(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 994 567**

(51) Int. Cl.:

C12Q 1/6886 (2008.01)
C12Q 1/6883 (2008.01)
G16B 20/30 (2009.01)
G16B 25/10 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **04.02.2020 PCT/US2020/016673**

(87) Fecha y número de publicación internacional: **13.08.2020 WO20163403**

(96) Fecha de presentación y número de la solicitud europea: **04.02.2020 E 20753046 (0)**

(97) Fecha y número de publicación de la concesión europea: **04.09.2024 EP 3921445**

(54) Título: **Detección de cáncer, tejido canceroso de origen y/o un tipo de célula cancerosa**

(30) Prioridad:

05.02.2019 US 201962801556 P
05.02.2019 US 201962801561 P
24.01.2020 US 202062965327 P
24.01.2020 US 202062965342 P
24.01.2020 WO PCT/US2020/015082

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
27.01.2025

(73) Titular/es:

GRAIL, INC. (100.00%)
1525 O'Brien Drive
Menlo Park, California 94025, US

(72) Inventor/es:

GROSS, SAMUEL S.;
VENN, OLIVER CLAUDE;
FIELDS, ALEXANDER P.;
LIU, QINWEN;
SCHELLENBERGER, JAN;
BREDNO, JOERG;
BEAUSANG, JOHN F.;
SHOJAEE, SEYEDMEHDI y
JAMSHIDI, ARASH

(74) Agente/Representante:

VALLEJO LÓPEZ, Juan Pedro

Observaciones:

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

DESCRIPCIÓN

Detección de cáncer, tejido canceroso de origen y/o un tipo de célula cancerosa

Antecedentes

Las células madre hematopoyéticas (HSC, por sus siglas en inglés) y las células progenitoras hematopoyéticas (HPC, por sus siglas en inglés) se dividen para producir células sanguíneas mediante un proceso de regeneración continuo. A medida que las células se dividen, son propensas a acumular mutaciones que, por lo general, no afectan a la función. Se ha determinado que alrededor del 3 al 5 % de las personas normales mayores de 50 años y aproximadamente 10 % de las personas de 70 a 80 años tienen hematopoyesis clonal de potencial indeterminado (CHIP, por sus siglas en inglés) definida por la presencia de mutaciones de bajo nivel en la sangre periférica en personas clínicamente normales.

Algunas mutaciones confieren ventajas en la autorrenovación, la proliferación o ambas, lo que resulta en la expansión clonal de las células que comprenden las mutaciones en cuestión. Aunque estas mutaciones no son necesariamente indicativas de una enfermedad hematológica, la acumulación de mutaciones durante la expansión clonal puede, eventualmente, conducir a un estado patológico (p. ej., cáncer). La hematopoyesis clonal se ha relacionado con un riesgo más de 10 veces mayor de desarrollar un cáncer de sangre. Por lo tanto, la detección de la hematopoyesis clonal puede permitir una detección temprana del cáncer, lo que a su vez permite un tratamiento más temprano y, por lo tanto, una mayor probabilidad de supervivencia. La diferenciación del CHIP de otros trastornos hematológicos, como la leucemia, el mieloma múltiple y el linfoma, permite además los tratamientos y las actividades profilácticas adecuados.

Estudios de secuenciación recientes han identificado un conjunto de mutaciones recurrentes en varios tipos de neoplasias malignas hematológicas (véase, p. ej., Mardis E R y col. The New England Journal of Medicine 2009; Bejar R y col. The New England Journal of Medicine 2011; Papaemmanuil E y col. The New England Journal of Medicine 2011; y Walter y col. Leukemia 2011). Sin embargo, se desconoce la frecuencia de estas mutaciones somáticas en la población general.

Por consiguiente, aún no se dispone de un método rentable para detectar con precisión diversos trastornos hematológicos mediante la detección de regiones metiladas diferencialmente.

Liu, L. y col. "Targeted methylation sequencing of plasma cell-free DNA for cancer detection and classification", ANNALS OF ONCOLOGY, vol. 29, n.º 6, 1 de junio de 2018 (2018-06-01), analizan el desarrollo de un ensayo de secuenciación de metilación dirigido a 9223 sitios CpG hipermetilados de forma consistente según The Cancer Genome Atlas. El estudio llevó a cabo una validación clínica del método utilizando muestras de ADNlc plasmático de 78 pacientes con cáncer colorrectal avanzado, cáncer de pulmón no microcítico (CPNM), cáncer de mama o melanoma, y comparó los resultados con los desenlaces clínicos de los pacientes.

Pidsley, R. y col. "Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling", *Genome Biology* (2016) 17:208 analiza usando la plataforma Illumina HumanMethylation450 (HM450) BeadChip con secuenciación de bisulfito de genoma completo (WGBS) realizar una evaluación crítica de la plataforma de micromatriz Illumina MethylationEPIC BeadChip.

Resumen

El alcance de la presente invención se expone en las reivindicaciones adjuntas, donde se describe:

Un método para detectar un trastorno hematológico (HD) en un sujeto, el método comprende:

(a) obtener lecturas de secuenciación del ADN libre de células (ADNlc) de una muestra de un sujeto, o productos de amplificación de los mismos, enriquecidos por contacto con una composición que comprende una pluralidad de diferentes oligonucleótidos cebo, en donde el ADNlc se convierte antes de la secuenciación mediante el tratamiento del ADNlc para convertir las citosinas no metiladas en uracilos, en donde:

i) cada oligonucleótido cebo en la pluralidad de oligonucleótidos cebo diferentes tiene al menos 45 nucleótidos de longitud, opcionalmente 45 a 300 nucleótidos de longitud;

ii) la pluralidad de oligonucleótidos cebo diferentes se hibrida colectivamente con al menos 100 regiones genómicas diana;

iii) las al menos 100 regiones genómicas diana están metiladas diferencialmente en al menos un HD en relación con un HD diferente;

iv) el al menos un HD y el HD diferente se seleccionan entre hematopoyesis clonal de potencial indeterminado (CHIP), leucemia, neoplasias linfoides, mieloma múltiple y una neoplasia mieloide; y

(b) usar un ordenador para aplicar un clasificador entrenado a las lecturas de secuenciación para determinar una presencia o ausencia del al menos un HD, en donde el clasificador se entrena usando secuencias de ADN convertido, en donde el ADN convertido se refiere al ADN que se ha tratado para convertir las citosinas no metiladas en uracilos, en donde:

i) el clasificador entrenado distingue entre el CHIP y uno o más HD seleccionados entre leucemia, neoplasias linfoides, mieloma múltiple y una neoplasia mieloide basándose en lecturas de secuenciación; y

ii) la detección de una serie de lecturas de secuenciación para una pluralidad de las al menos 100 regiones genómicas diana por encima de un umbral identifica la presencia del al menos un HD.

En la presente memoria también se proporcionan (pero no se reivindican) composiciones que comprenden una pluralidad de diferentes oligonucleótidos cebo, en donde la pluralidad de diferentes oligonucleótidos cebo se configuran para hibridarse colectivamente con moléculas de ADN derivadas de al menos 100 regiones genómicas diana, en donde cada región genómica de las al menos 100 regiones genómicas diana está metilada diferencialmente en al menos un primer trastorno hematológico o cáncer hematológico en relación con otro tipo de trastorno hematológico o cáncer no hematológico, en donde el primer trastorno hematológico y el otro trastorno hematológico se selecciona entre leucemia, neoplasias linfoides (p. ej., linfoma), mieloma múltiple y una neoplasia mieloide.

En algunas disposiciones, la pluralidad de oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 %, al menos 25 %, o al menos 50 % de las regiones genómicas diana de una cualquiera de las Listas 1-8. En algunas disposiciones, la pluralidad de oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos 20 %, al menos 25 % o al menos 50 % de las regiones genómicas diana de las listas 1-8. En algunas disposiciones, la pluralidad de oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 % de las regiones genómicas diana de las listas 1o 8. En algunas disposiciones, las moléculas de ADN se derivan de al menos 25 %, 30 %, 40 %, 50 %, 60 %, 70 %, o 80 % de las regiones genómicas diana de las listas 1 o 8. En algunas disposiciones, la pluralidad de oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 % de las regiones genómicas diana de una cualquiera de las listas 2-4. En algunas disposiciones, la pluralidad de oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 % de las regiones genómicas diana de las listas 2-4. En algunas disposiciones, la pluralidad de oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 25 %, 30 %, 40 %, 50 %, 60 %, 70 %, o 80 % de las regiones genómicas diana de las listas 2-4. En algunas disposiciones, la pluralidad de oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 % de las regiones genómicas diana de una cualquiera de las listas 5-7. En algunas disposiciones, la pluralidad de oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 % de las regiones genómicas diana de las listas 5-7. En algunas disposiciones, la pluralidad de oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 25 %, 30 %, 40 %, 50 %, 60 %, 70 %, o 80 % de las regiones genómicas diana de las listas 5-7. En algunas disposiciones, el primer trastorno hematológico y el otro trastorno hematológico se seleccionan entre una neoplasia linfoide, un mieloma múltiple y una neoplasia mieloide.

En la presente memoria también se proporcionan (pero no se reivindican) composiciones que comprenden una pluralidad de oligonucleótidos cebo diferentes configurados para hibridarse con moléculas de ADN derivadas de al menos el 20 % de las regiones genómicas diana de una cualquiera de las listas 1-7. En algunas disposiciones, la pluralidad de oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 % de las regiones genómicas diana de las listas 1o 8. En algunas disposiciones, las moléculas de ADN se derivan de al menos 25 %, 30 %, 40 %, 50 %, 60 %, 70 %, o 80 % de las regiones genómicas diana de las listas 1 o 8. En algunas disposiciones, la pluralidad de oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 % de las regiones genómicas diana de una cualquiera de las listas 2-4. En algunas disposiciones, la pluralidad de oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 % de las regiones genómicas diana de las listas 2-4. En algunas disposiciones, la pluralidad de oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 % de las regiones genómicas diana de una cualquiera de las listas 5-7. En algunas disposiciones, la pluralidad de oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 % de las regiones genómicas diana de las listas 5-7.

En la presente memoria también se proporcionan composiciones proporcionadas anteriormente, en donde la pluralidad de diferentes oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 % de las regiones genómicas diana de la lista 2. En algunas disposiciones, las moléculas de ADN se derivan de al menos el 25 %, 30 %, 40 %, 50 %, 60 %, 70 %, o 80 % de las regiones genómicas diana de la lista 2.

En la presente memoria también se proporcionan composiciones proporcionadas anteriormente, en donde la pluralidad de diferentes oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 %

de las regiones genómicas diana de la lista 3. En algunas disposiciones, las moléculas de ADN se derivan de al menos el 25 %, 30 %, 40 %, 50 %, 60 %, 70 %, o 80 % de las regiones genómicas diana de la lista 3.

En la presente memoria también se proporcionan composiciones proporcionadas anteriormente, en donde la pluralidad de diferentes oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 % de las regiones genómicas diana de la lista 4. En algunas disposiciones, las moléculas de ADN se derivan de al menos el 25 %, 30 %, 40 %, 50 %, 60 %, 70 %, o 80 % de las regiones genómicas diana de la lista 4.

En la presente memoria también se proporcionan composiciones proporcionadas anteriormente, en donde la pluralidad de diferentes oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 % de las regiones genómicas diana de la lista 5. En algunas disposiciones, las moléculas de ADN se derivan de al menos el 25 %, 30 %, 40 %, 50 %, 60 %, 70 %, o 80 % de las regiones genómicas diana de la lista 5.

En la presente memoria también se proporcionan composiciones proporcionadas anteriormente, en donde la pluralidad de diferentes oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 % de las regiones genómicas diana de la lista 6. En algunas disposiciones, las moléculas de ADN se derivan de al menos el 25 %, 30 %, 40 %, 50 %, 60 %, 70 %, o 80 % de las regiones genómicas diana de la lista 6.

En la presente memoria también se proporcionan composiciones proporcionadas anteriormente, en donde la pluralidad de diferentes oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 % de las regiones genómicas diana de la lista 7. En algunas disposiciones, las moléculas de ADN se derivan de al menos el 25 %, 30 %, 40 %, 50 %, 60 %, 70 %, o 80 % de las regiones genómicas diana de la lista 7.

En la presente memoria también se proporcionan composiciones proporcionadas anteriormente, en donde la pluralidad de diferentes oligonucleótidos cebo se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 % de las regiones genómicas diana de la lista 8. En algunas disposiciones, las moléculas de ADN se derivan de al menos el 25 %, 30 %, 40 %, 50 %, 60 %, 70 %, o 80 % de las regiones genómicas diana de la lista 8.

En la presente memoria también se proporcionan las composiciones proporcionadas anteriormente, en donde el tamaño total de las regiones genómicas diana es menor que 2000 kb, menor que 1500 kb, menor que 1200 kb, menor que 1000 kb, menor que 500 kb o menor que 300 kb.

En la presente memoria también se proporcionan las composiciones proporcionadas anteriormente, en donde las moléculas de ADN son fragmentos de ADNlc convertidos. En algunas disposiciones, las regiones genómicas diana son regiones hipermetiladas, regiones hipometiladas o regiones binarias que pueden estar hipermetiladas o hipometiladas, como se indica en la lista de secuencias. En algunas disposiciones, los oligonucleótidos cebo están configurados para hibridarse con moléculas de ADN convertidas hipermetiladas, moléculas de ADN convertidas hipometiladas o moléculas de ADN convertidas tanto hipermetiladas como hipometiladas derivadas de cada región genómica diana, como se indica en la lista de secuencias.

En la presente memoria también se proporcionan las composiciones proporcionadas anteriormente, en donde cada uno de los oligonucleótidos cebo está conjugado con un resto de afinidad. En algunas disposiciones, el resto de afinidad es biotina. En algunas disposiciones, cada uno de los oligonucleótidos cebo se conjuga con una superficie sólida. En algunas disposiciones, la superficie sólida es una micromatriz o chip.

En la presente memoria también se proporcionan composiciones proporcionadas anteriormente, en donde cada uno de los oligonucleótidos cebo tiene una longitud de 45 a 300 bases de nucleótidos, 75-200 bases de nucleótidos, 100-150 bases de nucleótidos o aproximadamente 120 bases de nucleótidos.

En la presente memoria también se proporcionan composiciones proporcionadas anteriormente, en donde los oligonucleótidos cebo comprenden una pluralidad de conjuntos de dos o más oligonucleótidos cebo, en donde cada oligonucleótido cebo dentro de un conjunto de oligonucleótidos cebo está configurado para unirse a las moléculas de ADN convertidas de la misma región genómica diana. En algunas disposiciones, cada conjunto de oligonucleótidos cebo comprende 1 o más pares de un primer oligonucleótido cebo y un segundo oligonucleótido cebo, cada oligonucleótido cebo comprende un extremo 5' y un extremo 3', una secuencia de al menos X bases de nucleótidos en el extremo 3' del primer oligonucleótido cebo es idéntica a una secuencia de X bases de nucleótidos en el extremo 5' del segundo oligonucleótido cebo, y X es al menos 25, 30, 35, 40, 45, 50, 60, 70, 75 o 100. En algunas disposiciones, el primer oligonucleótido cebo comprende una secuencia de al menos 31, 40, 50 o 60 bases de nucleótidos que no se superpone a una secuencia del segundo oligonucleótido cebo.

En la presente memoria también se proporcionan composiciones proporcionadas anteriormente, en donde las al menos 100 regiones diana comprenden al menos 200, al menos 500, al menos 1000, al menos 1500, al menos 2000, al menos 3000, al menos 4000, al menos 5000, al menos 8000, al menos 10.000, al menos 15.000 o al menos 20.000 regiones genómicas regiones genómicas.

En la presente memoria también se proporcionan las composiciones proporcionadas anteriormente, que comprenden además el ADNlc convertido de un sujeto de prueba.

En la presente memoria también se proporcionan las composiciones proporcionadas anteriormente, en donde el ADNlc del sujeto de prueba se convierte mediante un proceso que comprende el tratamiento con bisulfito o una citosina desaminasa.

