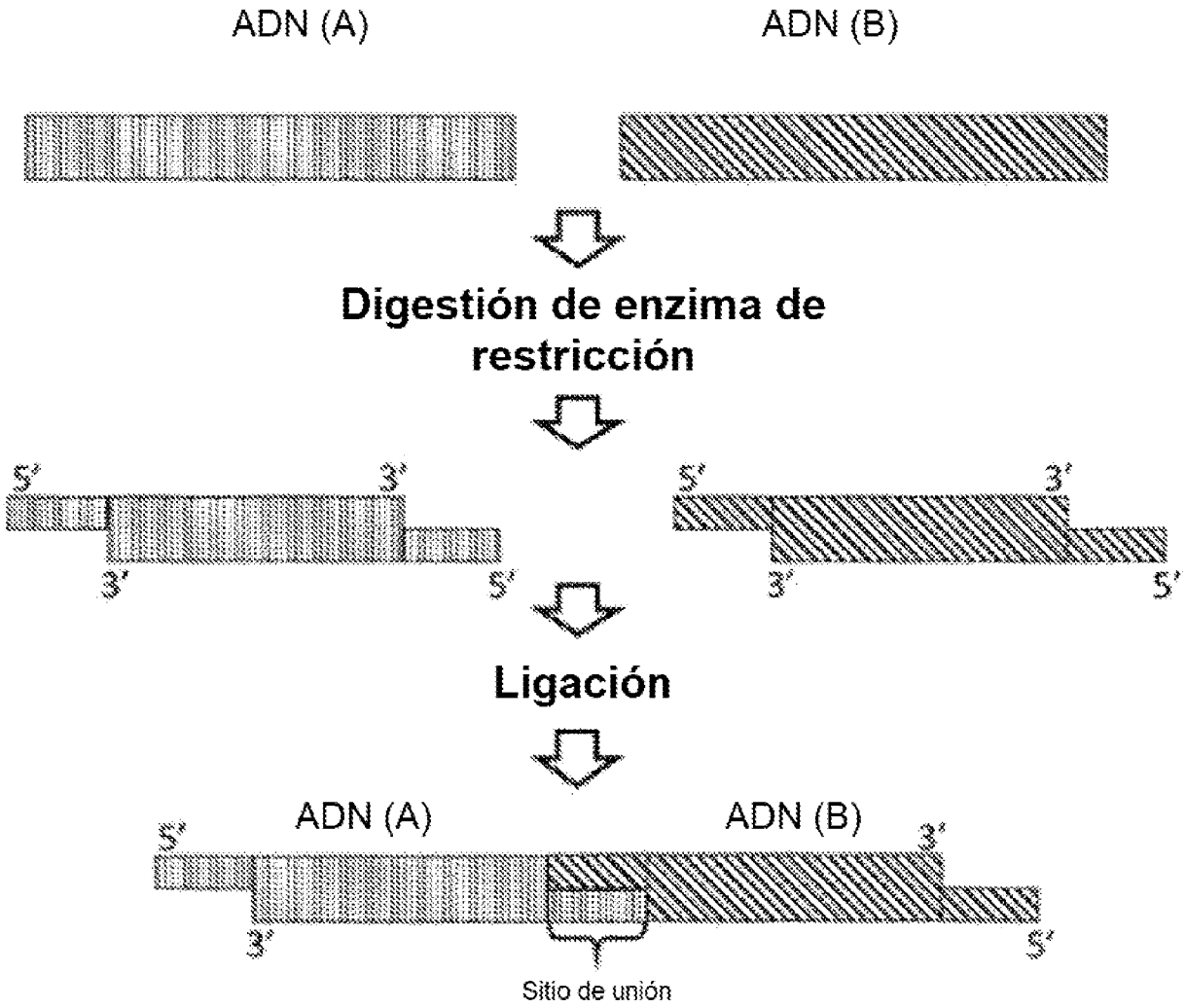

ADN (A)
ADN (B)
Digestión de enzima de restricción
5'
3'
3'
5'
5'
3'
3'
5'
Ligación
ADN (A)
ADN (B)
5'
3'
3'
5'
Sitio de unión

FIG. 113

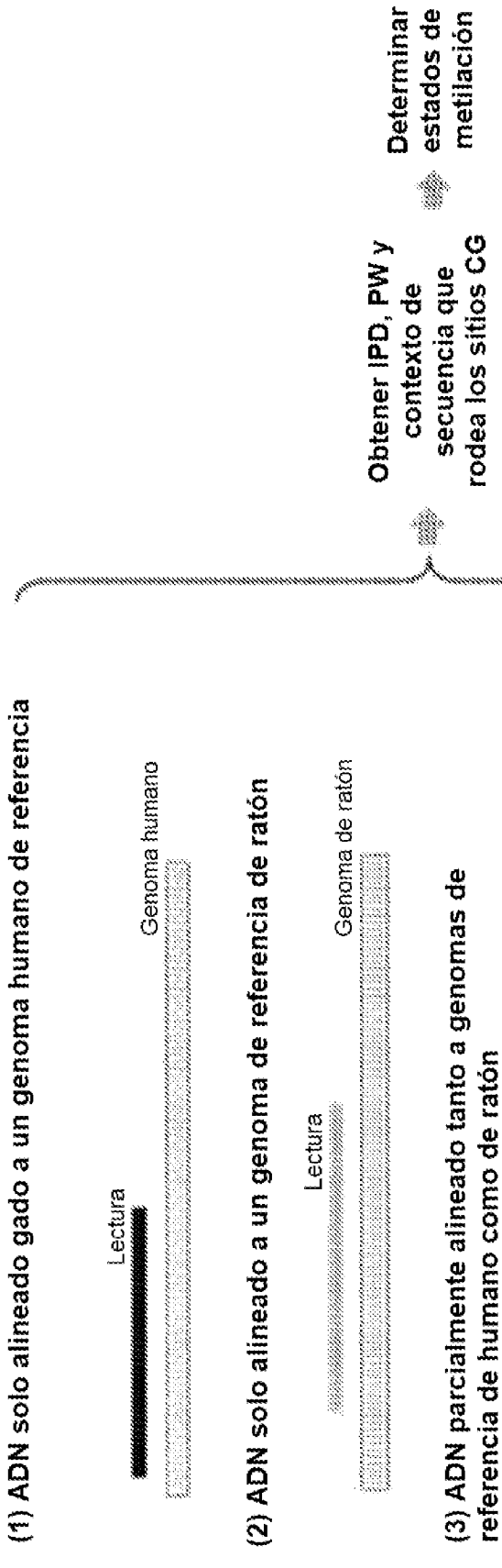

(1) ADN solo alineado gado a un genoma humano de referencia
Lectura
Genoma humano
(2) ADN solo alineado a un genoma de referencia de ratón
Lectura
Genoma de ratón
(3) ADN parcialmente alineado tanto a genomas de referencia de humano como de ratón
Sitio de unión Sitio RE 11410
Lectura
Parte de ADN de humano
Ratón
Parte de ADN de ratón
Genoma humano
Genoma de ratón
Obtener IPD, PW y contexto de secuencia que rodea los sitios CG
Determinar estados de metilación