En la presente memoria también se proporcionan (pero no se reivindican de forma aislada) métodos para enriquecer fragmentos de ADNlc convertidos que informen sobre un tipo de trastorno hematológico, cuyo método comprende: poner en contacto la composición de oligonucleótidos cebo proporcionada anteriormente con ADN derivado de un sujeto de prueba y enriquecer la muestra para obtener el ADNlc correspondiente a las regiones genómicas asociadas con el tipo de cáncer mediante captura por hibridación.

En la presente memoria también se proporcionan (pero no se reivindican de forma aislada) métodos para obtener información de secuencia informativa sobre la presencia o ausencia de un tipo de trastorno hematológico, un método que comprende enriquecer el ADN convertido de un sujeto de prueba poniendo en contacto el ADN con una composición de oligonucleótidos cebo proporcionada anteriormente y secuenciar el ADN convertido enriquecido.

En la presente memoria también se proporcionan métodos para determinar que un sujeto de prueba tiene un tipo de trastorno hematológico (HD, por sus siglas en inglés), un método que comprende capturar fragmentos de ADNlc del sujeto de prueba con una composición de oligonucleótidos cebo proporcionada anteriormente, secuenciar los fragmentos de ADNlc capturados y aplicar un clasificador entrenado a las secuencias de ADNlc para determinar que el sujeto de prueba tiene el tipo de HD. El alcance específico del método de la presente invención es como se describe en las reivindicaciones adjuntas. En la presente memoria también se proporcionan (pero no se citan explícitamente en las reivindicaciones) métodos para determinar que un sujeto de prueba tiene un tipo de trastorno hematológico (HD), un método que comprende capturar fragmentos de ADNlc del sujeto de prueba con una composición de oligonucleótidos cebo proporcionada anteriormente, detectar los fragmentos de ADNlc capturados mediante micromatrices de ADN y aplicar un clasificador entrenado a los fragmentos de ADN hibridados con la micromatriz de ADN para determinar que el sujeto de prueba tiene el tipo de HD.

En algunas disposiciones, el clasificador entrenado determina la presencia o ausencia de cáncer y, si el clasificador determina la presencia de cáncer, el clasificador determina un tipo de cáncer. En algunas disposiciones, el tipo de cáncer se selecciona del grupo de consiste en cáncer de útero, cáncer escamoso del tracto gastrointestinal superior, todos los demás cánceres del tracto gastrointestinal superior, cáncer de tiroides, sarcoma, cáncer renal urotelial, todos los demás cánceres renales, cáncer de próstata, cáncer de páncreas, cáncer de ovario, cáncer neuroendocrino, mieloma múltiple, melanoma, linfoma, cáncer de pulmón microcítico, adenocarcinoma de pulmón, todos los demás cánceres de pulmón, leucemia, carcinoma hepatobiliar, hepatobiliar biliar, cáncer de cabeza y cuello, cáncer colorrectal, cáncer de cuello uterino, cáncer de mama, cáncer de vejiga y cáncer anorrectal. En algunas disposiciones, el tipo de cáncer se selecciona del grupo que consiste en cáncer anal, cáncer de vejiga, cáncer colorrectal, cáncer de esófago, cáncer de cabeza y cuello, cáncer de hígado/conducto biliar, cáncer de pulmón, linfoma, cáncer de ovario, cáncer de páncreas, neoplasia de células plasmáticas y cáncer de estómago. En algunas disposiciones, el tipo de cáncer se selecciona del grupo que consiste en cáncer de tiroides, melanoma, sarcoma, neoplasia mieloide, cáncer renal, cáncer de próstata, cáncer de mama, cáncer de útero, cáncer de ovario, cáncer de vejiga, cáncer urotelial, cáncer de cuello uterino, cáncer anorrectal, cáncer de cabeza y cuello, cáncer colorrectal, cáncer de hígado, cáncer de vías biliares, cáncer de páncreas, cáncer de vesícula biliar, cáncer del tracto gastrointestinal superior, mieloma múltiple, neoplasia linfoide y cáncer de pulmón. En algunas disposiciones, el tipo de cáncer es un HD el HD se selecciona del grupo que consiste en CHIP, leucemia, neoplasias linfoides (p. ej., linfoma), mieloma múltiple y una neoplasia mieloide. En algunas disposiciones, el tipo de trastorno hematológico se selecciona de neoplasia linfoide, mieloma múltiple y neoplasia mieloide. En algunas disposiciones, el clasificador entrenado es un clasificador de modelo de mezcla. En algunas disposiciones, el clasificador se entrenó en secuencias de ADN convertidas derivadas de al menos 1000, al menos 2000, o al menos 4000 regiones genómicas diana seleccionadas de una cualquiera de las listas 1-8. En algunos arreglos, el clasificador entrenado determina la presencia o ausencia de cáncer o de un tipo de cáncer mediante: (i) generar un conjunto de características para la muestra, en donde cada característica del conjunto de características comprende un valor numérico; (ii) introducir el conjunto de características en el clasificador, en donde el clasificador comprende un clasificador multinomial; (c) basándose en el conjunto de características, determinar, en el clasificador, un conjunto de puntuaciones de probabilidad, en donde el conjunto de puntuaciones de probabilidad comprende una puntuación de probabilidad por clase de tipo de cáncer y por clase de tipo distinto de cáncer; y (iv) el umbral del conjunto de puntuaciones de probabilidad basándose en uno o más valores determinados durante el entrenamiento del clasificador para determinar una clasificación final del cáncer de la muestra. En algunas disposiciones, el conjunto de características comprende un conjunto de características binarizadas. En algunas disposiciones, el valor numérico comprende un único valor binario. En algunas disposiciones, el clasificador multinomial comprende un conjunto de regresión logística multinomial entrenado para predecir un tejido fuente del cáncer. En algunas disposiciones, el método comprende además determinar la clasificación final del cáncer basándose en un diferencial de puntuación de las dos probabilidades superiores en relación con un valor mínimo, en donde el valor mínimo corresponde a un porcentaje predefinido de muestras de cáncer de entrenamiento que habían sido asignadas al tipo de cáncer correcto como su puntuación más alta durante el entrenamiento del clasificador. En

algunas disposiciones, (i) según la determinación de que el diferencial de puntuación de las dos probabilidades superiores supera el valor mínimo, asigne una etiqueta de cáncer correspondiente a la puntuación de probabilidad más alta determinada por el clasificador como clasificación final del cáncer; y (ii) según la determinación de que el diferencial de puntuación de las dos probabilidades superiores no supera el valor mínimo, asignando una etiqueta de cáncer indeterminada como la clasificación final del cáncer. En algunas disposiciones, el tipo de trastorno hematológico se selecciona entre CHIP, leucemia, neoplasias linfoides (p. ej., linfoma), mieloma múltiple y una neoplasia mieloide. En algunas disposiciones, el tipo de trastorno hematológico se selecciona de neoplasia linfoide, mieloma múltiple y neoplasia mieloide. En algunas disposiciones, se determina que el sujeto tiene un cáncer y la especificidad es al menos 0.990. En algunas realizaciones, la relación entre la probabilidad de determinar con precisión un trastorno hematológico y la probabilidad de determinar de forma incorrecta un tumor sólido es al menos 25:1 o al menos 50:1. En algunas disposiciones, la relación entre la probabilidad de determinar con precisión un trastorno hematológico y la probabilidad de determinar de forma incorrecta un trastorno hematológico es de al menos 8:, al menos 12:1, o al menos 16:1. En algunas disposiciones, la probabilidad de determinar de forma incorrecta un tipo de cáncer es de al menos el 80 %, al menos el 85 % o al menos el 89 %. En algunos casos, el cáncer es un cáncer en estadio I y la probabilidad de determinar con precisión un tipo de cáncer es de al menos el 65 %, al menos el 70 %, al menos el 75 % o al menos el 80 %. En algunos casos, el cáncer es un cáncer en estadio II y la probabilidad de determinar con precisión un tipo de cáncer es de al menos el 75 %, al menos el 80 %, al menos el 85 % o al menos el 90 %. En algunos casos, el cáncer es un cáncer en estadio III o un cáncer en estadio IV y la probabilidad de determinar con precisión un tipo de cáncer es de al menos el 85 % o al menos el 90 %. En algunas disposiciones, la sensibilidad para el mieloma múltiple es de al menos el 55 %, al menos el 65 %, al menos el 75 % o al menos el 85 %. En algunos casos, la sensibilidad para el mieloma múltiple en estadio I es de al menos el 60 %, al menos el 65 % o al menos el 70 %. En algunos casos, la sensibilidad para el mieloma múltiple en estadio II es de al menos el 60 %, al menos el 75 % o al menos el 85 %. En algunas disposiciones, la composición de oligonucleótidos cebo está configurada para hibridarse con el ADNlc derivado de las regiones genómicas diana de la lista 3 o la lista 6. En algunas disposiciones, la sensibilidad para la neoplasia linfoide es de al menos el 55 %, al menos el 60 %, al menos el 65 % o al menos el 70 %. En algunas disposiciones, la sensibilidad para la neoplasia linfoide en estadio I es de al menos el 30 %. En algunas disposiciones, la sensibilidad para la neoplasia linfoide en estadio II es de al menos el 65 %, al menos el 75 %, al menos el 85 % o al menos el 90 %. En algunas disposiciones, la composición de oligonucleótidos cebo está configurada para hibridarse con el ADNlc derivado de las regiones genómicas diana de la lista 2 o la lista 5.

En la presente memoria también se proporcionan (pero no se reivindican) paneles de ensayo de trastorno hematológico (HD) que comprenden: al menos 500 pares de sondas, en donde cada par de los al menos 500 pares comprende dos sondas configuradas para superponerse entre sí mediante una secuencia de superposición, en donde la secuencia de superposición comprende una secuencia de al menos 30 nucleótidos, y en donde la secuencia de los al menos 30 nucleótidos está configurada para hibridarse con una molécula de ADNlc convertido correspondiente a, o derivada de una o más de las regiones genómicas, en donde cada una de las regiones genómicas comprende al menos cinco sitios de metilación, y en donde los al menos cinco sitios de metilación tienen un patrón de metilación anormal en muestras de HD. En algunas disposiciones, cada sonda de los al menos 5 pares de sondas comprende una secuencia no superpuesta de al menos 31 nucleótidos. En algunas disposiciones, las moléculas de ADNlc convertido comprenden moléculas de ADNlc tratadas para convertir la C (citosina) no metilada en U (uracilo). En algunas disposiciones, cada uno de los al menos 500 pares de sondas se conjuga con un resto de afinidad no nucleotídico. En algunas disposiciones, el resto de afinidad no nucleotídico es un resto de biotina. En algunas disposiciones, las muestras de HD provienen de sujetos que tienen un trastorno hematológico seleccionado del grupo que consiste en CHIP, leucemia, mieloma múltiple y linfoma. En algunas disposiciones, el patrón de metilación anormal tiene al menos un valor de p umbral de rareza en las muestras de HD. En algunas disposiciones, cada una de las sondas está diseñada para tener una homología de secuencia o complementariedad con menos de 20 regiones genómicas fuera de la diana. En algunas disposiciones, las menos de 20 regiones genómicas fuera de la diana se identifican usando una estrategia de siembra k-mero. En algunas disposiciones, las menos de 20 regiones genómicas fuera de la diana se identifican usando la estrategia de siembra k-mero combinada con la alineación local en las ubicaciones de siembra. En algunas disposiciones, el panel de ensayo de HD comprende al menos 1.000, 2.000, 5.000, 10.000, 50.000, 100.000, 150.000, 200.000, o 250.000 sondas. En algunas disposiciones, los al menos 500 pares de sondas juntos comprenden al menos 10.000, 20.000, 30.000, 40.000, 50.000, 60.000, 70.000, 80.000, 90.000, 100.000, 120.000, 140.000, 160.000, 180.000, 200.000, 240.000, 260.000, 280.000, 300.000, 320.000, 400.000, 450.000, 500.000, 550.000, 600.000, 650.000, 700.000, 750.000, 800.000, 850.000, 900.000, 1 millón, 1,5 millones, 2 millones, 2,5 millones, 3 millones, 3,5 millones, 4 millones, 4,5 millones o 5 millones de nucleótidos. En algunas disposiciones, de las sondas comprende al menos 50, 75, 100 o 120 nucleótidos. En algunas disposiciones, cada una de las sondas comprende menos de 300, 250, 200 o 150 nucleótidos. En algunas disposiciones, cada una de las sondas comprende 100-150 nucleótidos. En algunas disposiciones, cada una de las sondas comprende menos de 20, 15, 10, 8 o 6 sitios de metilación. En algunas disposiciones, al menos el 80, el 85, el 90, el 92, el 95 o el 98 % de los al menos cinco sitios de metilación están metilados o no metilados en las muestras de HD. En algunas disposiciones, al menos el 3 %, el 5 %, el 10 %, el 15 % o el 20 % de las sondas no contienen G (guanina). En algunas disposiciones, cada una de las sondas comprende múltiples sitios de unión a los sitios de metilación de la molécula de ADNlc convertido, en donde al menos el 80, el 85, el 90, el 92, el 95 o el 98 % de los múltiples sitios de unión comprenden exclusivamente CpG o CpA. En algunas disposiciones, cada una de las sondas está configurada para tener homología de secuencias o complementariedad de secuencias con menos de 15, 10 u 8 regiones genómicas fuera de la diana. En algunas disposiciones, al menos el 30 % de las regiones genómicas están en exones o intrones.

En algunas disposiciones, al menos el 15 % de las regiones genómicas están en exones. En algunas disposiciones, al menos el 20 % de las regiones genómicas están en exones. En algunas disposiciones, menos del 10 % de las regiones genómicas están en regiones intergénicas. En algunas disposiciones, las regiones genómicas se seleccionan de la lista 1. En algunas disposiciones, las regiones genómicas comprenden al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 %, el 95 % o el 100 % de las regiones genómicas en la lista 1. En algunas disposiciones, las regiones genómicas comprenden al menos 100, 200, 300, 400, 500, 1.000, 5.000, 10.000, 15.000, 16.000, 17.000, 18.000, 19.000, 20.000, 21.000 o 23.000 regiones genómicas de la lista 1.

En la presente memoria también se proporcionan métodos para detectar un trastorno hematológico (HD), que comprenden: recibir una muestra que comprende una pluralidad de moléculas de ADNlc; tratar la pluralidad de moléculas de ADNlc para convertir C (citosina) no metilada en U (uracilo), obteniendo así una pluralidad de moléculas de ADNlc convertido; aplicar un panel de ensayo de HD de una cualquiera de las disposiciones anteriores a la pluralidad de moléculas de ADNlc convertido, enriqueciendo así un subconjunto de las moléculas de ADNlc convertido; y secuenciar el subconjunto enriquecido de la molécula de ADNlc convertido, proporcionando de este modo un conjunto de lecturas de secuencia. En algunas disposiciones, el método comprende además la etapa de: determinar una condición de salud evaluando el conjunto de lecturas de secuencia, en donde la condición de salud es una presencia o ausencia de trastorno hematológico; una etapa de un trastorno hematológico; una presencia o ausencia de un tipo de cáncer de la sangre; o una presencia o ausencia de al menos 1, 2 o 3 tipos diferentes de trastornos hematológicos. En algunas disposiciones, la muestra que comprende una pluralidad de moléculas de ADNlc se obtuvo de un sujeto humano. En algunas disposiciones, el trastorno hematológico se selecciona del grupo que consiste en: neoplasia linfoide, mieloma múltiple y neoplasia mieloide.

En la presente memoria también se proporcionan métodos para detectar un trastorno hematológico (HD), que comprenden las etapas de: obtener un conjunto de lecturas de secuencia mediante secuenciación de un conjunto de fragmentos de ácido nucleico de un sujeto, en donde cada uno de los fragmentos de ácido nucleico corresponde a, o se derivan de una pluralidad de regiones genómicas seleccionadas de una cualquiera de las listas 1-8; para cada una de las lecturas de secuencia, determinar el estado de metilación en una pluralidad de sitios de CpG; y detectar un trastorno hematológico del sujeto evaluando el estado de metilación para las lecturas de secuencia, en donde el trastorno hematológico comprende uno o más de los siguiente: (i) una presencia o ausencia de un trastorno hematológico; (ii) un estadio de un trastorno hematológico; (iii) una presencia o ausencia de un tipo de cáncer de la sangre; y (iv) una presencia o ausencia de al menos 1, 2, o 3 tipos diferentes de trastornos hematológicos. En algunas disposiciones, la pluralidad de regiones genómicas comprende al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 %, el 95 % o el 100 % de las regiones genómicas de la lista 1. En algunas disposiciones, la pluralidad de regiones genómicas comprende 100, 200, 300, 400, 500, 1.000, 5.000, 10.000, 15.000, 16.000, 17.000, 18.000, 19.000, 20.000, 21.000 o 23.000 de las regiones genómicas de una cualquiera de las listas 1-8.

En la presente memoria también se proporcionan (pero no se reivindican) métodos para diseñar un panel de ensayo de trastorno hematológico (HD) que comprende las etapas de: identificar una pluralidad de regiones genómicas, en donde cada una de la pluralidad de regiones genómicas (i) comprende al menos 30 nucleótidos, y (ii) comprende al menos cinco sitios de metilación, seleccionar un subconjunto de las regiones genómicas, en donde la selección se realiza cuando las moléculas de ADNlc correspondientes a, o derivadas de cada una de las regiones genómicas en muestras de HD tienen un patrón de metilación anormal, en donde el patrón de metilación anormal comprende al menos cinco sitios de metilación hipometilados o hipermetilados, y diseñar un panel de ensayo de HD que comprende una pluralidad de sondas, en donde cada una de las sondas está configurada para hibridarse con una molécula de ADNlc convertido correspondiente a o derivada de uno o más del subconjunto de las regiones genómicas. En algunas disposiciones, las moléculas de ADNlc convertido comprenden moléculas de ADNlc tratadas para convertir las citosinas no metiladas en uracilos.

En la presente memoria también se proporcionan (pero no se reivindican) paneles de ensayo de trastorno hematológico (HD) que comprenden una pluralidad de sondas, en donde cada una de la pluralidad de sondas está configurada para hibridarse con una molécula de ADNlc convertido correspondiente a una o más de las regiones genómicas en la lista 1. En algunas disposiciones, las moléculas de ADNlc convertido comprenden moléculas de ADNlc tratadas para convertir las citosinas no metiladas en uracilos. En algunas disposiciones, la pluralidad de sondas está configurada para hibridarse con una pluralidad de moléculas de ADNlc convertido correspondientes a o derivadas de al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 % o el 90 %, el 95 % o el 100 % de las regiones genómicas de una cualquiera de las listas 1-8. En algunas disposiciones, la pluralidad de sondas está configurada para hibridarse con una pluralidad de moléculas de ADNlc convertido correspondientes a o derivadas de al menos 100, 200, 300, 400, 500, 1.000, 5.000, 10.000, 15.000, 16.000, 17.000, 18.000, 19.000, 20.000, 21.000 o 23.000 regiones genómicas de una cualquiera de las listas 1-8. En algunas disposiciones, al menos el 3 %, el 5 %, el 10 %, el 15 % o el 20 % de las sondas no contienen G (guanina). En algunas disposiciones, cada una de las sondas comprende múltiples sitios de unión a los sitios de metilación de la molécula de ADNlc convertido, en donde al menos el 80, el 85, el 90, el 92, el 95 o el 98 % de los múltiples sitios de unión comprenden exclusivamente CpG o CpA. En algunas disposiciones, cada una de las sondas se conjuga con un resto de afinidad no nucleotídico. En algunas disposiciones, el resto de afinidad no nucleotídico es un resto de biotina.

Breve descripción de los dibujos

Las características de la descripción se exponen con particularidad en las reivindicaciones adjuntas. Se obtendrá una mejor comprensión de las características y ventajas de la presente descripción mediante referencia a la siguiente descripción detallada que establece realizaciones ilustrativas, en las que se utilizan los principios de la descripción, y los dibujos adjuntos de los cuales:

La Figura 1A ilustra un diseño de sonda en forma de dos títulos, con tres sondas dirigidas a una región diana pequeña, donde cada base en una región diana (recuadrada en el rectángulo punteado) está cubierta por al menos dos sondas.

La Figura 1B ilustra un diseño de sonda en forma de dos títulos, con más de tres sondas dirigidas a una región diana más grande, donde cada base en una región diana (recuadrada en el rectángulo punteado) está cubierta por al menos dos sondas.

La Figura 1C ilustra el diseño de sondas dirigidas a fragmentos hipometilados y/o hipermetilados en regiones genómicas.

La Figura 2 ilustra un proceso para generar un panel de ensayo de Heme.

La **Figura 3A** es un diagrama de flujo que describe un proceso para crear una estructura de datos para un grupo de control.

La **Figura 3B** es un diagrama de flujo que describe una etapa adicional de validar la estructura de datos para el grupo de control de la Figura 3A.

La **Figura 4** es un diagrama de flujo que describe un proceso para seleccionar regiones genómicas para diseñar sondas para un panel de ensayo de HD.

La **Figura 5** es una ilustración de un cálculo de puntuación de valor de p de ejemplo.

La **Figura 6A** es un diagrama de flujo que describe un proceso de entrenamiento de un clasificador basándose en fragmentos hipometilados e hipermetilados indicativos de un trastorno hematológico.

La **Figura 6B** es un diagrama de flujo que describe un proceso de identificación de fragmentos indicativos del cáncer determinado por modelos probabilísticos.

La **Figura 7A** es un diagrama de flujo que describe un proceso de secuenciación de un fragmento de ADN libre de células (Ic).

La **Figura 7B** es una ilustración del proceso de la **Figura 7A** de secuenciación de un fragmento de ADN libre de células (Ic) para obtener un vector de estado de metilación.

La **Figura 8** es un gráfico de las cantidades de fragmentos de ADN que se hibridan a sondas dependiendo de los tamaños de superposiciones entre los fragmentos de ADN y las sondas.

La **Figura 9A** ilustra un diagrama de flujo de dispositivos para secuenciar muestras de ácido nucleico según una disposición. La **Figura 9B** ilustra un sistema analítico que analiza el estado de metilación del ADNIc.

La **Figura 10** es una curva operadora del receptor que compara la tasa de verdaderos positivos y la tasa de falsos positivos de detección del cáncer mediante un clasificador entrenado que utiliza información del estado de metilación de un 50 % aleatorio de las regiones genómicas objetivo de la lista 8.

Descripción detallada

Definiciones

A menos que se defina lo contrario, todos los términos técnicos y/o científicos usados en la presente memoria tienen el mismo significado que entiende comúnmente un experto en la técnica a la que pertenece la invención. Como se usa en la presente memoria, los siguientes términos tienen los significados atribuidos a continuación.

Como se usa en la presente memoria, cualquier referencia a "una realización" o "una realización" significa que un elemento, característica, estructura o característica particular descrita en relación con la realización se incluye en al menos una realización. Las apariciones de la frase "en una realización" en varios lugares de la especificación no se refieren necesariamente a la misma realización, proporcionando de este modo un marco para diversas posibilidades de las realizaciones descritas para funcionar en conjunto.

Como se usa en la presente memoria, los términos "comprende", "que comprende", "incluye", "que incluye", "tiene", "que tiene" o cualquier otra variación de los mismos, pretenden cubrir una inclusión no exclusiva. Por ejemplo, un proceso, método, artículo o aparato que comprende una lista de elementos no se limita necesariamente a solo esos elementos, sino que puede incluir otros elementos no expresamente enumerados o inherentes a dicho proceso, método, artículo o aparato. Además, a menos que se indique expresamente lo contrario, "o" se refiere a un inclusivo o y no a un o exclusivo o. Por ejemplo, una condición A o B se satisface por uno cualquiera de los siguientes: A es verdadero (o presente) y B es falso (o no está presente), A es falso (o no está presente) y B es verdadero (o presente), y tanto A como B son verdaderos (o presentes).

Además, el uso de "un" o "una" se emplea para describir elementos y componentes de los realizaciones en la presente memoria. Esto se hace simplemente por conveniencia y para dar un sentido general de la descripción. Esta descripción debe leerse para incluir uno o al menos uno y el singular también incluye el plural a menos que sea obvio que se pretende de otro modo.

Como se usa en la presente memoria, los intervalos y cantidades pueden expresarse como "aproximadamente" un valor o intervalo particular. También incluye la cantidad exacta. Por tanto, "aproximadamente 5 µg" significa "aproximadamente 5 µg" y también "5 µg." Generalmente, el término "aproximadamente" incluye una cantidad que se esperaría que esté dentro del error experimental. En algunas disposiciones, "aproximadamente" se refiere al número o valor mencionado, "+" o "-" 20 %, 10 % o 5 % del número o valor. Además, se entiende que los intervalos citados en la presente memoria son abreviados para todos los valores dentro del intervalo, incluidos los puntos finales enumerados. Por ejemplo, se entiende que un intervalo de 1 a 50 incluye cualquier número, combinación de números o subintervalo del grupo que consiste en 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, 40, 41, 42, 43, 44, 40, 41, 42, 43, 49 y 50.

El término "trastorno hematológico" o "HD", tal como se usa en la presente memoria, se refiere a un trastorno que afecta principalmente a la sangre, seleccionado del grupo que consiste en CHIP, leucemia, neoplasias linfoides (p. ej., linfoma), mieloma múltiple y neoplasia mieloide.

El término "metilación", como se usa en la presente memoria, se refiere a un proceso mediante el cual se añade un grupo metilo a una molécula de ADN. Por ejemplo, un átomo de hidrógeno en el anillo de pirimidina de una base de citosina se puede convertir en un grupo metilo, formando 5-metilcitosina. El término también se refiere a un proceso por el que se añade un grupo hidroximetilo a una molécula de ADN, por ejemplo mediante oxidación de un grupo metilo en el anillo de pirimidina de una base de citosina. La metilación y la hidroximetilación tienden a producirse en dinucleótidos de citosina y guanina a los que se hace referencia en la presente memoria como "sitios CpG"

El término "metilación" también puede referirse al estado de metilación de un sitio CpG. Un sitio CpG con un resto 5-metilcitosina está metilado. Un sitio CpG con un átomo de hidrógeno en el anillo de pirimidina de la base de citosina no está metilado.

En tales disposiciones, el ensayo de laboratorio húmedo utilizado para detectar la metilación puede variar de los descritos en la presente memoria como es bien conocido en la técnica.

El término "sitio de metilación" como se usa en la presente memoria se refiere a una región de una molécula de ADN donde se puede añadir un grupo metilo. Los sitios "CpG" son el sitio de metilación más común, pero los sitios de metilación no se limitan a sitios CpG. Por ejemplo, la metilación del ADN puede producirse en citosinas en CHG y CHH, donde H es adenina, citosina o timina. La metilación de citosina en forma de 5-hidroximetilcitosina también puede evaluarse (véase, p. ej., los documentos WO 2010/037001 y WO 2011/127136), y sus características, usando los métodos y procedimientos descritos en la presente memoria.

El término "sitio CpG" como se usa en la presente memoria se refiere a una región de una molécula de ADN donde un nucleótido de citosina es seguido por un nucleótido de guanina en la secuencia lineal de bases a lo largo de su dirección 5' a 3'. "CpG" es una abreviatura de 5'-C-fosfato-G-3', es decir, citosina y guanina separadas por un solo grupo fosfato. Las citosinas en los dinucleótidos CpG se pueden metilar para formar 5-metilcitosina.

El término "sitio de detección CpG" como se usa en la presente memoria se refiere a una región en una sonda que está configurada para hibridarse con un sitio CpG de una molécula de ADN diana. El sitio CpG en la molécula de ADN diana puede comprender citosina y guanina separadas por un grupo fosfato, donde citosina está metilada o no metilada. El sitio CpG en la molécula de ADN diana puede comprender uracilo y guanina separados por un grupo fosfato, donde el uracilo se genera por la conversión de citosina no metilada.

El término "UpG" es una abreviatura de 5'-U-fosfato-G-3', que es uracilo y guanina separados por un solo grupo fosfato. UpG puede generarse mediante un tratamiento con bisulfito de un ADN que convierte las citosinas no metiladas en uracilos. Las citosinas pueden convertirse en uracilos por otros métodos conocidos en la técnica, como la modificación química, la síntesis o la conversión enzimática.

Los términos "hipometilado" o "hipermetilado" como se usan en la presente memoria se refieren al estado de metilación de una molécula de ADN que contiene múltiples sitios CpG (p. ej., más de 3, 4, 5, 6, 7, 8, 9, 10, etc.) donde un alto porcentaje de los sitios CpG (p. ej., más del 80 %, el 85 %, el 90 % o el 95 %, o cualquier otro porcentaje dentro del intervalo del 50 %-100 %) no están metilados o están metilados, respectivamente.

Los términos "vector de estado de metilación" o "vector de estado de metilación" como se usan en la presente memoria se refieren a un vector que comprende múltiples elementos, donde cada elemento indica el estado de metilación de un sitio de metilación en una molécula de ADN que comprende múltiples sitios de metilación, en el orden en que aparecen de 5' a 3' en la molécula de ADN. Por ejemplo, $< M_x, M_{x+1}, M_{x+2} >$, $< M_x, M_{x+1}, U_{x+2} >$, ..., $< U_x, U_{x+1}, U_{x+2} >$ pueden ser vectores de metilación para moléculas de ADN que comprenden tres sitios de metilación, donde M representa un sitio de metilación metilado y U representa un sitio de metilación no metilado.

El término "patrón de metilación anormal" o "patrón de metilación anómalo" como se usa en la presente memoria se refiere al patrón de metilación de una molécula de ADN o un vector de estado de metilación que se espera que se encuentre en una muestra con menos frecuencia que un valor umbral. En una disposición proporcionada en la presente memoria, la expectativa de encontrar un vector de estado de metilación específico en un grupo de control sano que comprende individuos sanos está representado por un valor de p. Una puntuación de valor de p baja corresponde generalmente a un vector de estado de metilación que es relativamente inesperado en comparación con otros vectores de estado de metilación dentro de muestras de individuos sanos. Una puntuación de valor de p alta corresponde generalmente a un vector de estado de metilación que es relativamente más esperado en comparación con otros vectores de estado de metilación que se encuentran en muestras de individuos sanos en el grupo de control sano. Un vector de estado de metilación que tiene un valor de p inferior a un valor umbral (por ejemplo, 0,1, 0,01, 0,001, 0,0001, etc.) puede definirse como un patrón de metilación anormal/anómalo. Pueden usarse diversos métodos conocidos en la técnica para calcular un valor de p o expectativa de un patrón de metilación o un vector de estado de metilación. Los métodos ilustrativos proporcionados en la presente memoria implican el uso de una probabilidad de cadena de Markov que asume que los estados de metilación de los sitios CpG dependen de los estados de metilación de los sitios CpG vecinos. Los métodos alternativos proporcionados en la presente memoria calculan la expectativa de observar un vector de estado de metilación específico en individuos sanos utilizando un modelo de mezcla que incluye múltiples componentes de mezcla, siendo cada uno un modelo de sitios independientes donde se supone que la metilación en cada sitio CpG es independiente de los estados de metilación en otros sitios CpG.

El término "muestra de HD" como se usa en la presente memoria se refiere a una muestra que comprende ADN genómicos de un individuo diagnosticado con un trastorno hematológico. Los ADN genómicos pueden ser, pero no se limitan a, fragmentos de ADNlc o ADN cromosómico de un sujeto con un trastorno hematológico. Los ADN genómicos pueden secuenciarse (o de otro modo detectarse) y su estado de metilación puede evaluarse mediante métodos conocidos en la técnica, por ejemplo, la secuenciación por bisulfito. Cuando las secuencias genómicas se obtienen de una base de datos pública (p. ej., The Cancer Genome Atlas [TCGA]) o se obtienen experimentalmente secuenciando un genoma de un individuo diagnosticado con un trastorno hematológico, la muestra de HD puede referirse a ADN genómicos o fragmentos de ADNlc que tienen las secuencias genómicas. El término "muestras de HD" como un plural se refiere a muestras que comprenden ADN genómicos de múltiples individuos, cada individuo diagnosticado con un trastorno hematológico. En diversas disposiciones, se usan muestras de HD de más de 10, 20, 50, 100, 200, 300, 500, 1.000, 2.000, 5.000, 10.000, 20.000, 40.000, 50.000, o más individuos diagnosticados con un trastorno hematológico.