FIG. 114

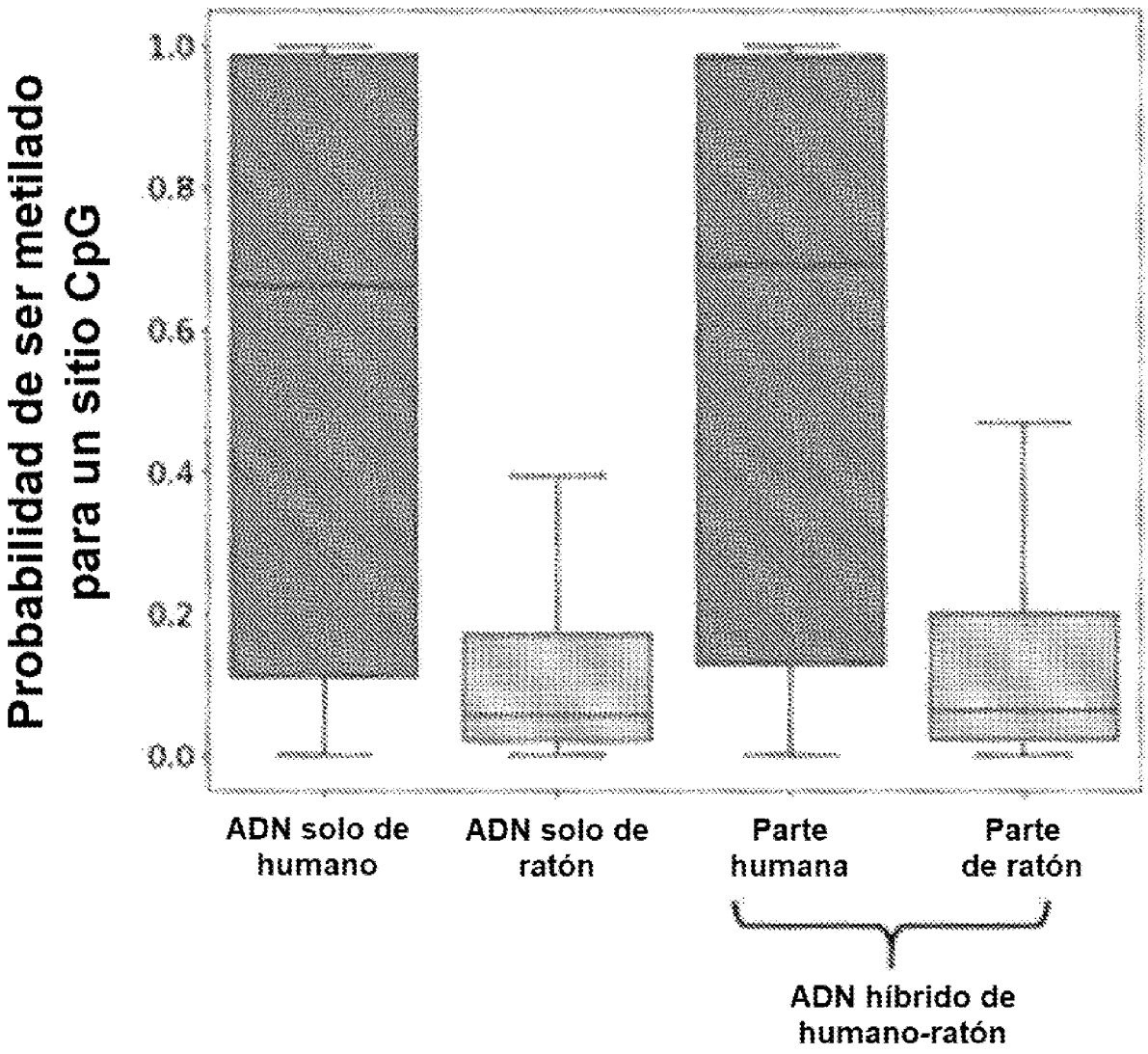

Probabilidad de ser metilado para un sitio CpG
1.0
0.8
0.6
0.4
0.2
0.0
ADN solo de humano
ADN solo de ratón
Parte humana
Parte de ratón
ADN híbrido de humano-ratón

FIG. 115

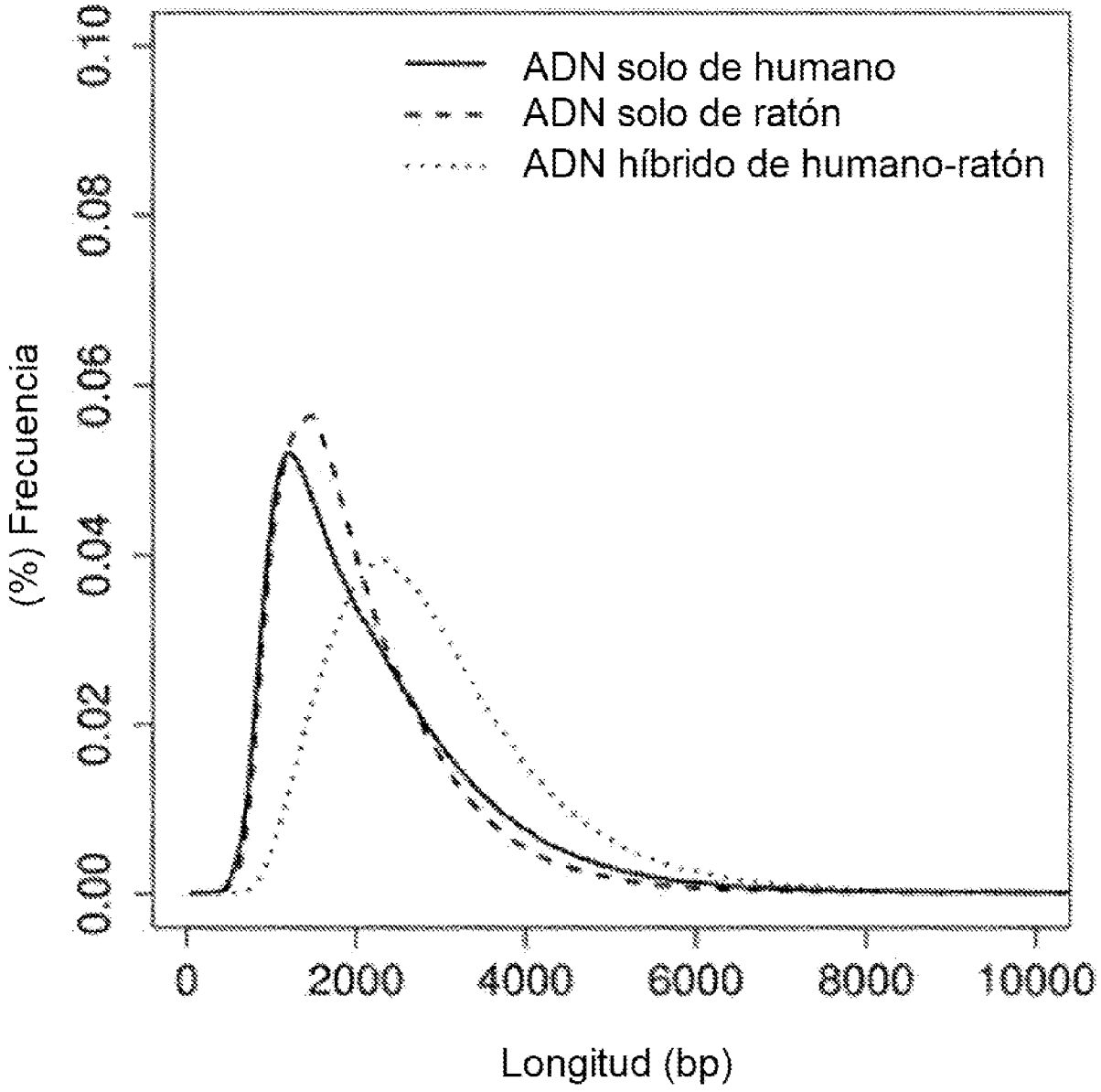

ADN solo de humano
ADN solo de ratón
ADN híbrido de humano-ratón
(%) Frecuencia
0.00
0.02
0.04
0.06
0.08
0.10
0
2000
4000
6000
8000
10000
Longitud (bp)

FIG. 116

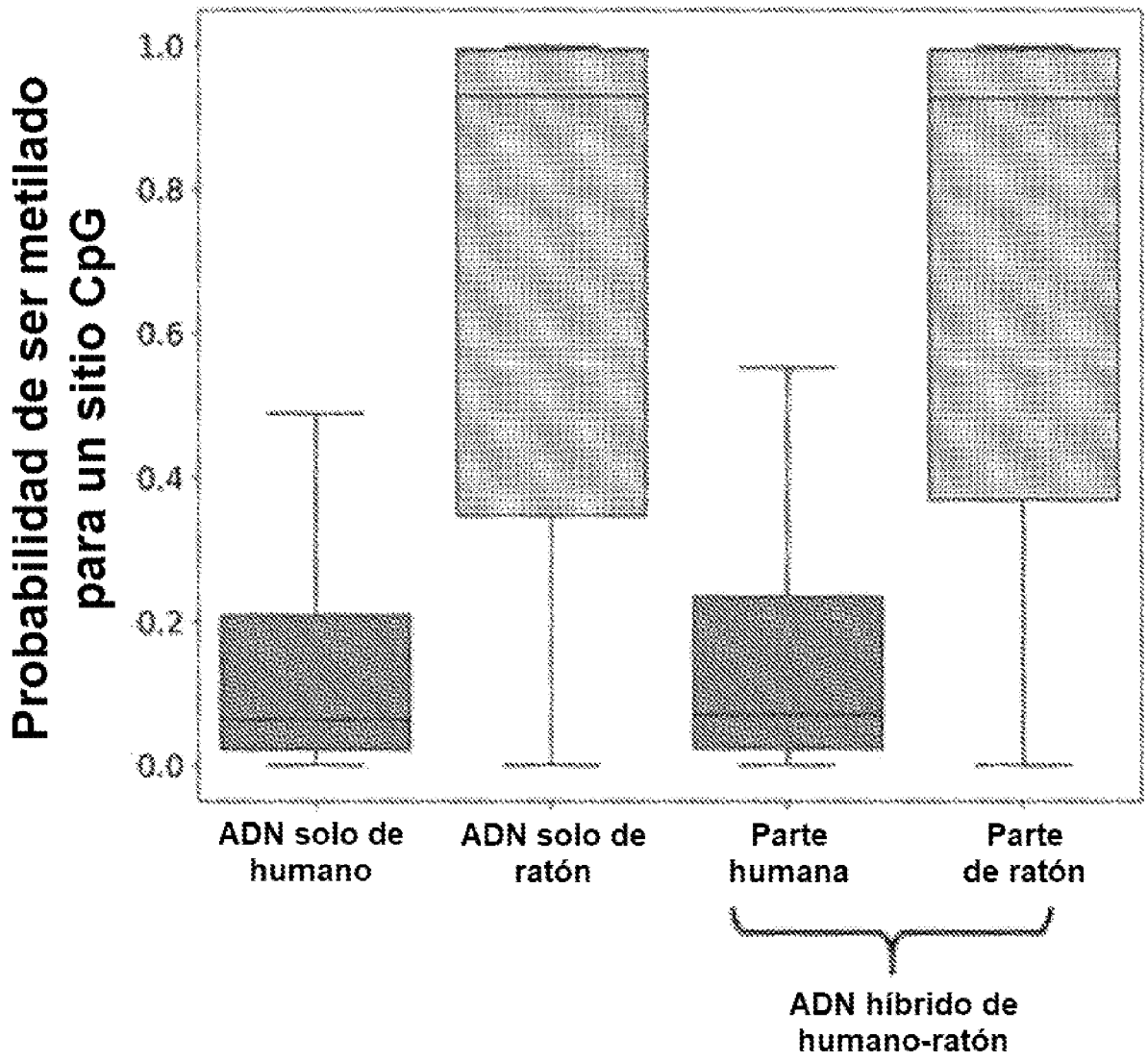

Probabilidad de ser metilado para un sitio CpG
1.0
0.8
0.6
0.4
0.2
0.0
ADN solo de humano
ADN solo de ratón
Parte humana
Parte de ratón
ADN híbrido de humano-ratón