El término "muestra no HD" o "muestra sana", como se usa en la presente memoria, se refiere a una muestra que comprende ADN genómicos de un individuo no diagnosticado con un trastorno hematológico. Los ADN genómicos pueden ser, pero no se limitan a, fragmentos de ADNlc o ADN cromosómico de un sujeto sin un trastorno hematológico (p. ej., un sujeto sano). Los ADN genómicos pueden secuenciarse (o de otro modo detectarse) y su estado de metilación puede evaluarse mediante métodos conocidos en la técnica, por ejemplo, la secuenciación por bisulfito. Cuando las secuencias genómicas se obtienen de una base de datos pública (p. ej., The Cancer Genome Atlas [TCGA]) o se obtienen experimentalmente secuenciando un genoma de un individuo sin un trastorno hematológico, la muestra sin HD puede referirse a ADN genómicos o fragmentos de ADNlc que tienen las secuencias genómicas. El término "muestras sin HD" como un plural se refiere a muestras que comprenden ADN genómicos de múltiples individuos, cada individuo está sin un trastorno hematológico. En diversas disposiciones, se usan muestras sanas de más de 10, 20, 50, 100, 200, 300, 500, 1.000, 2.000, 5.000, 10.000, 20.000, 40.000, 50.000, o más individuos sin un trastorno hematológico.

El término "muestra de entrenamiento" como se usa en la presente memoria se refiere a una muestra usada para entrenar un clasificador descrito en la presente memoria y/o para seleccionar una o más regiones genómicas para la detección de un trastorno hematológico. Las muestras de entrenamiento pueden comprender ADN genómicos o una modificación del mismo, de uno o más sujetos sanos y de uno o más sujetos que tienen un trastorno hematológico. Los ADN genómicos pueden ser, pero no se limitan a, fragmentos de ADNlc o ADN cromosómico. Los ADN genómicos pueden secuenciarse (o de otro modo detectarse) y su estado de metilación puede evaluarse mediante métodos conocidos en la técnica, por ejemplo, la secuenciación por bisulfito. Cuando las secuencias genómicas se obtienen de una base de datos pública (por ejemplo, The Cancer Genome Atlas [TCGA]) o se obtienen experimentalmente secuenciando el genoma de un individuo, una muestra de entrenamiento puede referirse a ADN genómico o fragmentos de ADNlc que tengan las secuencias genómicas.

El término "muestra de prueba" como se usa en la presente memoria se refiere a una muestra de un sujeto, cuya condición de salud se ha probado o se probará usando un clasificador y/o un panel de ensayo descrito en la presente memoria. La muestra de prueba puede comprender ADN genómicos o una modificación de la misma. Los ADN genómicos pueden ser, pero no se limitan a, fragmentos de ADNlc o ADN cromosómico.

El término "región genómica diana", como se usa en la presente memoria, se refiere a una región en un genoma seleccionado para el análisis en muestras de prueba. Se genera un panel de ensayo con sondas diseñadas para hibridarse con fragmentos de ácido nucleico (y opcionalmente eliminar) derivados de la región genómica diana o un fragmento de la misma. Un fragmento de ácido nucleico derivado de la región genómica diana se refiere a un fragmento de ácido nucleico generado por degradación, escisión, conversión con bisulfito u otro procesamiento del ADN de la región genómica diana.

Se describen varias regiones genómicas diana según su ubicación cromosómica en el listado de secuencias presentado aquí. El ADN cromosómico es bicatenario, por lo que una región genómica diana incluye dos cadenas de ADN: uno con la secuencia proporcionada en el listado y un segundo que es un complemento inverso a la secuencia en el listado. Las sondas pueden diseñarse para hibridarse con una o ambas secuencias. Opcionalmente, las sondas se hibridan con secuencias convertidas resultantes de, por ejemplo, tratamiento con bisulfito de sodio.

El término "región genómica fuera de diana", como se usa en la presente memoria, se refiere a una región en un genoma que no se ha seleccionado para su análisis en muestras de prueba, pero tiene una homología suficiente con una región genómica diana para potencialmente unirse y extraerse por una sonda diseñada para dirigirse a la región genómica diana. En una disposición, una región genómica fuera de la diana es una región genómica que se alinea con una sonda a lo largo de al menos 45 pb con al menos una tasa de coincidencia del 90 %.

Los términos "moléculas de ADN convertido", "moléculas de ADNlc convertido" y "fragmento modificado obtenido a partir del procesamiento de las moléculas de ADNlc" se refieren a moléculas de ADN obtenidas a partir del procesamiento de moléculas de ADN o ADNlc en una muestra con el fin de diferenciar un nucleótido metilado y un nucleótido no metilado en las moléculas de ADN o ADNlc. Por ejemplo, en una disposición, la muestra puede tratarse con ion bisulfito (p. ej., usando bisulfito de sodio), como es bien conocido en la técnica, para convertir citosinas no metiladas ("C") en uracilos ("U"). En otra disposición, la conversión de citosinas no metiladas en uracilos se logra usando una reacción de conversión enzimática, por ejemplo, usando una citidina desaminasa (tal como APOBEC). Después del tratamiento, las moléculas de ADN transformadas o las moléculas de ADNlc incluyen uracilos adicionales que no están presentes en la muestra de ADNlc original. Replicación por ADN polimerasa de una cadena de ADN que comprende un uracilo resulta en la adición de una adenina a la cadena complementaria naciente en lugar de la guanina añadida normalmente como complemento a una citosina o metilcitosina.

Los términos "ácido nucleico libre de células", "ADN libre de células" o "ADNlc" se refieren a fragmentos de ácido nucleico que circulan en el cuerpo de un individuo (p. ej., torrente sanguíneo) y se originan a partir de una o más células sanas y/o de una o más células de HD (es decir, células de un sujeto que tiene un trastorno hematológico). Además, el ADNlc puede provenir de otras fuentes tales como virus, fetos, etc.

El término "ADN tumoral circulante" o "ADNtc" se refiere a fragmentos de ácido nucleico que se originan a partir de células tumorales, que pueden liberarse en el torrente sanguíneo de un individuo como resultado de procesos biológicos tales como apoptosis o necrosis de células que mueren o se liberan activamente por células tumorales viables.

El término "fragmento" como se usa en la presente memoria puede referirse a un fragmento de una molécula de ácido nucleico. Por ejemplo, en una disposición, un fragmento puede referirse a una molécula de ADNlc en una muestra de sangre o plasma, o una molécula de ADNlc que se ha extraído de una muestra de sangre o plasma. Un producto de amplificación de una molécula de ADNlc también puede denominarse "fragmento". En otra disposición, el término "fragmento" se refiere a una lectura de secuencia o conjunto de lecturas de secuencia, que se han procesado para el análisis posterior (p. ej., para la clasificación basándose en aprendizaje automático), como se describe en la presente memoria. Por ejemplo, como se conoce bien en la técnica, las lecturas de secuencia sin procesar pueden alinearse con un genoma de referencia y las lecturas de secuencia de extremos emparejados coincidentes ensambladas en un fragmento más largo para el análisis posterior.

El término "individuo" se refiere a un individuo humano. El término "individuo sano" se refiere a un individuo que se supone que no tiene un trastorno hematológico.

El término "sujeto" se refiere a un individuo cuyo ADN se está analizando. Un sujeto puede ser un sujeto de prueba cuyo ADN se evalúa usando de un panel diana como se describe en la presente memoria para evaluar si la persona tiene un trastorno hematológico u otra enfermedad. Un sujeto también puede ser parte de un grupo de control que se sabe que no tiene trastorno hematológico u otra enfermedad. Un sujeto también puede formar parte de un grupo de trastorno hematológico u otra enfermedad conocido por tener un trastorno hematológico u otra enfermedad. Los grupos de control y de cáncer/enfermedad pueden usarse para ayudar a diseñar o validar el panel específico.

El término "lecturas de secuencia" como se usa en la presente memoria se refiere a las lecturas de secuencias de nucleótidos de una muestra. Las lecturas de secuencia se pueden obtener a través de diversos métodos proporcionados en la presente memoria o como se conoce en la técnica.

El término "profundidad de secuenciación", como se usa en la presente memoria, se refiere al recuento del número de veces que se ha secuenciado un ácido nucleico diana determinado dentro de una muestra (p. ej., el recuento de lecturas de secuencia en una región diana dada). El aumento de la profundidad de secuenciación puede reducir las cantidades requeridas de ácidos nucleicos necesarios para evaluar un estado de enfermedad (p. ej., estado de la enfermedad hematológica).

El término "tejido de origen" o "TOO" como se usa en la presente memoria se refiere al órgano, grupo de órganos, región de cuerpo o tipo de célula del que surge o se origina un trastorno hematológico. La identificación de un tejido de origen o de células cancerosas típicamente permite la identificación de las siguientes etapas más apropiadas en el continuo de cuidado del cáncer para detectar, diagnosticar aún más, estadificar y decidir sobre el tratamiento.

El término "transición" generalmente se refiere a cambios en la composición base de una purina a otra purina, o de una pirimidina a otra pirimidina. Por ejemplo, los siguientes cambios son transiciones C→U, U→C, G→A, A→G, C→T y T→C.

"Una totalidad de las sondas" de un panel o conjunto de cebos o "una totalidad de sondas que contienen polinucleótidos" de un panel o conjunto de cebos generalmente se refiere a todas las sondas administradas con un panel específico o conjunto de cebos. Por ejemplo, en algunas disposiciones, un panel o conjunto de cebos puede incluir (1) sondas que tienen características especificadas en la presente memoria (p. ej., sondas para unirse a fragmentos de ADN libres de células correspondientes o derivadas de regiones genómicas expuestas en la presente memoria en una o más listas) y (2) sondas adicionales que no contienen tal(es) característica(s). La totalidad de las sondas de un panel generalmente se refiere a todas las sondas suministradas con el panel o conjunto de cebos, incluyendo tales sondas que no contienen la(s) característica(s) especificada(s).

Panel de ensayo de HD

En una primera disposición, la presente descripción proporciona un panel de ensayo de HD que comprende una pluralidad de sondas o una pluralidad de pares de sondas. Los paneles de ensayo descritos en la presente memoria pueden denominarse alternativamente conjuntos cebos o como composiciones que comprenden oligonucleótidos cebo. Las sondas pueden ser sondas que contienen polinucleótidos que están diseñadas específicamente para dirigirse a una o más moléculas de ácido nucleico correspondientes a, o derivadas de, regiones genómicas metiladas diferencialmente entre muestras de HD y sin HD, entre diferentes tipos de HD, entre CHIP y otras muestras de HD, entre diferentes tipos de tejido canceroso de origen (TOO, por sus siglas en inglés) o entre muestras de diferentes estadios de HD. En algunas disposiciones, las regiones genómicas diana (o moléculas de ácido nucleico derivadas de las mismas) se seleccionan para maximizar la precisión de la clasificación, sujeto a un presupuesto de tamaño (que se determina por presupuesto de secuenciación y profundidad de secuenciación deseada).

El diseño y la utilidad del panel de ensayo HD se describen generalmente en la **Figura 2.** Para diseñar el panel de ensayo de HD, un sistema de análisis recopila información sobre el estado de metilación de los sitios CpG de los fragmentos de ácido nucleico de las muestras correspondientes a los diversos resultados considerados, p. ej., muestras que se sabe que tienen HD, muestras con o sin CHIP, muestras con HD distintas de CHIP, muestras consideradas sanas, muestras de un TOO conocido, etc. Estas muestras pueden procesarse usando uno o más métodos conocidos en la técnica para determinar el estado de metilación de los sitios CpG (p. ej., con secuenciación con bisulfito del genoma completo [WGBS, por sus siglas en inglés]), o la información puede obtenerse de una base de datos pública (p. ej., TCGA). El sistema de análisis puede ser cualquier sistema informático genérico con un procesador de ordenador y un medio de almacenamiento legible por ordenador con instrucciones para ejecutar el procesador de ordenador para realizar cualquiera o todas las operaciones descritas en esta presente descripción.

La metodología ilustrativa para diseñar un panel de ensayo de trastorno hematológico se describe generalmente en la figura 2. Por ejemplo, para diseñar un panel de ensayo de trastorno hematológico, un sistema de análisis puede recoger información sobre el estado de metilación de los sitios CpG de fragmentos de ácido nucleico de muestras correspondientes a diversos resultados en consideración, p. ej., muestras que se sabe que tienen trastorno hematológico, muestras consideradas para ser sanas, etc. Estas muestras pueden procesarse (p. ej., con secuenciación con bisulfito de genoma completo [WGBS]) para determinar el estado de metilación de los sitios CpG, o la información puede obtenerse de TCGA. El sistema de análisis puede ser cualquier sistema informático genérico con un procesador de ordenador y un medio de almacenamiento legible por ordenador con instrucciones para ejecutar el procesador de ordenador para realizar cualquiera o todas las operaciones descritas en esta presente descripción.

El sistema de análisis puede seleccionar regiones genómicas diana basadas en patrones de metilación de fragmentos de ácido nucleico. Un enfoque considera la capacidad de distinción por pares entre pares de resultados para regiones (o más específicamente para sitios CpG dentro de regiones). Otro enfoque considera la capacidad de distinción de las

regiones (o más específicamente para los sitios CpG dentro de las regiones) cuando se considera cada resultado contra los resultados restantes. A partir de las regiones genómicas diana seleccionadas con alta potencia de distinción, el sistema de análisis puede diseñar sondas para dirigirse a fragmentos de las regiones genómicas seleccionadas. El sistema de análisis puede generar tamaños variables del panel de ensayo de trastorno hematológico, por ejemplo, donde un panel de ensayo de trastorno hematológico de pequeño tamaño incluye sondas dirigidas a las regiones genómicas más informativas, un panel de ensayo de trastorno hematológico de tamaño medio incluye sondas del panel de ensayo de trastorno hematológico de pequeño tamaño y sondas adicionales dirigidas a un segundo nivel de regiones genómicas informativas, y un panel de ensayo de trastorno hematológico de gran tamaño incluye sondas de los paneles de ensayo de trastorno hematológico de tamaño pequeño y de tamaño mediano junto con incluso más sondas dirigidas a un tercer nivel de regiones genómicas informativas. Con los datos obtenidos tales paneles de ensayo de trastorno hematológico (p. ej., el estado de metilación en ácidos nucleicos derivados de los paneles de ensayo de trastorno hematológico), el sistema de análisis puede entrenar clasificadores con diversas técnicas de clasificación para predecir la probabilidad de una muestra de tener un resultado o estado particular, p. ej., trastorno hematológico, otro trastorno, otra enfermedad, etc.

En algunas disposiciones, el panel de ensayo de HD comprende al menos 500 pares de sondas, en donde cada par de al menos 500 pares comprende dos sondas configuradas para superponerse entre sí mediante una secuencia superpuesta, en donde la secuencia superpuesta comprende al menos 30 nucleótidos, y en donde cada sonda está configurada para hibridarse a una molécula de ADN convertido (p. ej., un ADNIc) correspondiente a una o más regiones genómicas. En algunas disposiciones, cada una de las regiones genómicas comprende al menos cinco sitios de metilación, y en donde los al menos cinco sitios de metilación tienen un patrón de metilación anormal en muestras de HD o un patrón de metilación diferente entre muestras de un HD diferente. Por ejemplo, en una disposición, los al menos cinco sitios de metilación se metilan diferencialmente entre las muestras de HD y las sin HD, entre diferentes tipos de HD, entre las muestras de CHIP y otras muestras de HD, entre el cáncer de sangre y el cáncer sólido, entre diferentes tipos de tejido canceroso de origen (TOO) o entre muestras de diferentes estadios de HD. En algunas disposiciones, cada par de sondas comprende una primera sonda y una segunda sonda, en donde la segunda sonda difiere de la primera sonda. La segunda sonda puede solaparse con la primera sonda mediante una secuencia superpuesta que es al menos 30, al menos 40, al menos 50 o al menos 60 nucleótidos de longitud.