FIG. 117

		Secuenciamiento de bisulfito		Secuenciamiento de PacBio	
		No. de sitios CG	Densidad de metilación (%)	No. de sitios CG	Densidad de metilación (%)
1) Solo de humano		2.330.407	41.4	16.226.014	56.0
2) Solo de ratón		2.726.499	1.6	9.388.340	10.7
3) ADN híbrido de humano-ratón	Parte humana	73.780	46.8	4.838.454	57.4
	Parte de ratón	76.312	2.3	4.385.048	12.1

FIG. 118

		Secuenciamiento de bisulfito		Secuenciamiento de PacBio	
		No. de sitios CG	Densidad de metilación (%)	No. de sitios CG	Densidad de metilación (%)
1) Solo de humano		2.938.088	1.6	14.503.548	11.6
2) Solo de ratón		1.513.671	62.4	11.348.555	71.5
3) ADN híbrido de humano-ratón	Parte humana	67.371	1.8	5.824.379	13.1
	Parte de ratón	58.242	67.4	5.093.097	72.2

FIG. 119

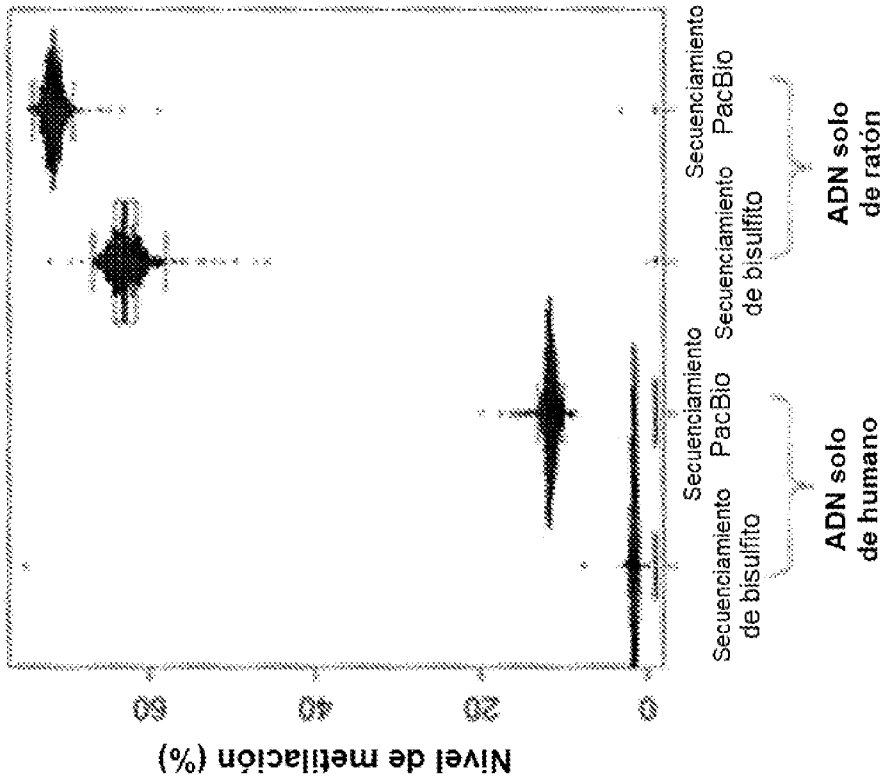


FIG. 120B

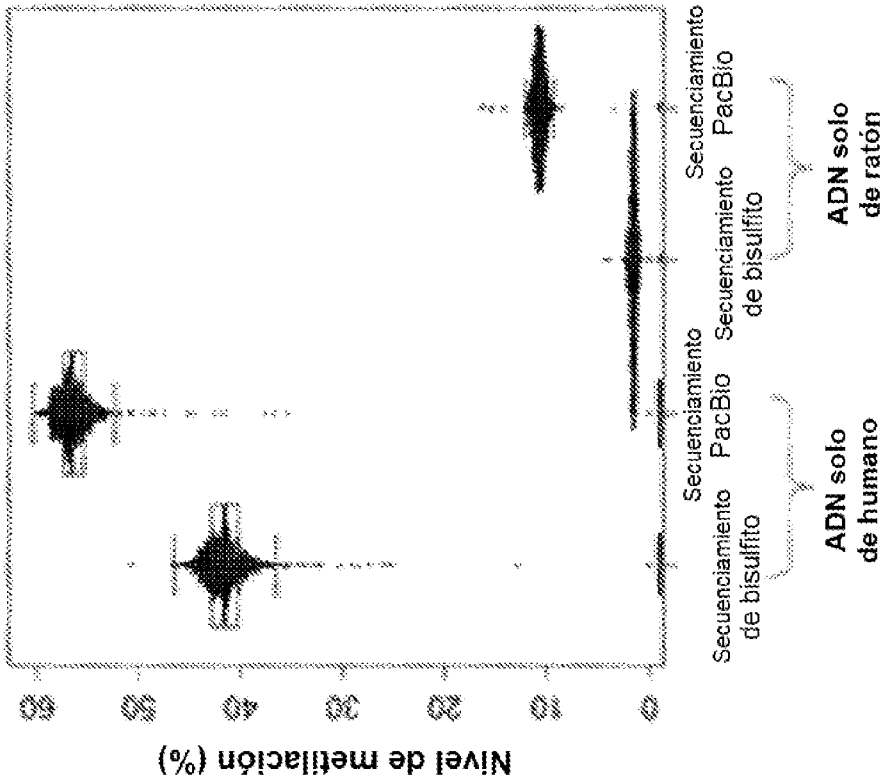


FIG. 120A

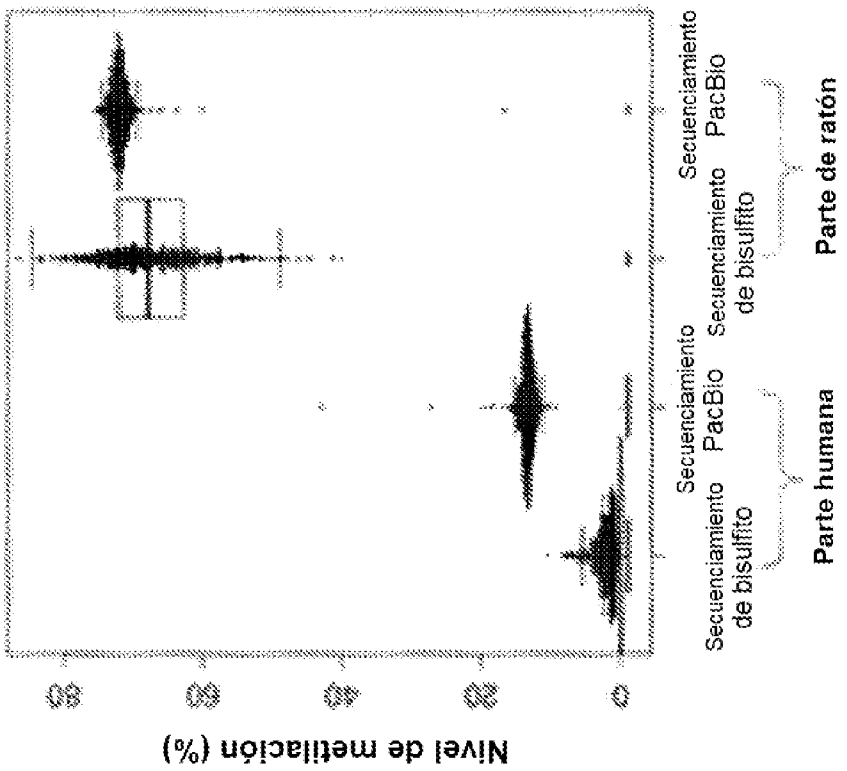


FIG. 121B

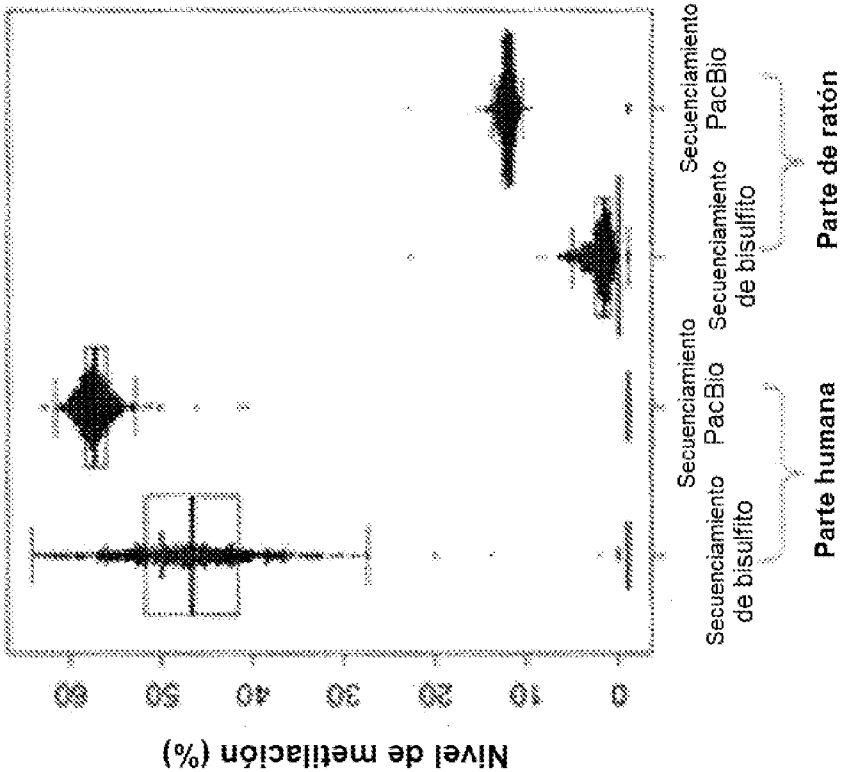


FIG. 121A

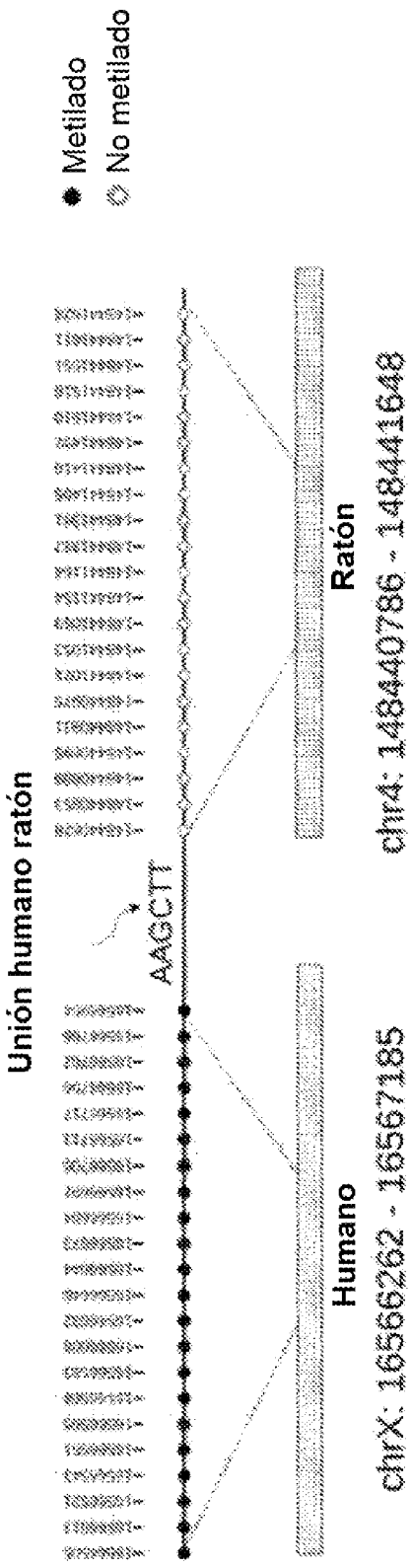


FIG. 122A

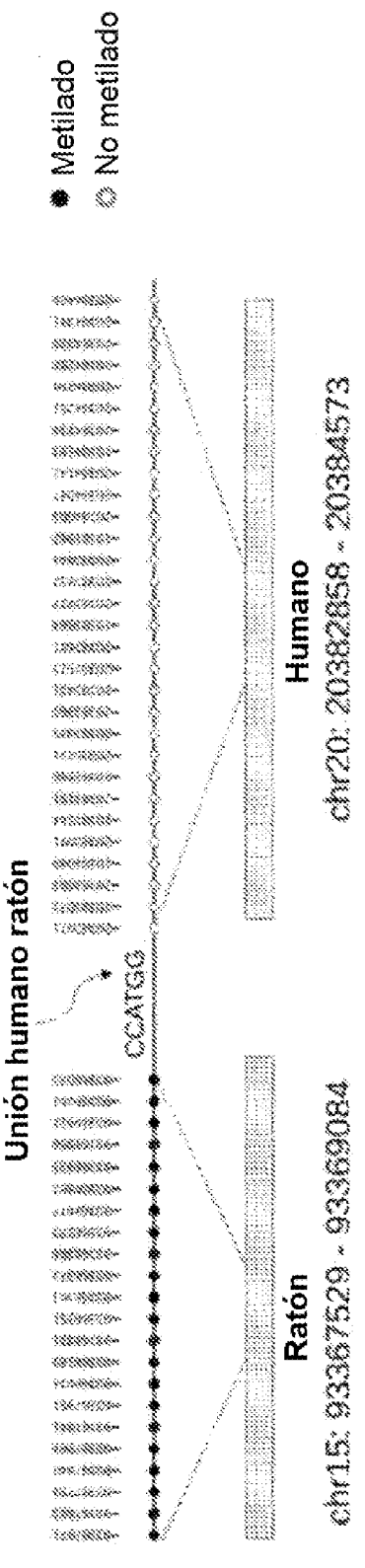


FIG. 122B

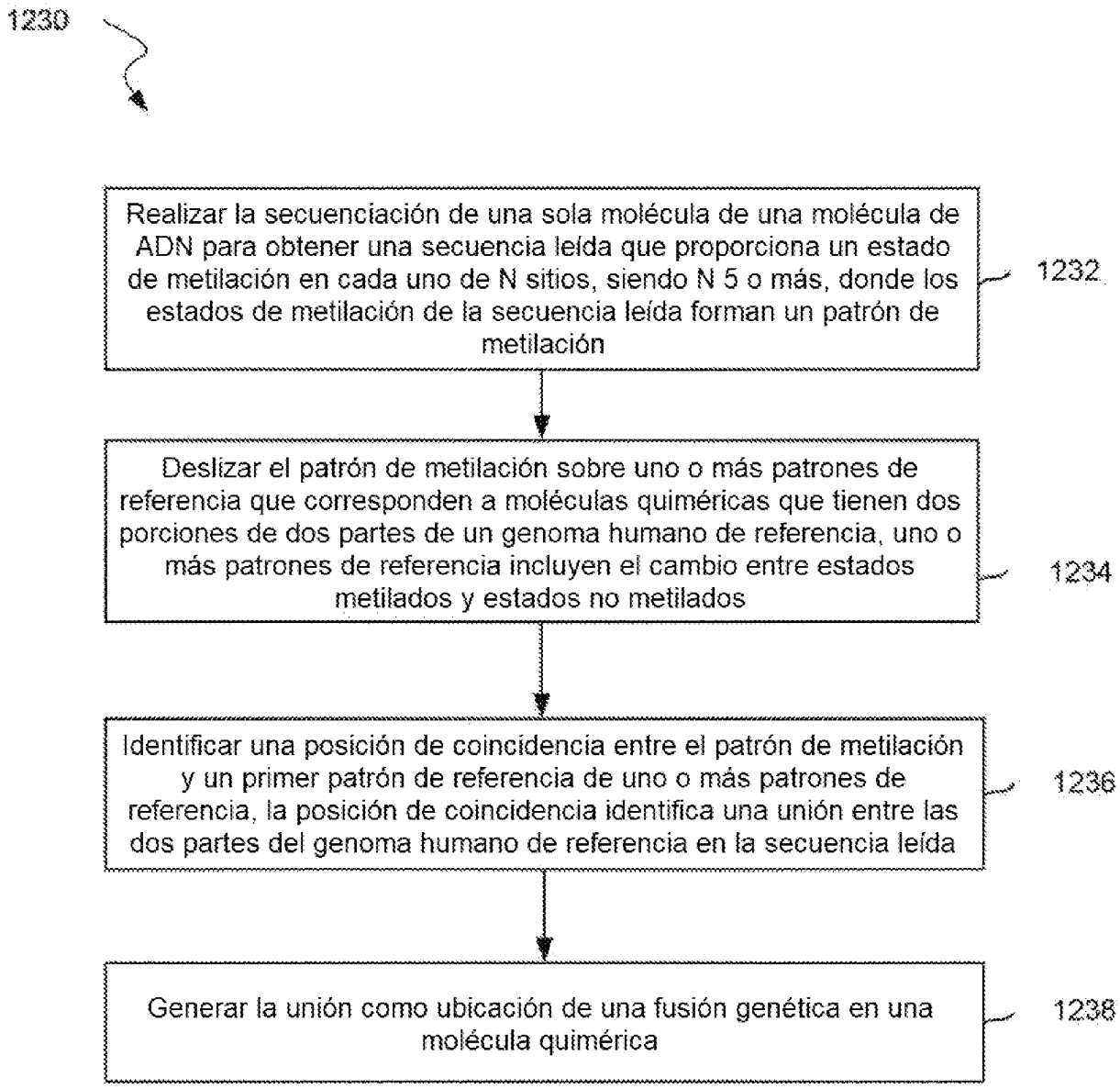

1230
Realizar la secuenciación de una sola molécula de una molécula de ADN para obtener una secuencia leída que proporciona un estado de metilación en cada uno de N sitios, siendo N 5 o más, donde los estados de metilación de la secuencia leída forman un patrón de metilación
1232
Deslizar el patrón de metilación sobre uno o más patrones de referencia que corresponden a moléculas quiméricas que tienen dos porciones de dos partes de un genoma humano de referencia, uno o más patrones de referencia incluyen el cambio entre estados metilados y estados no metilados
1234
Identificar una posición de coincidencia entre el patrón de metilación y un primer patrón de referencia de uno o más patrones de referencia, la posición de coincidencia identifica una unión entre las dos partes del genoma humano de referencia en la secuencia leída
1236
Generar la unión como ubicación de una fusión genética en una molécula quimérica
1238