Las regiones genómicas diana se pueden seleccionar de una cualquiera de las listas 1-8 (**TABLA 1**). En algunas disposiciones, el panel de ensayo de HD comprende una pluralidad de sondas, en donde cada una de la pluralidad de sondas está configurada para hibridarse con una molécula de ADNIc convertido correspondiente a una o más de las regiones genómicas en una cualquiera de las listas 1-8. En algunas disposiciones, la pluralidad de oligonucleótidos cebo diferentes se configura para hibridarse con moléculas de ADN derivadas de al menos el 20 % de las regiones genómicas diana de una cualquiera de las listas 1-8. En algunas disposiciones, la pluralidad de oligonucleótidos cebo diferentes se configura para hibridarse con moléculas de ADN derivadas de al menos el 30 %, el 40 %, el 50 %, el 60 %, el 70 % o el 80 % de las regiones genómicas diana de una cualquiera de las listas 1-8. Por ejemplo, la pluralidad de diferentes oligonucleótidos cebo puede configurarse para hibridarse con moléculas de ADN derivadas de al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 % o el 80 % de las regiones genómicas diana de las listas 2-4, o de al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 % o el 80 % de las regiones genómicas diana de las listas 5-7.

Las regiones genómicas diana pueden seleccionarse de la lista 1. Las regiones genómicas diana pueden seleccionarse de la lista 2. En algunas disposiciones, un método para detectar la neoplasia linfoide comprende evaluar el estado de metilación para secuenciar las lecturas derivadas de las regiones genómicas diana de la lista 2. Las regiones genómicas diana pueden seleccionarse de la lista 3. En algunas disposiciones, un método para detectar el mieloma múltiple comprende evaluar el estado de metilación para secuenciar las lecturas derivadas de las regiones genómicas diana de la lista 3. Las regiones genómicas diana pueden seleccionarse de la lista 4. En algunas disposiciones, un método para detectar la neoplasia mieloide comprende evaluar el estado de metilación para secuenciar las lecturas derivadas de las regiones genómicas diana de la lista 4. Las regiones genómicas diana pueden seleccionarse de la lista 5. En algunas disposiciones, un método para detectar la neoplasia linfoide comprende evaluar el estado de metilación para secuenciar las lecturas derivadas de las regiones genómicas diana de la lista 5. Las regiones genómicas diana pueden seleccionarse de la lista 6. En algunas disposiciones, un método para detectar el mieloma múltiple comprende evaluar el estado de metilación para secuenciar las lecturas derivadas de las regiones genómicas diana de la lista 6. Las regiones genómicas diana pueden seleccionarse de la lista 7. En algunas disposiciones, un método para detectar la neoplasia mieloide comprende evaluar el estado de metilación para secuenciar las lecturas derivadas de las regiones genómicas diana de la lista 7. Las regiones genómicas diana pueden seleccionarse de la lista 8. En algunas disposiciones, un método para detectar la neoplasia mieloide comprende evaluar el estado de metilación para secuenciar las lecturas derivadas de las regiones genómicas diana de la lista 8. En algunas disposiciones, las regiones genómicas pueden seleccionarse entre dos o más, tres o más, cuatro o más, cinco o más, seis o más, de las listas 1-8.

Dado que las sondas están configuradas para hibridarse con una molécula de ADN o ADNIc convertido correspondiente a, o derivada de, una o más regiones genómicas, las sondas pueden tener una secuencia diferente de la región genómica diana. Por ejemplo, un ADN que contiene un sitio CpG no metilado se convertirá para incluir UpG en lugar de CpG porque las citosinas no metiladas se convierten en uracilos mediante una reacción de conversión

(p. ej., tratamiento con bisulfito). Como resultado, una sonda se configura para hibridarse con una secuencia que incluye UpG en lugar de un CpG no metilado existente natural. Por consiguiente, un sitio complementario en la sonda al sitio de no metilación puede comprender CpA en lugar de CpG, y algunas sondas dirigidas a un sitio hipometilado donde todos los sitios de metilación no son metilados pueden no tener bases de guanina (G). En algunas disposiciones, al menos el 3 %, el 5 %, el 10 %, el 15 % o el 20 % de las sondas no comprenden secuencias CpG.

El panel de ensayo de HD puede usarse para detectar la presencia o ausencia de HD en general y/o proporcionar una clasificación de HD, como un tipo de HD o un estadio de HD. En algunas disposiciones, el panel de ensayo de HD puede usarse para proporcionar una clasificación del cáncer, como el tipo de cáncer, el estadio del cáncer, como ausencia del cáncer, estadio I del cáncer, estadio II del cáncer, estadio III del cáncer o estadio IV del cáncer. El panel puede incluir sondas dirigidas a regiones genómicas derivadas de ácidos nucleicos metiladas diferencialmente entre muestras de HD y sin HD, entre diferentes tipos de HD, entre muestras de CHIP y otras muestras de HD, entre diferentes tipos de tejido canceroso de origen (TOO) o entre muestras de diferentes estadios de HD. Por ejemplo, en algunas disposiciones, un panel de ensayo de HD está diseñado para enriquecer los ácidos nucleicos derivados de regiones genómicas metiladas diferencialmente basándose en los datos de secuenciación con bisulfito generados a partir del ADNlc de individuos con y sin HD.

Cada sonda, par de sondas o conjunto de sondas puede diseñarse para dirigirse a fragmentos de ácido nucleico correspondientes a o derivados de una o más regiones genómicas diana. Las regiones genómicas diana se seleccionan basándose en varios criterios diseñados para aumentar el enriquecimiento selectivo de fragmentos de ácidos nucleicos informativos mientras se reducen el ruido y las uniones no específicas.

En un ejemplo, un panel puede incluir sondas que pueden hibridar selectivamente (es decir, unirse a) y opcionalmente enriquecer fragmentos de ADNlc que están diferencialmente metilados en muestras de HD. En este caso, la secuencia de los fragmentos enriquecidos puede proporcionar información relevante para la detección de HD. Además, las sondas están diseñadas para dirigirse a regiones genómicas que se determina que tienen un patrón de metilación anormal en muestras de HD, o en muestras de un tipo específico de HD. En una disposición, las sondas se diseñan para dirigirse a regiones genómicas determinadas para ser hipermetiladas o hipometiladas en ciertos HD o tejido canceroso de orígenes para proporcionar selectividad y especificidad adicionales de la detección. En algunas disposiciones, un panel comprende sondas dirigidas a fragmentos hipometilados. En algunas disposiciones, un panel comprende sondas dirigidas a fragmentos hipermetilados. En algunas disposiciones, un panel comprende tanto un primer conjunto de sondas que se dirigen a fragmentos hipermetilados como un segundo conjunto de sondas dirigidas a fragmentos hipometilados. (**Figura 1C**) En algunas disposiciones, la relación entre el primer conjunto de sondas dirigidas a fragmentos hipermetilados y el segundo conjunto de sondas dirigidas a fragmentos hipometilados (relación HiperHipo) oscila entre 0,4 y 2, entre 0,5 y 1,8, entre 0,5 y 1,6, entre 1,4 y 1,6, entre 1,2 y 1,4, entre 1 y 1,2, entre 0,8 y 1, entre 0,6 y 0,8 o entre 0,4 y 0,6. En la presente memoria se proporcionan en detalle métodos para identificar regiones genómicas (es decir, regiones genómicas que dan lugar a moléculas de ADN metiladas diferencialmente o moléculas de ADN metiladas de forma anómala) entre muestras de HD y sin HD, entre diferentes tipos de HD, entre muestras CHIP y otras de HD, entre diferentes tipos de tejido canceroso de origen (TOO) o entre muestras de diferentes estadios de HD (p. ej., ausencia de cáncer, estadio I del cáncer, estadio II del cáncer, estadio III del cáncer, estadio IV del cáncer) y se proporcionan en detalle en la presente memoria métodos para identificar moléculas o fragmentos de ADN anómalamente metilados que se identifican como indicativos de HD.

En un segundo ejemplo, las regiones genómicas pueden seleccionarse cuando las regiones genómicas dan lugar a moléculas de ADN anormalmente metiladas en muestras de HD o muestras con tipos conocidos de HD (p. ej., CHIP, cáncer de la sangre). Por ejemplo, como se describe en la presente memoria, puede usarse un modelo de Markov entrenado en un conjunto de muestras sin HD para identificar regiones genómicas que dan lugar a moléculas de ADN anormalmente metiladas (es decir, moléculas de ADN que tienen un patrón de metilación por debajo de un umbral de valor de p).

Cada una de las sondas puede dirigirse a una región genómica que comprende al menos 30 pb, 35 pb, 40 pb, 45 pb, 50 pb, 60 pb, 70 pb, 80 pb, 90 pb, 100 pb o más. En algunas disposiciones, las regiones genómicas pueden seleccionarse para tener menos de 30, 25, 20, 15, 12, 10, 8 o 6 sitios de metilación.

Las regiones genómicas pueden seleccionarse cuando al menos el 80, el 85, el 90, el 92, el 95 o el 98 % de los al menos cinco sitios de metilación (p. ej., CpG) dentro de la región están metilados o no metilados en muestras sin HD o HD, muestras de un tipo específico de HD (p. ej., muestras de CHIP o muestras de cáncer de un tejido de origen [TOO]) o muestras de un estadio específico de HD.

Las regiones genómicas pueden filtrarse adicionalmente para seleccionar solo aquellas que probablemente sean informativas basándose en sus patrones de metilación, por ejemplo, sitios CpG que están metilados diferencialmente entre muestras HD o sin HD (p. ej., anormalmente metilados o no metilados en HD vs. sin HD), entre tipos de HD diferentes, entre muestras CHIP y otras muestras de HD o entre muestras de diferentes estadios de HD. Para la selección, el cálculo se puede realizar con respecto a cada CpG o una pluralidad de sitios CpG. Por ejemplo, se determina un primer recuento que es el número de muestras que contienen HD (recuento_HD) que incluyen un fragmento que se superpone con ese CpG, y se determina un segundo recuento que es el número de muestras totales

que contienen fragmentos que se superponen con ese sitio CpG (total). Las regiones genómicas pueden seleccionarse basándose en criterios correlacionados positivamente con el número de muestras que contienen HD (recuento_HD) que incluyen un fragmento indicativo de HD superpuesto a ese sitio CpG, e inversamente correlacionados con el número total de muestras que contienen fragmentos indicativos de HD superpuestos a ese sitio CpG (total). En una realización, se cuenta el número de muestras de sin HD ($n_{no\text{-}HD}$) y el número de muestras de HD (n_{HD}) que tienen un fragmento que se superpone a un sitio CpG. A continuación, se estima la probabilidad de que una muestra sea HD, por ejemplo, como $(n_{HD} + 1) / (n_{HD} + n_{no\text{-}HD} + 2)$.

Los sitios CpG puntuados por esta métrica se clasifican y se añaden suavemente a un panel hasta que el presupuesto del tamaño del panel se agota. El proceso de selección de regiones genómicas indicativas de HD se detalla adicionalmente en la presente memoria.

Se pueden seleccionar diferentes regiones diana dependiendo de si el ensayo pretende ser un ensayo pan-HD o un ensayo de HD único, o qué tipo de flexibilidad que se desee. Un panel para detectar un tipo específico de HD puede diseñarse usando un proceso similar. En esta disposición, para cada tipo de HD, y para cada sitio CpG, la ganancia de información se calcula para determinar si incluir una sonda que se dirige a ese sitio CpG. La ganancia de información se puede calcular para muestras con un HD dado en comparación con todas las demás muestras. Por ejemplo, considerar dos variables aleatorias, “AF” y “CT”. “AF” es una variable binaria que indica si hay un fragmento anormal que se superpone a un sitio CpG particular en una muestra particular (sí o no). La “CT” es una variable aleatoria binaria que indica si el HD es de un tipo particular (p. ej., CHIP, leucemia, neoplasias linfoides (p. ej., linfoma), mieloma múltiple y neoplasia mieloide). Se puede calcular la información mutua con respecto a “CT” dada “AF”. Es decir, cuántos bits de información sobre el tipo de HD (p. ej., CHIP vs. cáncer de la sangre) se obtienen si se sabe si hay un fragmento anómalo que se superpone a un sitio CpG particular. Esto puede usarse para clasificar CpG basándose en cómo son específicos de CHIP. Este procedimiento se repite para una pluralidad de tipos de HD. Si una región particular está metilada diferencialmente solo en CHIP (y no cáncer de la sangre), CpG en esa región tenderían a tener altas ganancias de información para CHIP. Para cada tipo de HD, los sitios CpG se clasifican por esta métrica de ganancia de información y después se añaden con avidez a un panel hasta que el presupuesto de tamaño para ese tipo de HD se agota.

Se puede realizar una filtración adicional para seleccionar sondas con alta especificidad para el enriquecimiento (es decir, alta eficiencia de unión) de ácidos nucleicos derivados de regiones genómicas específicas. Las sondas pueden filtrarse para reducir la unión inespecífica (o unión fuera de la diana) a ácidos nucleicos derivados de regiones genómicas no dirigidas. Por ejemplo, las sondas pueden filtrarse para seleccionar solo aquellas sondas que tengan menos de un umbral establecido de acontecimientos de unión fuera de la diana. En una disposición, las sondas pueden alinearse con un genoma de referencia (p. ej., un genoma de referencia humana) para seleccionar sondas que se alinean a menos de un umbral establecido de regiones a través del genoma. Por ejemplo, se pueden seleccionar sondas que se alinean a menos de 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9 u 8 regiones fuera de la diana a través del genoma de referencia. En otros casos, la filtración se realiza para eliminar las regiones genómicas cuando la secuencia de las regiones genómicas diana aparece más de 5, 10, 15, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34 o 35 veces en un genoma. Puede realizarse un filtrado adicional para seleccionar regiones genómicas diana cuando una secuencia de sonda, o un conjunto de secuencias de sonda que son homólogas en el 90 %, el 91 %, el 92 %, el 93 %, el 94 %, el 95 %, el 96 %, el 97 %, el 98 % o el 99 % a las regiones genómicas diana, aparecen menos de 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9 u 8 veces en un genoma de referencia, o para eliminar regiones genómicas diana cuando la secuencia de la sonda, o un conjunto de secuencias de la sonda diseñadas para enriquecer la región genómica diana son homólogas en el 90 %, el 91 %, el 92 %, el 93 %, el 94 %, el 95 %, el 96 %, el 97 %, el 98 % o el 99 % a las regiones genómicas diana, aparecen más de 5, 10, 15, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34 o 35 veces en un genoma de referencia. Esto es para excluir sondas repetitivas que pueden tirar de fragmentos fuera de diana, que no son deseables y pueden afectar la eficiencia del ensayo.

En algunas disposiciones, se demostró que una superposición de fragmento-sonda de al menos 45 pb es efectiva para lograr una cantidad no despreciable de extracción (aunque un experto en la técnica apreciaría este número puede muy) como se proporciona en el ejemplo 1. En algunas disposiciones, más de un 10 % de tasa de emparejamiento erróneo entre la sonda y las secuencias de fragmentos en la región de solapamiento es suficiente para interrumpir en gran medida la unión y, por tanto, la eficiencia de la extracción. Por tanto, las secuencias que pueden alinearse con la sonda a lo largo de al menos 45 pb con al menos una tasa de coincidencia del 90 % pueden ser candidatas para la extracción fuera de la diana. Por tanto, en una realización, se puntúa el número de tales regiones. Las mejores sondas tienen una puntuación de 1, lo que significa que coinciden en un solo lugar (la región diana prevista). Las sondas con una puntuación intermedia (digamos, menos de 5 o 10) pueden aceptarse en algunos casos y, en algunos casos, se descartan cualquier sonda por encima de una puntuación particular. Pueden usarse otros valores de corte para muestras específicas.

Una vez que las sondas hibridan y capturan fragmentos de ADN correspondientes a una región genómica diana o derivados de ella, los intermedios sonda-fragmento de ADN hibridado se extraen (o aíslan), y el ADN diana se amplifica y su estado de metilación se determina, por ejemplo, mediante secuenciación o hibridación con una micromatriz, etc. La lectura de la secuencia proporciona información relevante para la detección de HD. Para este fin, un panel se

diseña para incluir una pluralidad de sondas que pueden capturar fragmentos que juntos pueden proporcionar información relevante para la detección del HD. En algunas disposiciones, un panel incluye al menos 500, 1.000, 2.000, 2.500, 5.000, 6.000, 7.500, 10.000, 15.000, 20.000, 25.000, 30.000, 35.000, 40.000, 50.000, 60.000, 70.000, 80.000, 90.000, 100.000, 110.000 o 120.000 pares de sondas. En otras disposiciones, un panel incluye al menos 1.000, 2.000, 5.000, 10.000, 50.000, 100.000, 150.000, 200.000, 250.000, 300.000, 400.000, 500.000, 550.000, 600.000, 700.000 u 800.000 sondas. La pluralidad de sondas juntas puede comprender al menos 10.000, 20.000, 30.000, 40.000, 50.000, 60.000, 70.000, 80.000, 90.000, 100.000, 120.000, 140.000, 160.000, 180.000, 200.000, 240.000, 260.000, 280.000, 300.000, 320.000, 400.000, 450.000, 500.000, 550.000, 600.000, 650.000, 700.000, 750.000, 800.000, 850.000, 900.000, 1 millón, 1,5 millones, 2 millones, 2,5 millones, 3 millones, 3,5 millones, 4 millones, 4,5 millones o 5 millones de nucleótidos.

Las regiones genómicas diana seleccionadas pueden ubicarse en varias posiciones en un genoma, que incluyen, pero no se limitan a, exones, intrones, regiones intergénicas y otras partes. En algunas disposiciones, pueden añadirse sondas dirigidas a regiones genómicas no humanas, tales como las que dirigen las regiones genómicas virales.

En algunos casos, pueden usarse cebadores para amplificar específicamente diana/biomarcadores de interés (por ejemplo, por PCR), enriqueciendo así la muestra para dianas/biomarcadores deseados (opcionalmente sin captura de hibridación). Por ejemplo, pueden prepararse cebadores directos e inversos para cada región genómica de interés y usarse para amplificar fragmentos que corresponden a o se derivan de la región genómica deseada. Por tanto, aunque la presente descripción presta particular atención a los paneles de ensayo de HD y conjuntos de cebos para captura de hibridación, la descripción es lo suficientemente amplia como para abarcar otros métodos para el enriquecimiento del ADN libre de células. No se reivindican otros métodos de este tipo. Por consiguiente, un experto en la materia, con el beneficio de esta descripción, reconocerá que los métodos análogos a los descritos en la presente memoria en relación con la captura de hibridación pueden lograrse alternativamente reemplazando la captura de hibridación con alguna otra estrategia de enriquecimiento, tal como la amplificación por PCR de fragmentos de ADN libres de células que se corresponden con regiones genómicas de interés. En algunas disposiciones, la captura de sonda de candado con bisulfito se usa para enriquecer regiones de interés, tal como se describe en Zhang y col. (documento US 2016/0340740). En algunas disposiciones, se usan métodos adicionales o alternativos para el enriquecimiento (p. ej., enriquecimiento no dirigido) tal como secuenciación de bisulfito de representación reducida, secuenciación de enzimas de restricción de metilación, secuenciación de inmunoprecipitación de ADN de metilación, secuenciación de proteínas de dominio de unión a metil-CpG, secuenciación de captura de ADN de metilo o PCR de microgotas.

Sondas

El panel de ensayo de HD proporcionado en la presente memoria es un panel que incluye un conjunto de sondas de hibridación (también denominadas en la presente memoria “sondas”) diseñadas para, durante el enriquecimiento, dirigir y eliminar fragmentos de ácido nucleico de interés para el ensayo. En algunas disposiciones, las sondas están diseñadas para hibridarse y enriquecer moléculas de ADN o ADNlc de muestras de HD que han sido tratadas para convertir citosinas (C) no metiladas en uracilos (U). En otras disposiciones, las sondas están diseñadas para hibridarse y enriquecer moléculas de ADN o ADNlc de un tipo específico de HD que han sido tratadas para convertir citosinas (C) no metiladas en uracilos (U). Las sondas pueden diseñarse para hibridarse (o hibridarse) con una cadena diana (complementaria) de ADN o ARN. La cadena diana puede ser la cadena “positiva” (p. ej., la cadena transcrita en el ARNm y posteriormente traducida en una proteína) o la cadena “negativa” complementaria. En una disposición particular, un panel de ensayo de HD puede incluir conjuntos de dos sondas, una sonda dirigida a la cadena positiva y la otra sonda dirigida a la cadena negativa de una región genómica diana.

Para cada región genómica diana, se pueden diseñar cuatro posibles secuencias de sonda. Las moléculas de ADN correspondientes a, o derivadas de, cada región diana es bicatenaria, y como tal, una sonda o conjunto de sonda puede dirigirse a la cadena “positiva” o hacia adelante o su complemento inverso (la cadena “negativa”). Además, en algunas disposiciones, las sondas o conjuntos de sondas se diseñan para enriquecer moléculas de ADN o fragmentos que se han tratado para convertir citosinas (C) no metiladas en uracilos (U). Dado que las sondas o conjuntos de sondas están diseñados para enriquecer moléculas de ADN correspondientes a las regiones diana o derivadas de ellas tras la conversión, la secuencia de la sonda puede diseñarse para enriquecer moléculas de ADN de fragmentos en los que las C no metiladas se han convertido en U (usando A en lugar de G en sitios que son citosinas no metiladas en moléculas o fragmentos de ADN correspondientes a la región diana o derivadas de ella). En una disposición, las sondas están diseñadas para unirse, o hibridarse con, moléculas de ADN o fragmentos de regiones genómicas que se sabe que contienen patrones de metilación específicos de HD (p. ej., moléculas de ADN hipermetiladas o hipometiladas), enriqueciendo (o detectando) así las moléculas o fragmentos de ADN específicos del HD. Las regiones genómicas específicas, o los patrones de metilación específicos de HD, pueden ser ventajosas, lo que permite enriquecer específicamente moléculas de ADN o fragmentos identificados como informativos para pan-HD o un tipo específico de HD, y por lo tanto, disminuir las necesidades de detección y los costos (p. ej., reducir los costos de secuenciación). En otras disposiciones, se pueden diseñar dos secuencias sonda por una región genómica diana (una para cada cadena de ADN).

En todavía otros casos, las sondas están diseñadas para enriquecer todas las moléculas de ADN o fragmentos correspondientes a, o derivadas de, una región dirigida (es decir, independientemente del estado de cadena o de

metilación). Esto podría deberse a que el estado de metilación del HD no está altamente metilado o no metilado, o porque las sondas están diseñadas para dirigirse a mutaciones pequeñas u otras variaciones en lugar de cambios de metilación, con estas otras variaciones de modo similar indicativas de la presencia o ausencia de un HD o la presencia o ausencia de un HD específico. En ese caso, las cuatro secuencias de sonda posibles pueden incluirse por una región genómica diana.

En algunas disposiciones, algunas sondas están diseñadas para detectar variantes y mutaciones indicativas de la presencia o ausencia de un HD o la presencia o ausencia de un HD específico. Tales sondas están diseñadas para enriquecer moléculas o fragmentos de ADN correspondientes a o derivados de una región diana que puede incluir tales variantes o mutaciones. Algunas de las variantes o mutaciones pueden ser uno o más loci que se sabe que están asociados o se sospecha que están asociados con el CHIP u otro HD. Algunas de las variantes o mutaciones pueden ser uno o más loci identificados como indicativos de CHIP u otro HD mediante los métodos descritos en 4.5.

Las sondas pueden variar en longitud de 10 s, 100 s, 200 s o 300 s de pares de bases. Las sondas pueden comprender al menos 50, 75, 100 o 120 nucleótidos. Las sondas pueden comprender menos de 300, 250, 200 o 150 nucleótidos. En una disposición, las sondas comprenden 100-150 nucleótidos. En una disposición, las sondas comprenden 120 nucleótidos.

En algunas disposiciones, las sondas están diseñadas en una forma "de 2 títulos" para cubrir porciones superpuestas de una región diana. Cada sonda se solapa opcionalmente en cobertura al menos parcialmente con otra sonda en la biblioteca. En tales disposiciones, el panel contiene múltiples pares de sondas, con cada sonda en un par que se superpone al otro en al menos 25, 30, 35, 40, 45, 50, 60, 70, 75 o 100 nucleótidos. En algunas disposiciones, la secuencia de superposición puede diseñarse para ser complementaria a una región genómica diana (o a partir de ADNlc derivado de la misma) o para ser complementaria a una secuencia con homología con una región diana o ADNlc. Por tanto, en algunas disposiciones, al menos dos sondas son complementarias a la misma secuencia dentro de una región genómica diana, y un fragmento de nucleótido correspondiente o derivado de la región genómica diana puede unirse y extraerse por al menos una de las sondas. Son posibles otros niveles de titulación, tales como tres títulos, cuatro títulos, etc., en donde cada nucleótido en una región diana puede unirse a más de dos sondas.

En una disposición, cada base en una región genómica diana se superpone con exactamente dos sondas, como se ilustra en la figura 1A. Un solo par de sondas es suficiente para extraer una región genómica si la superposición entre las dos sondas es más larga que la región genómica diana y se extiende más allá de ambos extremos de la región genómica diana. En algunos casos, incluso las regiones diana relativamente pequeñas pueden dirigirse con tres sondas (véase la figura 1A). Un conjunto de sondas que comprende tres o más sondas se usa opcionalmente para capturar una región genómica más grande (véase la figura 1B). En algunas disposiciones, los subconjuntos de sondas se extenderán colectivamente a través de una región genómica completa (p. ej., puede ser complementaria a fragmentos no convertidos o convertidos de la región genómica). Un conjunto de sonda de mosaico comprende opcionalmente sondas que incluyen colectivamente al menos dos sondas que se solapan cada nucleótido en la región genómica. Esto se hace para asegurar que los ADNlc que comprenden una pequeña porción de una región genómica diana en un extremo tendrán una superposición sustancial que se extiende en la región genómica adyacente no dirigida con al menos una sonda, para proporcionar una captura eficiente.

Por ejemplo, puede garantizarse que un fragmento de ADNlc de 100 pb que comprende una región genómica diana de 30 nt tenga al menos 65 pb de superposición con al menos una de las sondas solapantes. Son posibles otros niveles de titulación. Por ejemplo, para aumentar el tamaño objetivo y añadir más sondas en un panel, las sondas pueden diseñarse para expandir una región diana de 30 pb en al menos 70 pb, 65 pb, 60 pb, 55 pb o 50 pb. Para capturar cualquier fragmento que se superponga a la región diana (aunque sólo sea 1 pb), las sondas pueden diseñarse para que se extiendan más allá de los extremos de la región diana a ambos lados.

Las sondas están diseñadas para analizar el estado de metilación de regiones genómicas diana (p. ej., del ser humano o de otro organismo) que se sospecha que están correlacionadas con la presencia o ausencia de HD en general, la presencia o ausencia de determinados tipos de HD, el estadio del HD, o la presencia o ausencia de otros tipos de enfermedades (p. ej., otros tipos de cáncer como un cáncer sólido).

Además, las sondas están diseñadas para hibridarse eficazmente a (o unirse a) y opcionalmente extraer fragmentos de ADNlc que contengan una región genómica diana. En algunas disposiciones, las sondas están diseñadas para cubrir partes solapadas de una región diana, de modo tal que cada sonda tiene una cobertura "titulada" tal que cada sonda se solapa en cobertura al menos parcialmente con otra sonda de la biblioteca. En tales disposiciones, el panel contiene múltiples pares de sondas, donde cada par comprende al menos dos sondas superpuestas entre sí por una secuencia superpuesta de al menos 25, 30, 35, 40, 45, 50, 60, 70, 75 o 100 nucleótidos. En algunas disposiciones, la secuencia de superposición puede diseñarse para tener una homología de secuencia con o para ser complementaria a una región genómica diana (o una versión convertida de la misma), por tanto, un fragmento de nucleótido derivado de o correspondiente a la región genómica diana puede unirse y extraerse opcionalmente por al menos una de las sondas.

En una disposición, la región genómica diana más pequeña es de 30 pb. Cuando se añade una nueva región diana al panel (basándose en la selección de voraz como se describió anteriormente), la nueva región diana de 30 pb se puede centrar en un sitio CpG específico de interés. Luego, se comprueba si cada borde de esta nueva diana está lo suficientemente cerca como para otros objetivos de manera que puedan fusionarse. Esto se basa en un parámetro de "distancia de fusión" que puede ser de 200 pb por defecto pero puede ajustarse. Esto permite que las regiones objetivo cercanas pero distintas se enriquezcan con sondas superpuestas. Dependiendo de si existen dianas suficientemente cercanas a la izquierda o a la derecha de la nueva diana, la nueva diana puede fusionarse sin nada (aumentando el número de dianas de panel en uno), fusionado con solo una diana a la izquierda o a la derecha (no cambiando el número de dianas de panel), o fusionado con dianas existentes tanto a la izquierda como a la derecha (reduciendo el número de dianas de panel en uno).

Métodos de selección de regiones genómicas diana basándose en el estado de metilación

En otra disposición (no reivindicada), se proporcionan métodos para seleccionar regiones genómicas diana para detectar HD y/o un tipo o estadio específico de HD. Las regiones genómicas dirigidas pueden usarse para diseñar y fabricar sondas para un panel de ensayo de HD. El estado de metilación de las moléculas de ADN o ADNlc correspondientes a, o derivados de, las regiones genómicas diana pueden seleccionarse usando el panel de ensayo de HD. Los métodos alternativos, por ejemplo, por WGBS u otros métodos conocidos en la técnica, también pueden implementarse para detectar el estado de metilación de moléculas de ADN o fragmentos correspondientes a, o derivados de, las regiones genómicas diana.

Procesamiento de muestra

La **Figura 7A** es un diagrama de flujo de un proceso 100 para procesar una muestra de ácido nucleico y generar vectores de estado de metilación para fragmentos de ADN, según una disposición. Si bien la presente descripción presta especial atención a los enfoques basados en la secuenciación para detectar ácidos nucleicos y determinar el estado de metilación, la descripción es lo suficientemente amplia como para abarcar otros métodos para determinar el estado de metilación de las secuencias de ácidos nucleicos (tales como los enfoques de secuenciación sensibles a la metilación descritos en WO 2014/043763). Como se describe en la Figura 7A, el método incluye, pero no se limita a, las siguientes etapas. Por ejemplo, cualquier etapa del método puede comprender una subetapa de cuantificación para el control de calidad u otros procedimientos de ensayo de laboratorio conocidos por un experto en la técnica.

En la etapa 105, una muestra de ácido nucleico (ADN o ARN) se extrae de un sujeto. En la presente descripción, el ADN y el ARN pueden usarse indistintamente a menos que se indique lo contrario. Es decir, las disposiciones descritas en la presente memoria pueden ser aplicables tanto a tipos de ADN como a ARN de secuencias de ácido nucleico. Sin embargo, los ejemplos descritos en la presente memoria pueden centrarse en ADN para fines de claridad y explicación. La muestra puede ser cualquier subconjunto del genoma humano, incluyendo el genoma completo. La muestra puede incluir sangre, plasma, suero, orina, heces, saliva, otros tipos de fluidos corporales o cualquier combinación de los mismos. En algunas disposiciones, los métodos para extraer una muestra de sangre (p. ej., jeringa o punción de dedo) pueden ser menos invasivos que los procedimientos para obtener una biopsia de tejido, que puede requerir cirugía. La muestra extraída puede comprender ADNlc y/o ADNct. Para individuos sanos, el cuerpo humano puede eliminar naturalmente el ADNlc y otros restos celulares. Si un sujeto tiene un cáncer o enfermedad, el ADNlc y/o ADNct en una muestra extraída pueden estar presentes a un nivel suficiente para detectar el trastorno hematológico.

En la etapa 110, los fragmentos de ADNlc se tratan para convertir citosinas no metiladas en uracilos. En una disposición, el método usa un tratamiento con bisulfito del ADN que convierte las citosinas no metiladas en uracilos sin convertir las citosinas metiladas. Por ejemplo, se usa un kit comercial tal como el kit EZ DNA Methylation™ - Gold, EZ DNA Methylation™ - Direct o un kit EZ DNA Methylation™ - Lightning (comercializado por Zymo Research Corp [Irvine, CA]) para la conversión con bisulfito. En otra disposición, la conversión de citosinas no metiladas en uracilos se logra usando de una reacción enzimática. Por ejemplo, la conversión puede usar un kit disponible comercialmente para la conversión de citosinas no metiladas en uracilos, tales como APOBEC-Seq (NEBiolabs, Ipswich, MA).