FIG. 123

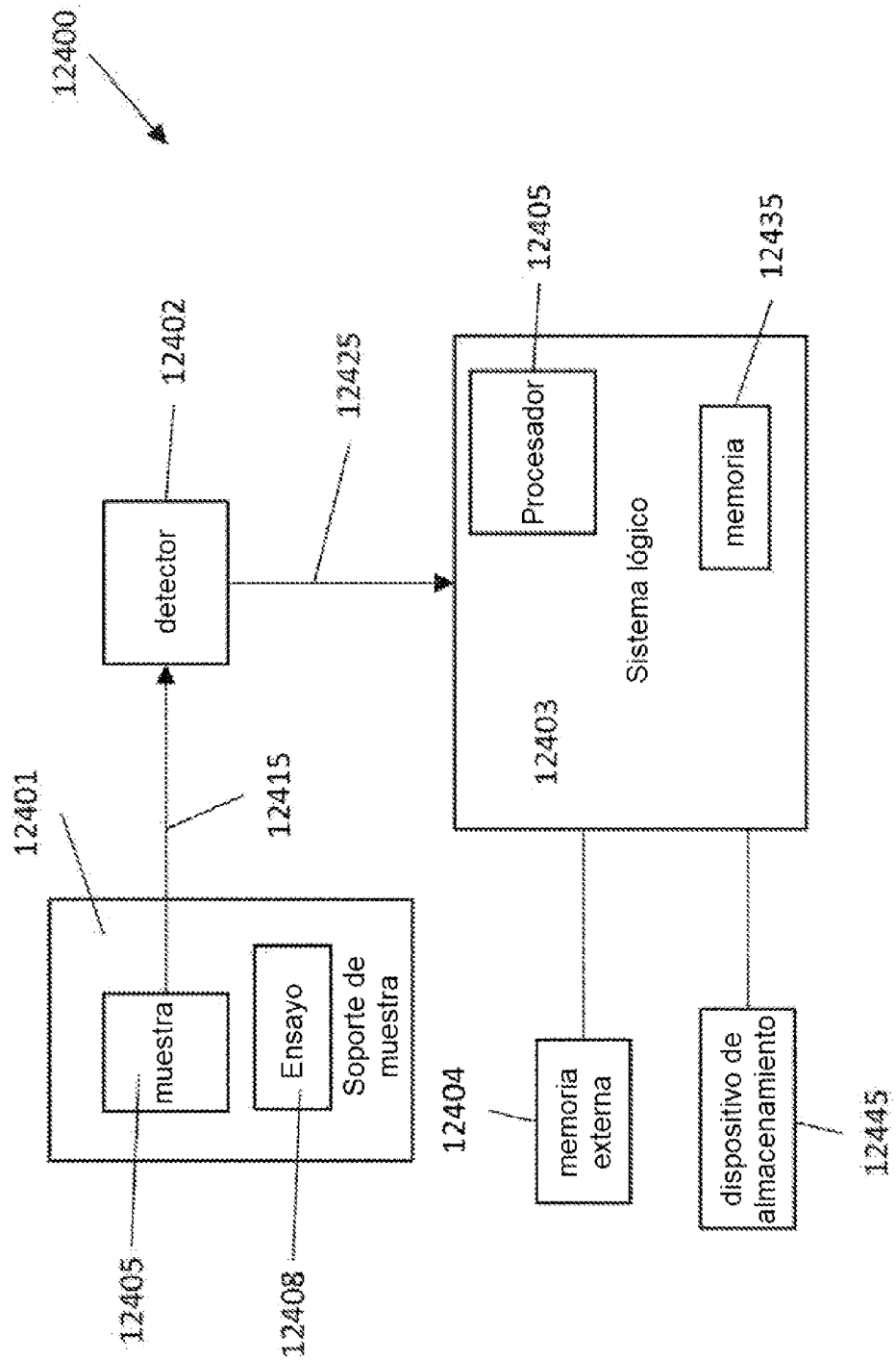

12400
12402
12425
12405
12435
detector
Procesador
memoria
Sistema lógico
12403
12401
12415
muestra
Ensayo
Soporte de muestra
12404
memoria externa
dispositivo de almacenamiento
12445
12405
12408

FIG. 124

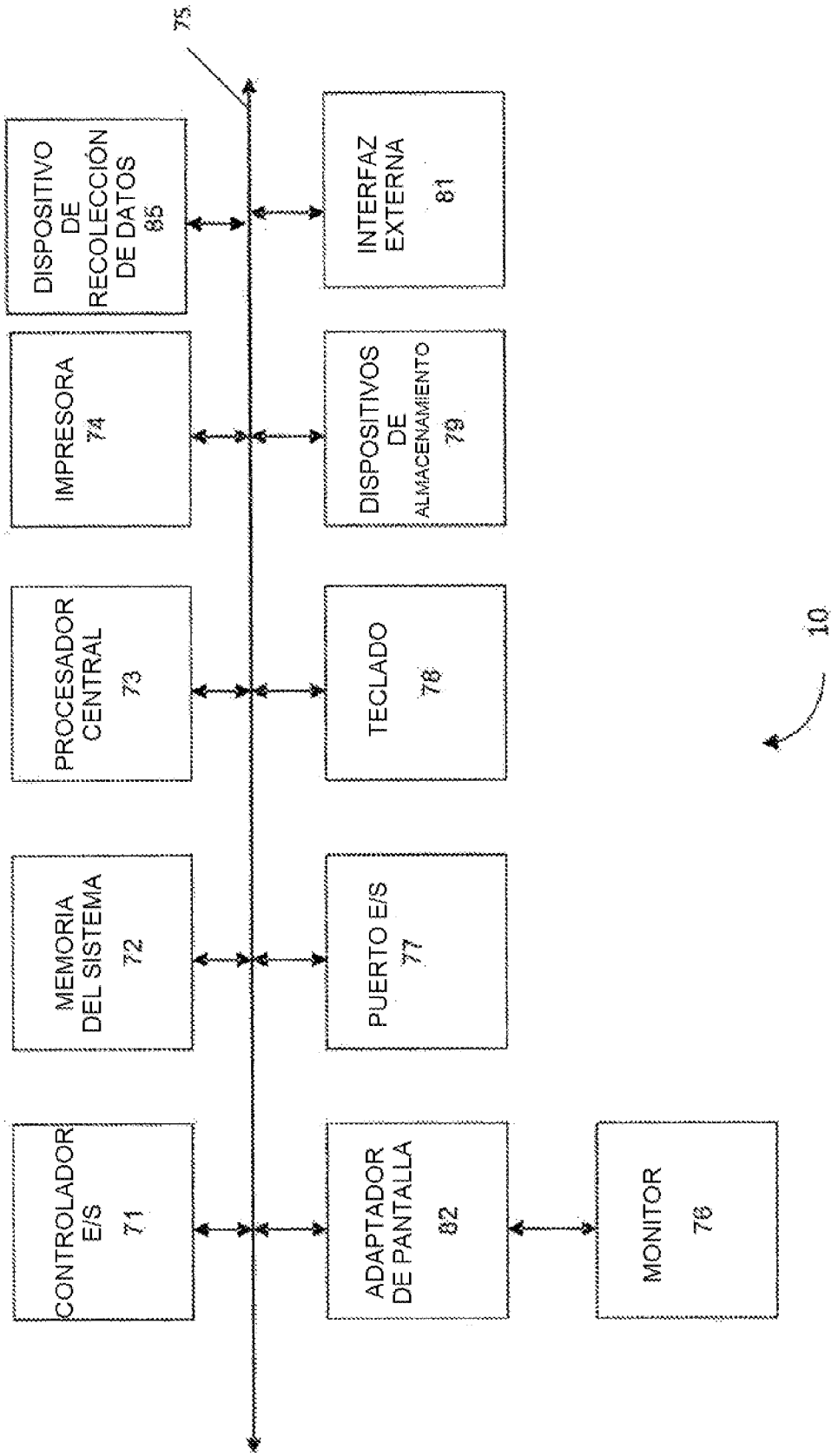

75
DISPOSITIVO DE RECOLECCIÓN DE DATOS 85
INTERFAZ EXTERNA 81
IMPRESORA 74
DISPOSITIVOS DE ALMACENAMIENTO 79
PROCESADOR CENTRAL 73
TECLADO 78
MEMORIA DEL SISTEMA 72
PUERTO E/S 77
CONTROLADOR E/S 71
ADAPTADOR DE PANTALLA 82
MONITOR 76
10

FIG. 125

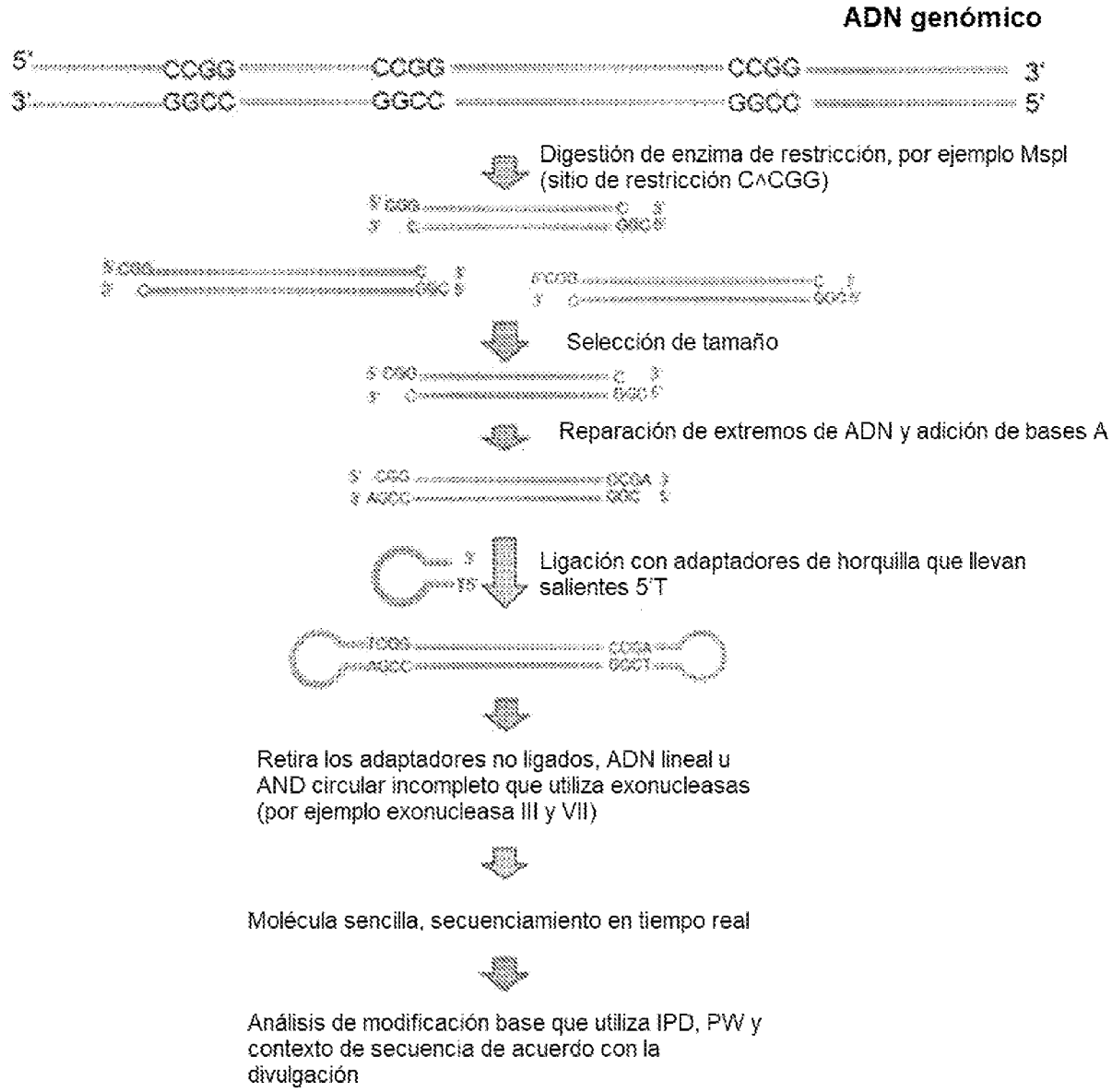

ADN genómico
5' CCGG CCGG CCGG 3'
3' GGCC GGCC GGCC 5'
Digestión de enzima de restricción, por ejemplo MspI (sitio de restricción C^CGG)
Selección de tamaño
Reparación de extremos de ADN y adición de bases A
Ligación con adaptadores de horquilla que llevan salientes 5'T
Retira los adaptadores no ligados, ADN lineal u AND circular incompleto que utiliza exonucleasas (por ejemplo exonucleasa III y VII)
Molécula sencilla, secuenciamiento en tiempo real
Análisis de modificación base que utiliza IPD, PW y contexto de secuencia de acuerdo con la divulgación

FIG. 126

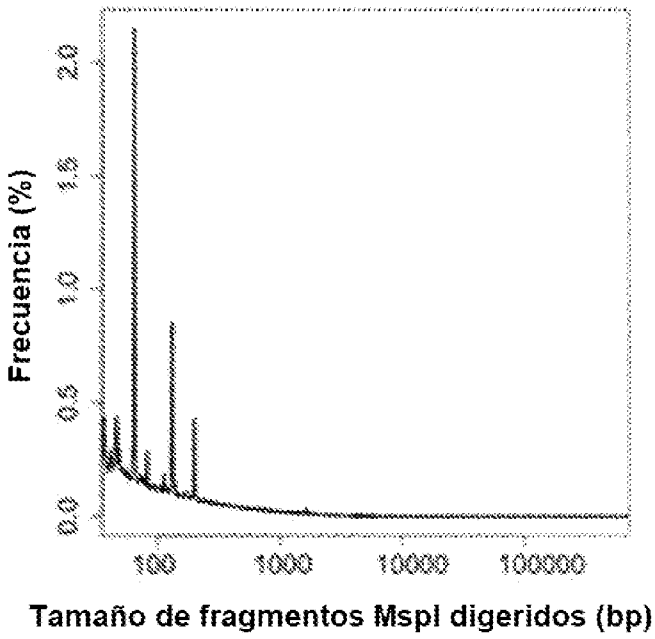


FIG. 127A

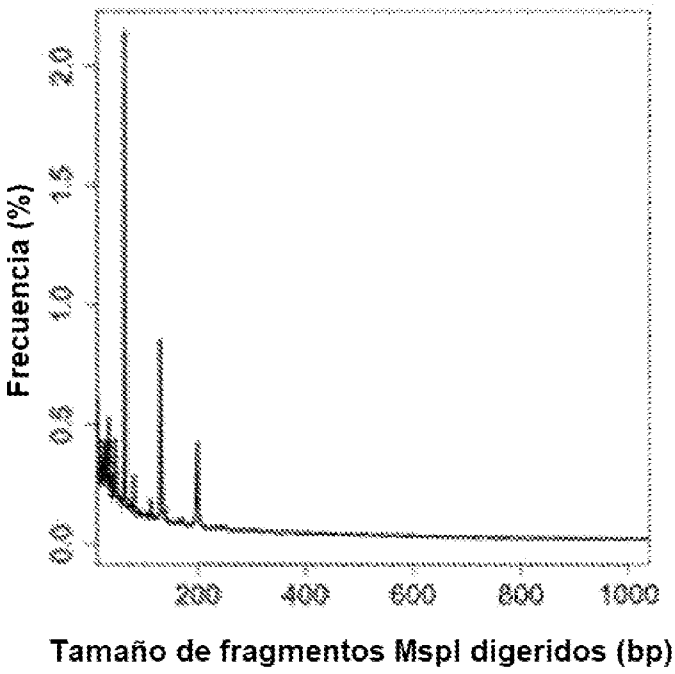


FIG. 127B

Rangos de tamaño (bp)	No. de moléculas	Porcentaje de moléculas dentro de un rango de tamaño relativo a los fragmentos totales (%)	No. de moléculas dentro de un rango de tamaño de se sobrepone a islas CpG	Porcentaje de moléculas dentro de un rango de tamaño que se sobrepone a islas CpG (%)	No. de sitios CpG que están secuenciados	No. de sitios CpG que caen dentro de islas CpG	Porcentaje de sitios CpG dirigidos por selección de tamaño y que caen dentro de las islas CpG (%)
93-200	526,543	23.03	104,059	19.76	2,358,020	885,041	37.53
200-400	269,562	11.79	23,927	8.88	1,781,568	353,097	19.82
400-600	177,776	7.77	7,369	4.15	1,468,561	107,130	7.29
600-800	133,927	5.86	3,673	2.74	1,326,844	46,851	3.53
800-1000	104,976	4.59	2,168	2.07	1,193,233	25,821	2.16
1000-2000	311,596	13.63	4,536	1.47	4,610,504	58,288	1.26
2000-3000	149,468	6.54	1,771	1.18	3,936,961	25,106	0.63
3000-4000	86,760	3.79	809	0.93	2,169,171	10,785	0.50
5000-6000	36,831	1.62	266	0.72	1,242,712	3,412	0.27
6000-7000	25,027	1.09	202	0.81	947,874	3,334	0.35
7000-8000	17,597	0.77	86	0.49	736,830	791	0.11
8000-9000	12,668	0.55	76	0.60	583,680	993	0.17
9000-10000	9,194	0.40	48	0.52	461,935	591	0.13
10000-15000	20,780	0.91	97	0.47	1,255,731	2,003	0.16
15000-20000	5,111	0.22	15	0.31	414,400	163	0.04
20000-25000	1,441	0.06	6	0.42	147,731	34	0.02