En la etapa 115, se prepara una biblioteca de secuenciación. En una primera etapa, se añade un adaptador de ADNmc al extremo 3'-OH de una molécula de ADNmc convertida con bisulfito usando una reacción de ligamiento de ADNmc. En una disposición, la reacción de ligamiento de ADNmc usa CircLigase II (Epicentre) para ligar el adaptador de ADNmc al extremo 3'-OH de una molécula de ADNmc convertida en bisulfito, en donde el extremo 5' del adaptador está fosforilado y el ADNmc convertido en bisulfito ha sido desfosforilado (es decir, el extremo 3' tiene un grupo hidroxilo). En otra disposición, la reacción de ligamiento de ADNmc usa la ligasa termoestable 5' AppDNA/RNA (disponible de New England BioLabs [Ipswich, MA]) para ligar el adaptador de ADNmc al extremo 3'-OH de una molécula de ADNmc convertida con bisulfito. En este ejemplo, el primer adaptador UMI se adenila en el extremo 5' y se bloquea en el extremo 3'. En otra disposición, la reacción de ligamiento de ADNmc usa una ligasa de ARN T4 (disponible de New England BioLabs) para ligar el adaptador de ADNmc al extremo 3'-OH de una molécula de ADNmc convertida con bisulfito. En una segunda etapa, se sintetiza un ADN de segunda cadena en una reacción de extensión. Por ejemplo, un cebador de extensión, que se hibrida con una secuencia de cebador incluida en el adaptador de ADNmc, se usa en una reacción de extensión de cebador para formar una molécula de ADN bicatenario convertida

con bisulfito. Opcionalmente, en una disposición, la reacción de extensión usa una enzima que es capaz de leer a través de residuos de uracilo en la cadena de molde convertida con bisulfito. Opcionalmente, en una tercera etapa, se añade un adaptador de ADNbc a la molécula de ADN bicatenario convertida con bisulfito. Finalmente, el ADN bicatenario convertido con bisulfito se amplifica para añadir adaptadores de secuenciación. Por ejemplo, la amplificación por PCR usando un cebador directo que incluye una secuencia P5 y un cebador inverso que incluye una secuencia P7 se usa para añadir las secuencias P5 y P7 al ADN convertido con bisulfito. Opcionalmente, durante la preparación de la biblioteca, pueden añadirse identificadores moleculares únicos (UMI) a las moléculas de ácido nucleico (por ejemplo, moléculas de ADN) mediante ligamiento de adaptador. Las UMI son secuencias cortas de ácido nucleico (por ejemplo, 4-10 pares de bases) que se añaden a los extremos de los fragmentos de ADN durante el ligamiento del adaptador. En algunas realizaciones, las UMI son pares de bases degenerados que sirven como una etiqueta única que puede usarse para identificar lecturas de secuencia que se originan en un fragmento de ADN específico. Durante la amplificación por PCR después de la ligamiento del adaptador, las UMI se replican junto con el fragmento de ADN unido, lo que proporciona una forma de identificar lecturas de secuencia que provienen del mismo fragmento original en el análisis posterior.

En la etapa 120, las secuencias de ADN diana pueden enriquecerse de la biblioteca. Esto se usa, por ejemplo, donde se realiza un ensayo de panel objetivo en las muestras. Durante el enriquecimiento, las sondas de hibridación (también denominadas en la presente memoria "sondas") se usan para dirigirse, y extraer, fragmentos de ácido nucleico informativos para la presencia o ausencia de HD (o enfermedad), estadio del HD o una clasificación de HD (p. ej., tipo de HD o tejido de origen). Para un flujo de trabajo determinado, las sondas pueden diseñarse para hibridar (o hibridarse) con una cadena diana (complementaria) de ADN o ARN. La cadena diana puede ser la cadena "positiva" (p. ej., la cadena transcrita en el ARNm y posteriormente traducida en una proteína) o la cadena "negativa" complementaria. Las sondas pueden variar en longitud de 10 s, 100 s o 1000 s de pares de bases. Además, las sondas pueden cubrir porciones superpuestas de una región diana.

Después de una etapa de hibridación 120, los fragmentos de ácido nucleico hibridados se capturan y también se pueden amplificar usando PCR (enriquecimiento 125). Por ejemplo, las secuencias diana pueden enriquecerse para obtener secuencias enriquecidas que se pueden secuenciar posteriormente. En general, cualquier método conocido en la técnica puede usarse para aislar, y enriquecer para, ácidos nucleicos diana hibridados por sonda. Por ejemplo, como es bien conocido en la técnica, puede añadirse un resto de biotina al extremo 5' de las sondas (es decir, biotiniladas) para facilitar el aislamiento de ácidos nucleicos diana hibridados con sondas usando una superficie recubierta con estreptavidina (p. ej., perlas recubiertas con estreptavidina).

En la etapa 130, se generan lecturas de secuencia a partir de las secuencias de ADN enriquecidas, por ejemplo, secuencias enriquecidas. Los datos de secuenciación pueden adquirirse a partir de las secuencias de ADN enriquecidas por medios conocidos en la técnica. Por ejemplo, el método puede incluir técnicas de secuenciación de próxima generación (NGS) que incluyen tecnología de síntesis (Illumina), pirosecuenciación (454 Life Sciences), tecnología de semiconductores de iones (secuenciación Ion Torrent), secuenciación en tiempo real de una sola molécula (Pacífico Biosciences), secuenciación por ligamiento (secuenciación SOLiD), secuenciación de nanoporos (Oxford Nanopore Technologies) o secuenciación de extremos emparejados. En algunas disposiciones, la secuenciación masivamente paralela se realiza mediante secuenciación por síntesis con terminadores de colorante reversibles. En otras disposiciones, como entenderá fácilmente un experto en la técnica, puede usarse cualquier medio conocido para detectar ácidos nucleicos y determinar el estado de metilación. Por ejemplo, las secuencias pueden detectarse y determinarse el estado de metilación usando una secuenciación compatible con la metilación conocida (véase, p. ej., WO 2014/043763), una micromatriz de ADN (p. ej., con sondas marcadas adheridas o conjugadas a una superficie sólida o un chip de matriz de ADN), etc.

En la etapa 140, se generan vectores de estado de metilación a partir de las lecturas de secuencia. Para hacerlo, una lectura de secuencia se alinea con un genoma de referencia. El genoma de referencia ayuda a proporcionar el contexto en cuanto a qué posición en un genoma humano se origina el fragmento ADNlc En un ejemplo simplificado, la lectura de secuencia se alinea de manera que los tres sitios CpG se correlacionan con los sitios CpG 23, 24 y 25 (identificadores de referencia arbitrarios usados por conveniencia de la descripción). Después de la alineación, hay información tanto en el estado de metilación de todos los sitios CpG en el fragmento de ADNlc como en qué posición en el genoma humano se correlaciona con los sitios CpG. Con el estado y ubicación de metilación, se puede generar un vector de estado de metilación para el fragmento ADNlc.

Generación de la estructura de datos

La Figura 3A es un diagrama de flujo que describe un proceso 300 para generar una estructura de datos para un grupo de control sano, según una disposición. Para crear una estructura de datos del grupo de control saludable, el sistema de análisis obtiene información relacionada con el estado de metilación de una pluralidad de sitios CpG en lecturas de secuencia derivadas de una pluralidad de moléculas de ADN o fragmentos de una pluralidad de sujetos sanos. El método proporcionado en la presente memoria para crear una estructura de datos del grupo de control saludable puede realizarse de modo similar para sujetos con HD, sujetos con un tipo específico de HD o sujetos con otro estadio de enfermedad conocido. Se genera un vector de estado de metilación para cada molécula o fragmento de ADN, por ejemplo, a través del proceso 100.

Con el vector de estado de metilación de cada fragmento, el sistema de análisis subdivide 310 el vector de estado de metilación en cadenas de sitios CpG. En una disposición, el sistema de análisis subdivide 310 el vector de estado de metilación de tal modo que las cadenas resultantes son todas menores que una longitud dada. Por ejemplo, un vector de estado de metilación de longitud 11 puede subdividirse en cadenas de longitud menor o igual a 3 daría como resultado 9 cadenas de longitud 3, 10 cadenas de longitud 2 y 11 cadenas de longitud 1. En otro ejemplo, un vector de estado de metilación de longitud 7 se subdivide en cadenas de longitud menor o igual a 4 daría como resultado 4 cadenas de longitud 4, 5 cadenas de longitud 3, 6 cadenas de longitud 2 y 7 cadenas de longitud 1. Si un vector de estado de metilación es más corto que o la misma longitud que la longitud de cadena especificada, entonces el vector de estado de metilación puede convertirse en una sola cadena que contiene todos los sitios CpG del vector.

El sistema de análisis consiste en 320 las cadenas contando, para cada sitio CpG posible y la posibilidad de estados de metilación en el vector, el número de cadenas presentes en el grupo de control que tiene el sitio CpG especificado como el primer sitio CpG en la cadena y que tiene esa posibilidad de estados de metilación. Por ejemplo, en un sitio CpG determinado y considerando las longitudes de cadena de 3, hay 2 ^ 3 u 8 configuraciones de cadena posibles. En ese sitio CpG dado, para cada una de las 8 configuraciones de cadena posibles, el sistema de análisis calcula 320 cuántas ocurrencias de cada posibilidad de vector del estado de metilación aparecen en el grupo de control. Continuando con este ejemplo, esto puede implicar calcular las siguientes cantidades: $< M_x, M_{x+1}, M_{x+2} >$, $< M_x, M_{x+1}, U_{x+2} >$, ..., $< U_x, U_{x+1}, U_{x+2} >$ para cada sitio CpG inicial x en el genoma de referencia. El sistema de análisis crea 330 la estructura de datos que almacena los recuentos calculados para cada sitio de CpG inicial y la posibilidad de cadena.

Existen varias ventajas para establecer un límite superior en la longitud de la cuerda. En primer lugar, dependiendo de la longitud máxima para una cadena, el tamaño de la estructura de datos creada por el sistema de análisis puede aumentar drásticamente el tamaño. Por ejemplo, una longitud máxima de cadena de 4 significa que cada sitio CpG tiene al menos 2^4 números para calcular cadenas de longitud 4. Aumentar la longitud máxima de cadena a 5 significa que cada sitio CpG tiene 2^4 o 16 números adicionales para calcular, lo que duplica los números para calcular (y requiere memoria de computadora) en comparación con la longitud de cadena anterior. La reducción del tamaño de la cadena ayuda a mantener la creación y el rendimiento de la estructura de datos (p. ej., usarlos para un acceso posterior, como se describe a continuación), en términos de cálculo y almacenamiento. En segundo lugar, una consideración estadística para limitar la longitud máxima de las cadenas es evitar sobreajustar los modelos corriente abajo que usan los recuentos de cadenas. Si las cadenas largas de sitios CpG no tienen, biológicamente, un efecto fuerte en el resultado (p. ej., las predicciones de anomalía que predicen la presencia de HD), calcular las probabilidades basándose en grandes cadenas de sitios CpG puede resultar problemático, ya que requiere una cantidad significativa de datos que puede no estar disponible y, por lo tanto, sería demasiado escasa para que un modelo funcione adecuadamente. Por ejemplo, calcular una probabilidad de anomalía/HD condicionado en los 100 sitios CpG anteriores requeriría recuentos de cadenas en la estructura de datos de longitud 100, idealmente alguna coincidencia exactamente con los 100 estados de metilación anteriores. Si solo hay recuentos dispersos de cadenas de longitud 100, habrá datos insuficientes para determinar si una cadena dada de longitud de 100 en una muestra de prueba es anómala o no.

Validación de la estructura de datos

Una vez que se ha creado la estructura de datos, el sistema de análisis puede buscar validar 340 la estructura de datos y/o cualquier modelo aguas abajo que haga uso de la estructura de datos. Un tipo de validación verifica la coherencia dentro de la estructura de datos del grupo de control. Por ejemplo, si hay sujetos, muestras y/o fragmentos atípicos dentro de un grupo de control, entonces el sistema de análisis puede realizar varios cálculos para determinar si se debe excluir algún fragmento de una de esas categorías. En un ejemplo representativo, el grupo de control sano puede contener una muestra que no está diagnosticada pero que tiene un HD tal que la muestra contiene fragmentos anómalamente metilados. Este primer tipo de validación asegura que las posibles muestras de HD se eliminen del grupo de control sano para no afectar la pureza del grupo de control.

Un segundo tipo de validación verifica el modelo probabilístico usado para calcular los valores de p con los recuentos de la propia estructura de datos (es decir, del grupo de control sano). Un proceso para el cálculo de valor de p se describe a continuación junto con la **figura 5.** Una vez que el sistema de análisis genera un valor de p para los vectores de estado de metilación en el grupo de validación, el sistema de análisis construye una función de densidad acumulativa (CDF) con los valores de p. Con el CDF, el sistema de análisis puede realizar diversos cálculos en la CDF para validar la estructura de datos del grupo de control. Una prueba utiliza el hecho de que la CDF debe estar idealmente en o por debajo de una función de identidad, de manera que $CDF(x) \leq x$. En el contrario, por encima de la función de identidad revela alguna deficiencia dentro del modelo probabilístico usado para la estructura de datos del grupo de control. Por ejemplo, si 1/100 de fragmentos tienen una puntuación de valor de p de 1/1000 que significa CDF (1/1000) = 1/100>1/1000, entonces el segundo tipo de validación falla indicando un problema con el modelo probabilístico.

Un tercer tipo de validación usa un conjunto saludable de muestras de validación distintas de las que se usan para crear la estructura de datos, lo que prueba si la estructura de datos se ha creado correctamente y si el modelo funciona. Un proceso de ejemplo para llevar a cabo este tipo de validación se describe a continuación junto con la **figura 3B**. El

tercer tipo de validación puede cuantificar cómo el grupo de control sano generaliza la distribución de muestras sanas. Si falla el tercer tipo de validación, entonces el grupo de control sano no generaliza bien en la distribución saludable.

Un cuarto tipo de pruebas de validación con muestras de un grupo de validación no saludable. El sistema de análisis calcula valores de p y construye el CDF para el grupo de validación no saludable. Con un grupo de validación no saludable, el sistema de análisis espera ver el CDF(x)>x para al menos algunas muestras o, indicadas de manera diferente, la inversa de lo que se esperaba en el segundo tipo de validación y el tercer tipo de validación con el grupo de control sano y el grupo de validación saludable. Si falla el cuarto tipo de validación, esto es indicativo de que el modelo no identifica adecuadamente la anomalía que estaba diseñado para identificar.

La **Figura 3B** es un diagrama de flujo que describe la etapa adicional 340 de validar la estructura de datos para el grupo de control de la **Figura 3A**, según una realización. En esta disposición de la etapa 340 de validar la estructura de datos, el sistema de análisis realiza el cuarto tipo de prueba de validación como se ha descrito anteriormente que utiliza un grupo de validación con una composición supuestamente similar de sujetos, muestras y/o fragmentos como el grupo de control. Por ejemplo, si el sistema de análisis seleccionó sujetos sanos sin HD para el grupo de control, entonces el sistema de análisis también usa sujetos sanos sin HD en el grupo de validación.

El sistema de análisis toma el grupo de validación y genera 100 un conjunto de vectores del estado de metilación como se describe en la **Figura 3A** . El sistema de análisis realiza un cálculo del valor de p para cada vector de estado de metilación del grupo de validación. El proceso de cálculo de valor de p se describirá adicionalmente junto con las **Figuras 4 y 5** . Para cada posibilidad del vector de estado de metilación, el sistema de análisis calcula una probabilidad de la estructura de datos del grupo de control. Una vez que se calculan las probabilidades para las posibilidades de los vectores de estado de metilación, el sistema de análisis calcula 350 una puntuación de valor de p para ese vector de estado de metilación basándose en las probabilidades calculadas. La puntuación de valor de p representa una expectativa de encontrar ese vector de estado de metilación específico y otros vectores de estado de metilación posibles que tienen probabilidades incluso menores en el grupo de control. Una puntuación de valor de p baja, por lo tanto, corresponde generalmente a un vector de estado de metilación que es relativamente inesperado en comparación con otros vectores de estado de metilación dentro del grupo de control, donde una puntuación de valor de p alta corresponde generalmente a un vector de estado de metilación que es relativamente más esperado en comparación con otros vectores de estado de metilación encontrados en el grupo de control. Una vez que el sistema de análisis genera una puntuación de valor de p para los vectores de estado de metilación en el grupo de validación, el sistema de análisis construye 360 una función de densidad acumulativa (CDF) con las puntuaciones de valor de p del grupo de validación. El sistema de análisis valida la coherencia 370 del CDF como se ha descrito anteriormente en el cuarto tipo de pruebas de validación.