FIG. 128

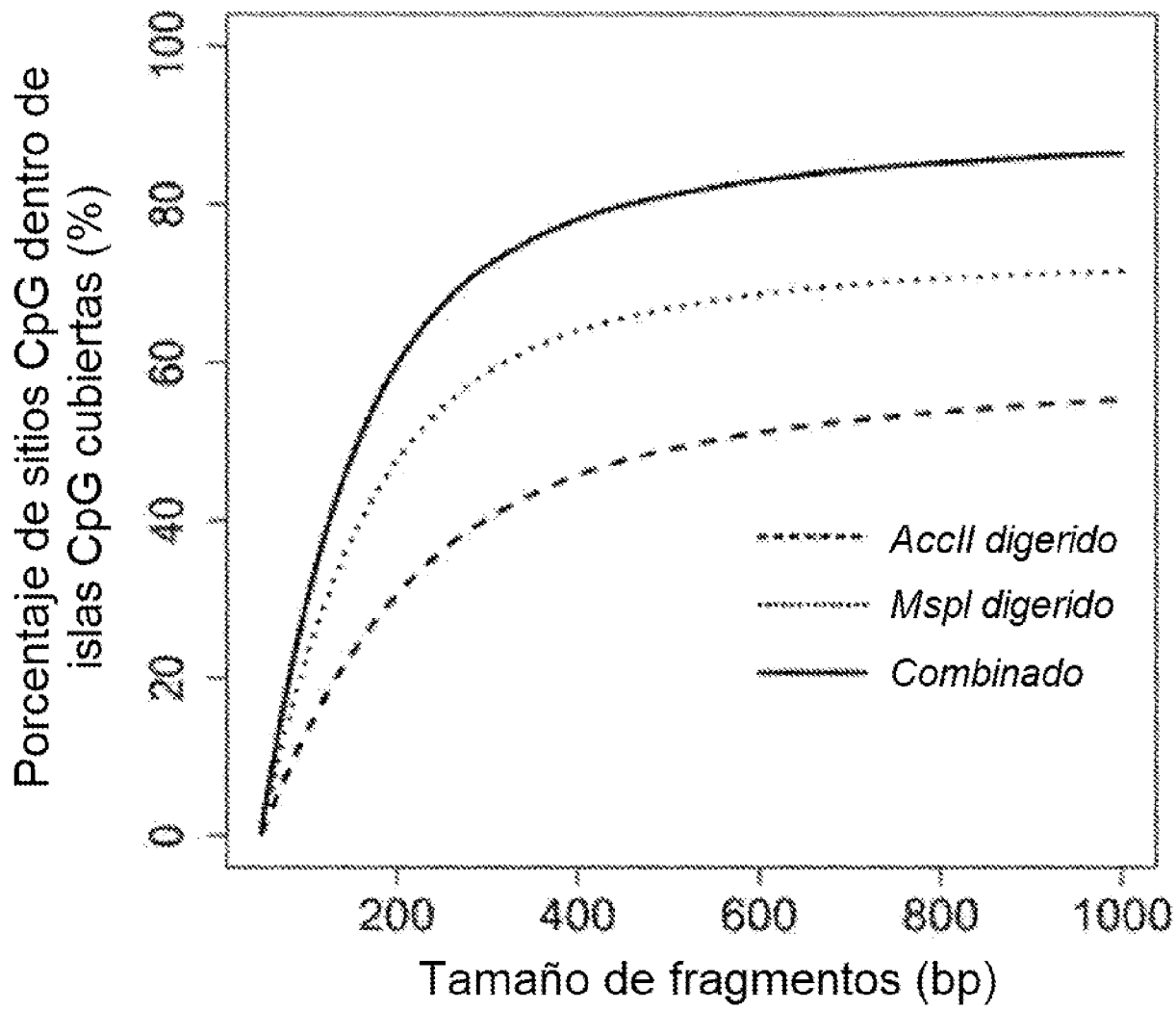

Porcentaje de sitios CpG dentro de islas CpG cubiertas (%)
0
20
40
60
80
100
200
400
600
800
1000
Tamaño de fragmentos (bp)
AccII digerido
MspI digerido
Combinado

FIG. 129

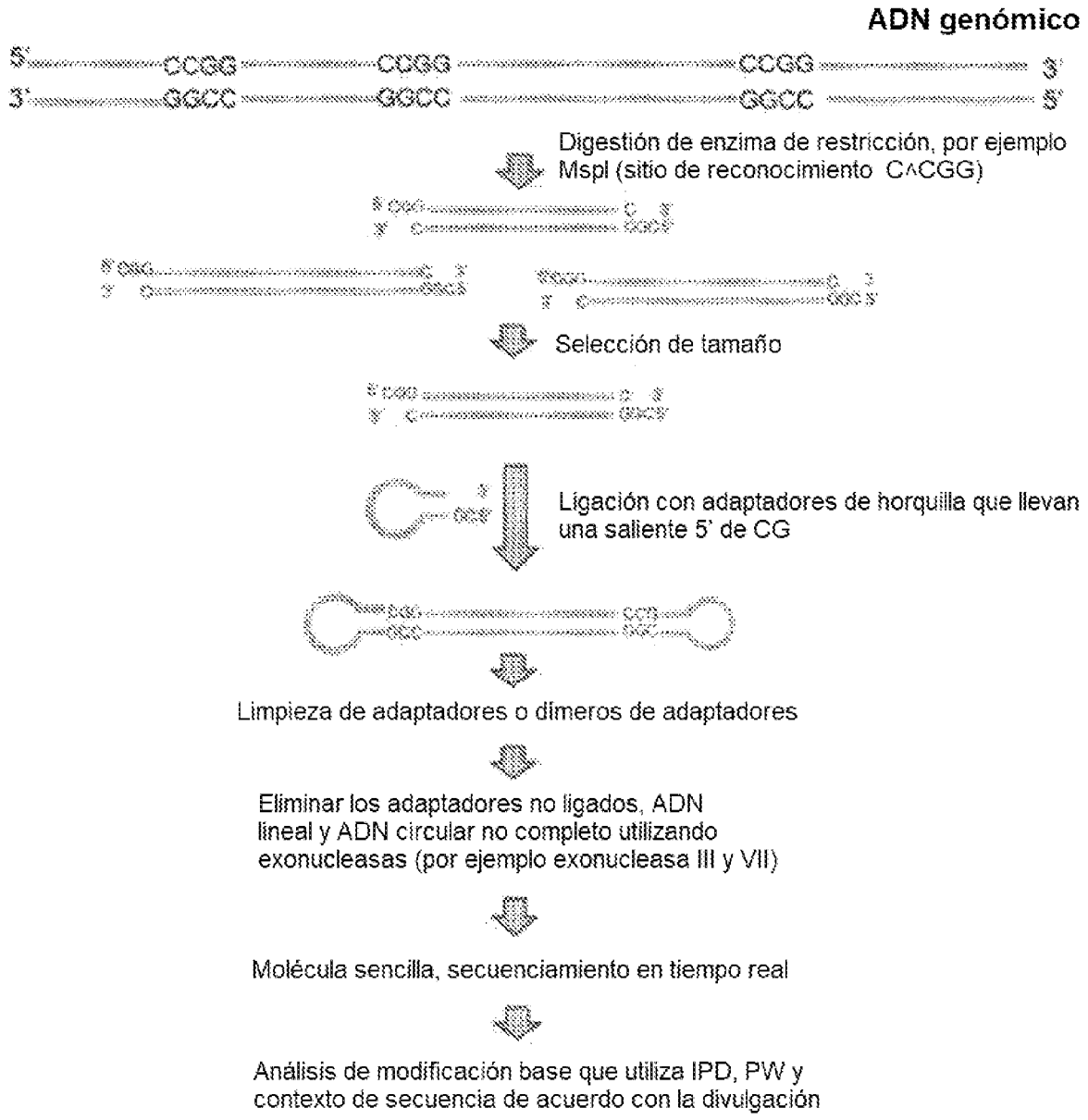

ADN genómico
5' CCGG CCGG CCGG 3'
3' GGCC GGCC GGCC 5'
Digestión de enzima de restricción, por ejemplo MspI (sitio de reconocimiento C^CGG)
Selección de tamaño
Ligación con adaptadores de horquilla que llevan una saliente 5' de CG
Limpieza de adaptadores o dímeros de adaptadores
Eliminar los adaptadores no ligados, ADN lineal y ADN circular no completo utilizando exonucleasas (por ejemplo exonucleasa III y VII)
Molécula sencilla, secuenciamiento en tiempo real
Análisis de modificación base que utiliza IPD, PW y contexto de secuencia de acuerdo con la divulgación

FIG. 130

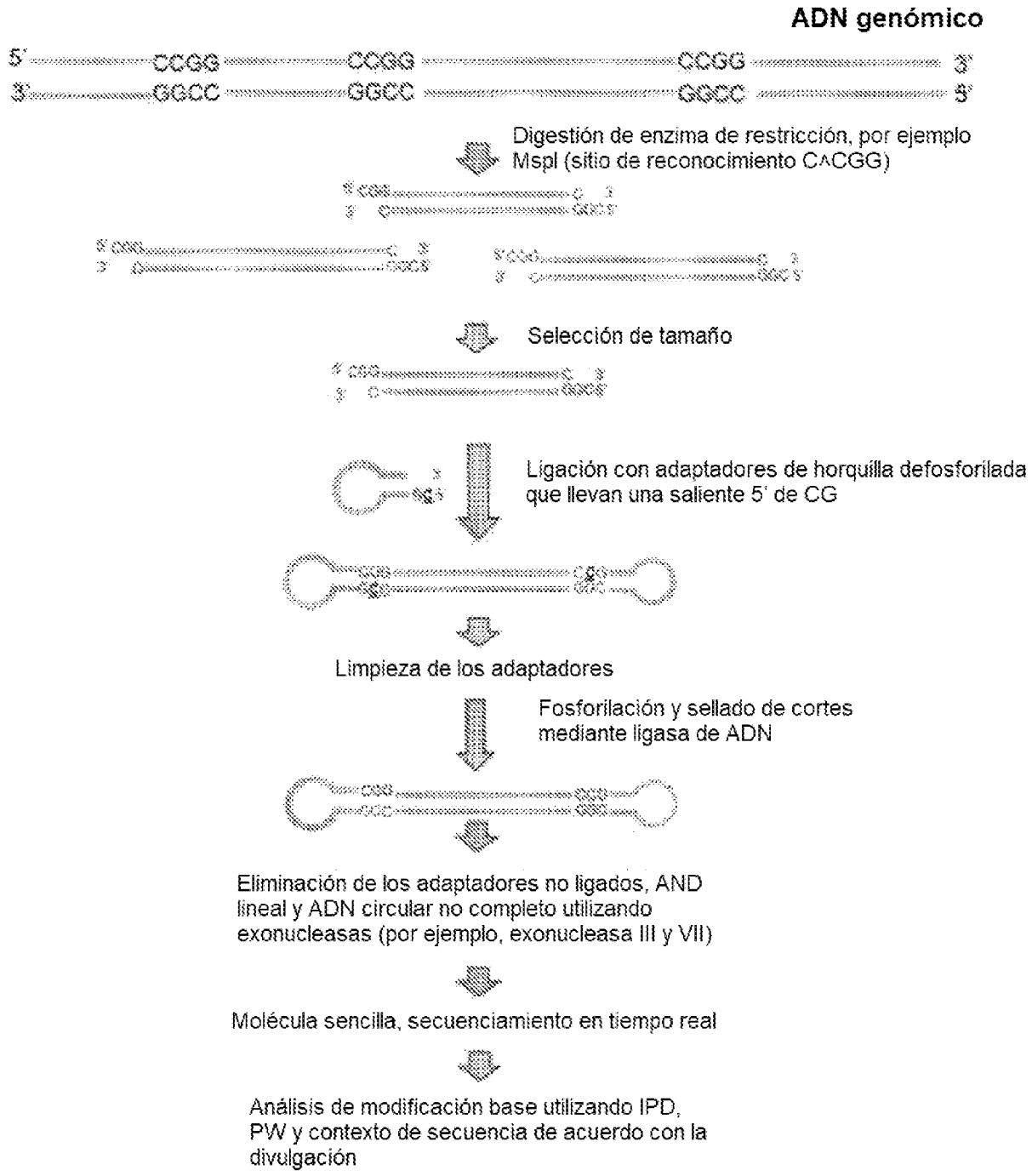

ADN genómico
5' CCGG CCGG CCGG 3'
3' GGCC GGCC GGCC 5'
Digestión de enzima de restricción, por ejemplo MspI (sitio de reconocimiento C^CGG)
Selección de tamaño
Ligación con adaptadores de horquilla defosforilada que llevan una saliente 5' de CG
Limpieza de los adaptadores
Fosforilación y sellado de cortes mediante ligasa de ADN
Eliminación de los adaptadores no ligados, AND lineal y ADN circular no completo utilizando exonucleasas (por ejemplo, exonucleasa III y VII)
Molécula sencilla, secuenciamiento en tiempo real
Análisis de modificación base utilizando IPD, PW y contexto de secuencia de acuerdo con la divulgación

FIG. 131

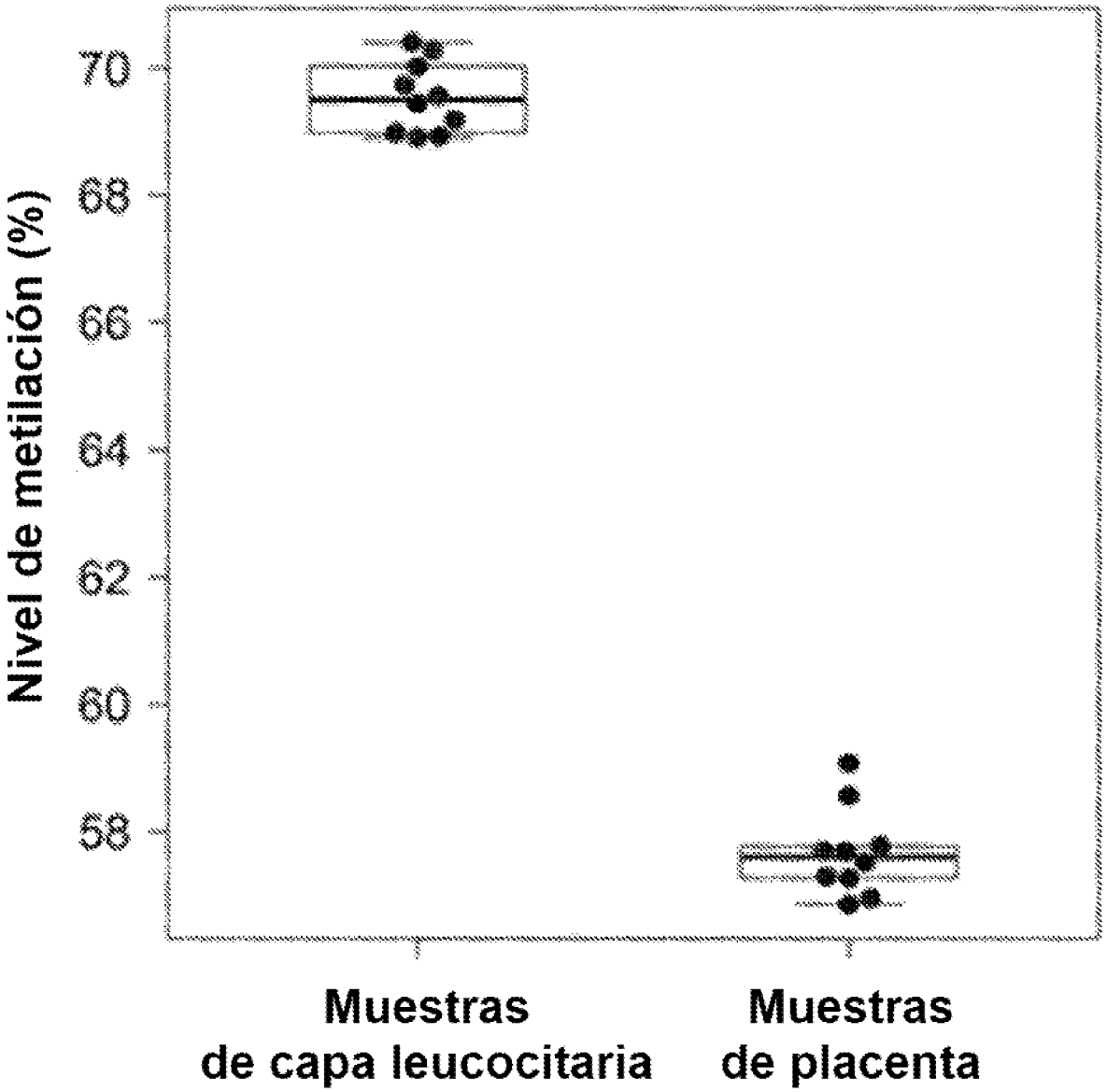

Nivel de metilación (%)
70
68
66
64
62
60
58
Muestras de capa leucocitaria
Muestras de placenta

FIG. 132

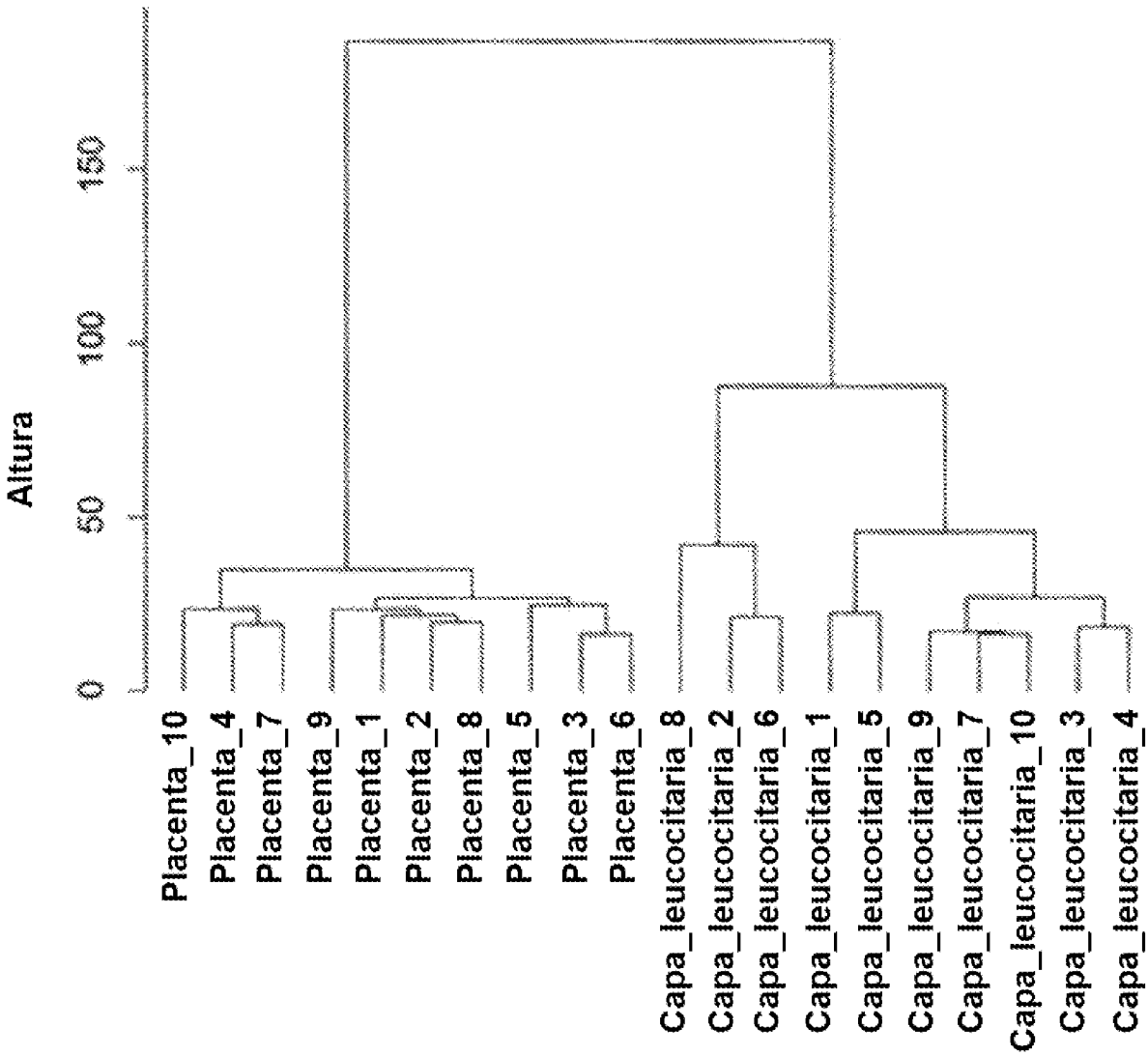

Altura
0
50
100
150
Placenta_10
Placenta_4
Placenta_7
Placenta_9
Placenta_1
Placenta_2
Placenta_8
Placenta_5
Placenta_3
Placenta_6
Capa_leucocitaria_8
Capa_leucocitaria_2
Capa_leucocitaria_6
Capa_leucocitaria_1
Capa_leucocitaria_5
Capa_leucocitaria_9
Capa_leucocitaria_7
Capa_leucocitaria_10
Capa_leucocitaria_3
Capa_leucocitaria_4

FIG. 133