Fragmentos finamente metilados

Los fragmentos anómalamente metilados que tienen patrones de metilación anormales en muestras de HD, sujeto con un tipo específico de HD o sujetos con otro estadio de la enfermedad conocido, se seleccionan como regiones genómicas diana, según una realización como se describe en la **Figura 4** . Los procesos ilustrativos de fragmentos aleatorios seleccionados de forma anómala 440 se ilustran visualmente en la **Figura 5** y se describe adicionalmente a continuación la descripción de la **Figura 4**. En el proceso 400, el sistema de análisis genera 100 Vectores de estado de metilación a partir de fragmentos de ADNlc de la muestra. El sistema de análisis maneja cada vector de estado de metilación de la siguiente manera.

Para un vector de estado de metilación dado, el sistema de análisis enumera 410 todas las posibilidades de los vectores de estado de metilación que tienen el mismo sitio CpG inicial y la misma longitud (es decir, conjunto de sitios CpG) en el vector de estado de metilación. Como cada estado de metilación puede estar metilado o no metilado, solo hay dos estados posibles en cada sitio CpG y, por tanto, el recuento de posibilidades distintas de vectores de estado de metilación depende de una potencia de 2, de tal modo que un vector de estado de metilación de longitud n estaría asociado con 2n posibilidades de los vectores de estado de metilación.

El sistema de análisis calcula 420 la probabilidad de observar cada posibilidad del vector de estado de metilación para el sitio CpG inicial identificado/longitud del vector de estado de metilación accediendo a la estructura de datos del grupo de control saludable. En una disposición, calcular la probabilidad de observar una posibilidad dada usa una probabilidad de cadena de Markov para modelar el cálculo de probabilidad conjunta que se describirá con mayor detalle con respecto a la **Figura 5** a continuación. En otras disposiciones, se usan métodos de cálculo distintos de las probabilidades de cadena de Markov para determinar la probabilidad de observar cada posibilidad del vector de estado de metilación.

El sistema de análisis calcula 430 una puntuación de valor de p para el vector de estado de metilación usando las probabilidades calculadas para cada posibilidad. En una disposición, esto incluye identificar la probabilidad calculada correspondiente a la posibilidad que coincida con el vector de estado de metilación en cuestión. Específicamente, esto es la posibilidad que tiene el mismo conjunto de sitios CpG, o similar al mismo sitio de CpG inicial y longitud como vector de estado de metilación. El sistema de análisis suma las probabilidades calculadas de cualquier posibilidades que tenga probabilidades inferiores o iguales a la probabilidad identificada para generar la puntuación de valor de p.

Este valor de p representa la probabilidad de observar el vector de estado de metilación del fragmento u otros vectores de estado de metilación incluso menos probables en el grupo de control sano. Una puntuación de valor de p baja, por lo tanto, corresponde generalmente a un vector de estado de metilación que es raro en un sujeto sano, y que hace que el fragmento se marque anormalmente metilado, en relación con el grupo de control sano. Se espera que una puntuación de valor de p alta generalmente se refiera a un vector de estado de metilación, en un sentido relativo, en un sujeto sano. Si el grupo de control sano es un grupo sin HD, por ejemplo, un valor de p bajo indica que el fragmento está anormalmente metilado en relación con el grupo sin HD y, por tanto, posiblemente indicativo de la presencia de HD en el sujeto de prueba.

Como anteriormente, el sistema de análisis calcula las puntuaciones de valor de p para cada uno de una pluralidad de vectores de estado de metilación, cada uno de los cuales representa un fragmento de ADNlc en la muestra de prueba. Para identificar cuál de los fragmentos están anormalmente metilados, el sistema de análisis puede filtrar 440 el conjunto de vectores de estado de metilación basándose en sus puntuaciones de valor de p. En una realización, el filtrado se realiza comparando las puntuaciones de los valores de p con respecto a un umbral y manteniendo solo aquellos fragmentos por debajo del umbral. Esta puntuación de valor de p umbral podría ser del orden de 0,1, 0,01, 0,001, 0,0001 o similar.

Cálculo de la puntuación de valor de p

La Figura **5** es una ilustración 500 de un cálculo de puntuación de valor de p de ejemplo, según una disposición. Para calcular una puntuación de valor de p dada un vector 505 de estado de metilación de prueba, el sistema de analítica toma ese vector 505 de estado de metilación de prueba y enumera 410 posibilidades de vectores de estado de metilación. En un ejemplo ilustrativo, el vector de estado de metilación de la prueba 505 es < M23, M24, M25, U26 >. Como la longitud del vector de estado de metilación de prueba 505 es 4, hay 2^4 posibilidades de vectores de estado de metilación que abarcan los sitios CpG 23 - 26. En un ejemplo genérico, el número de posibilidades de vectores de estado de metilación es 2^n, donde n es la longitud del vector de estado de metilación de prueba o, alternativamente, la longitud de la ventana deslizante (descrita más adelante).

El sistema de análisis calcula 420 probabilidades 515 para las posibilidades enumeradas de vectores de estado de metilación. Como la metilación depende condicionalmente del estado de metilación de los sitios CpG cercanos, una forma de calcular la probabilidad de observar una posibilidad de vector de estado de metilación dada es usar el modelo de cadena de Markov. Generalmente, un vector de estado de metilación tal como <S_1, S_2, …, S_n>, donde S indica el estado de metilación ya sea metilado (indicado como M), no metilado (indicado como U) o indeterminado (indicado como I), tiene una probabilidad conjunta que puede expandirse usando la regla de cadena de probabilidades como:

$$P(< S_1, S_2, \dots, S_n >) = P(S_n | S_1, \dots, S_{n-1} >) * P(S_{n-1} | S_1, \dots, S_{n-2} >) * \dots * P(S_2 | S_1) * P(S_1) \quad (1)$$

El modelo de cadena de Markov se puede usar para hacer el cálculo de las probabilidades condicionales de cada posibilidad más eficiente. En una realización, el sistema de análisis selecciona un orden de cadena de Markov *k* que corresponde a cuántos sitios CpG anteriores en el vector (o ventana) considerar en el cálculo de probabilidad condicional, de tal modo que la probabilidad condicional se modela como P(S_n| S_1, …, S_{n-1}) ~ P(S_n| S_{n-k-2}, …, S_{n-1}).

Para calcular cada probabilidad modelada por Markov para una posibilidad de vector de estado de metilación, el sistema de análisis accede a la estructura de datos del grupo de control, específicamente los recuentos de varias cadenas de sitios y estados CpG. Para calcular P(M_n | S_{n-k-2}, …, S_{n-1}), el sistema de análisis toma una relación del recuento almacenado del número de cadenas de estructura de datos que coincide < S_{n-k-2}, …, S_{n-1}, M_n > dividido entre la suma del recuento almacenado del número de cadenas de la estructura de datos que coincide < S_{n-k-2}, …, S_{n-1}, M_n > y < S_{n-k-2}, …, S_{n-1}, U_n>. Por tanto, P(M_n | S_{n-k-2}, …, S_{n-1}), se calcula una relación que tiene la forma:

$$\frac{\text{Núm.de} < S_{n-k-2},\dots,S_{n-1},M_n >}{\text{Núm.de} < S_{n-k-2},\dots,S_{n-1},M_n > + \#\, of < S_{n-k-2},\dots,S_{n-1},U_n >} \quad (2)$$

El cálculo puede implementar adicionalmente un suavizado de los recuentos aplicando una distribución previa. En una disposición, la distribución previa es un valor uniforme antes del suavizado de Laplace. Como ejemplo de esto, se añade una constante al numerador y otra constante (por ejemplo, dos veces la constante en el numerador) se añade al denominador de la ecuación anterior. En otras realizaciones, se usa una técnica algorítmica tal como suavizado de Knesser-Ney.

En la ilustración, las fórmulas indicadas anteriormente se aplican al vector de estado de metilación de prueba 505 que cubre los sitios 23-26. Una vez que se completan las probabilidades calculadas 515, el sistema de análisis calcula 430 una puntuación de valor de p 525 que suma las probabilidades que son menores o iguales a la probabilidad de posibilidad de que el vector de estado de metilación coincida con el vector de estado de metilación de prueba 505.

En una disposición, la carga computacional de las probabilidades de cálculo y/o las puntuaciones de valor de p puede reducirse aún más al almacenar en caché al menos algunos cálculos. Por ejemplo, el sistema analítico puede

almacenar en caché cálculos de probabilidades de memoria transitoria o persistente para posibilidades de vectores de estado de metilación (o ventanas de los mismos). Si otros fragmentos tienen los mismos sitios CpG, el almacenamiento en caché de las probabilidades de posibilidad permite un cálculo eficiente de las puntuaciones de valor de p sin necesidad de volver a calcular las probabilidades de posibilidad subyacente. De manera equivalente, el sistema de análisis puede calcular puntuaciones de valor de p para cada una de las posibilidades de los vectores de estado de metilación asociados con un conjunto de sitios CpG del vector (o ventana del mismo). El sistema de análisis puede almacenar en caché las puntuaciones de valor de p para su uso en la determinación de las puntuaciones de valor de p de otros fragmentos que incluyen los mismos sitios CpG. Generalmente, las puntuaciones de valor de p de las posibilidades de los vectores de estado de metilación que tienen los mismos sitios CpG pueden usarse para determinar la puntuación de valor de p de una diferente de las posibilidades del mismo conjunto de sitios CpG.

Ventana deslizante

En una disposición, el sistema de análisis usa 435 una ventana deslizante para determinar las posibilidades de los vectores de estado de metilación y calcular los valores de p. En lugar de enumerar las posibilidades y calcular los valores de p para todos los vectores de estado de metilación, el sistema de análisis enumera las posibilidades y calcula los valores de p solo para una ventana de sitios CpG secuenciales, donde la ventana es más corta en longitud (de sitios CpG) que al menos algunos fragmentos (de lo contrario, la ventana no serviría). La longitud de la ventana puede ser estática, determinada por el usuario, dinámica o seleccionada de otro modo.

Al calcular los valores de p para un vector de estado de metilación mayor que la ventana, la ventana identifica el conjunto secuencial de sitios CpG del vector dentro de la ventana comenzando desde el primer sitio CpG en el vector. El sistema analítico calcula una puntuación de valor de p para la ventana que incluye el primer sitio CpG. El sistema de análisis "desliza" la ventana al segundo sitio CpG en el vector, y calcula otra puntuación de valor de p para la segunda ventana. Por lo tanto, para un tamaño de ventana *l* y longitud del vector de metilación *m,* cada vector de estado de metilación generará *m-l+1* puntuaciones de valor de p. Después de completar los cálculos de valor de p para cada parte del vector, la puntuación de valor de p más baja de todas las ventanas deslizantes se toma como la puntuación de valor de p global para el vector de estado de metilación. En otra disposición, el sistema de análisis agrega las puntuaciones de valor de p para los vectores de estado de metilación para generar una puntuación de valor de p general.

Usando la ventana deslizante ayuda a reducir el número de posibilidades enumeradas de vectores de estado de metilación y sus cálculos de probabilidad correspondientes que de otro modo necesitaría realizarse. Los cálculos de probabilidad de ejemplo se muestran en la **figura 5**, pero generalmente el número de posibilidades de vectores de estado de metilación aumenta exponencialmente en un factor de 2 con el tamaño del vector de estado de metilación. Para dar un ejemplo realista, es posible que los fragmentos tengan hacia arriba de 54 sitios CpG. En lugar de probabilidades informáticas para 2A54 (~1,8x10^16) posibilidades para generar un solo valor de p, el sistema de análisis puede usar en cambio una ventana de tamaño 5 (por ejemplo) que da como resultado los cálculos de 50 p para cada una de las 50 Ventanas del vector de estado de metilación para ese fragmento. Cada uno de los 50 cálculos enumera 2A5 (32) posibilidades de vectores de estado de metilación, cuyos resultados totales dan como resultado 50 x 2^5 (1,6 x 10^3) cálculos de probabilidad. Esto da como resultado que se realice una gran reducción de los cálculos, sin impacto significativo para la identificación precisa de fragmentos anómalos. Esta etapa adicional también se puede aplicar cuando se valida 340 el grupo de control con los vectores de estado de metilación del grupo de validación.

Identificar fragmentos indicativos del HD

El sistema de análisis identifica 450 fragmentos de ADN indicativos del HD del conjunto filtrado de fragmentos anormalmente metilados.

Fragmentos hipometilados e hipermetilados

Según un primer método, el sistema de análisis puede identificar fragmentos de ADN que se consideran hipometilados o hipermetilados como fragmentos indicativos del HD del conjunto filtrado de fragmentos anormalmente metilados. Los fragmentos hipometilados y hipermetilados pueden definirse como fragmentos de una cierta longitud de sitios CpG (por ejemplo, más de 3, 4, 5, 6, 7, 8, 9, 10, etc.) con un alto porcentaje de sitios CpG metilados (por ejemplo, más del 80 %, el 85 %, el 90 % o el 95 %, o cualquier otro porcentaje dentro del intervalo del 50 % -100 %) o un alto porcentaje de sitios CpG no metilados (por ejemplo, más del 80 %, 85 %, 90 % o 95 %, o cualquier otro porcentaje dentro del intervalo del 50 % -100 %).

Modelos probabilísticos

Según un segundo método, el sistema de análisis identifica fragmentos indicativos del HD que utilizan modelos probabilísticos de patrones de metilación ajustados a cada tipo de HD y tipo sin HD. El sistema de análisis calcula relaciones de probabilidad logarítmica para una muestra usando fragmentos de ADN en las regiones genómicas considerando los diversos tipos de HD con los modelos probabilísticos ajustados para cada tipo de HD y tipo sin HD. El sistema de análisis puede determinar que un fragmento de ADN es indicativo del HD basándose en si al menos una

de las relaciones de probabilidad logarítmica consideradas contra los diversos tipos de HD está por encima de un valor umbral.

En una disposición de partición del genoma, el sistema de análisis divide el genoma en regiones por múltiples etapas. En una primera etapa, el sistema de análisis separa el genoma en bloques de sitios CpG. Cada bloque se define cuando hay una separación entre dos sitios CpG adyacentes que excede algún umbral, por ejemplo, más de 200 pb, 300 pb, 400 pb, 500 pb, 600 pb, 700 pb, 800 pb, 900 pb o 1.000 pb. A partir de cada bloque, el sistema de análisis subdivide en una segunda etapa cada bloque en regiones de una cierta longitud, por ejemplo, 500 pb, 600 pb, 700 pb, 800 pb, 900 pb, 1.000 pb, 1.100 pb, 1.200 pb, 1.300 pb, 1.400 pb o 1.500 pb. El sistema de análisis puede superponerse adicionalmente a regiones adyacentes por un porcentaje de la longitud, por ejemplo, el 10 %, el 20 %, el 30 %, el 40 %, el 50 % o el 60 %.

El sistema de análisis analiza lecturas de secuencia derivadas de fragmentos de ADN para cada región. El sistema de análisis puede procesar muestras de tejido y/o ADNlc de alta señal. Las muestras de ADNlc de alta señal pueden determinarse mediante un modelo de clasificación binaria, por estadio de HD o por otra métrica.

Para cada tipo de HD y sin HD, el sistema de análisis se ajusta a un modelo probabilístico separado para fragmentos. En un ejemplo, cada modelo probabilístico es un modelo de mezcla que comprende una combinación de una pluralidad de componentes de la mezcla con cada componente de la mezcla que es un modelo de sitios independientes donde se supone que la metilación en cada sitio CpG es independiente de los estados de metilación en otros sitios CpG.

En disposiciones alternativas, el cálculo se realiza con respecto a cada sitio CpG. Específicamente, se determina un primer recuento que es el número de muestras HD (recuento_HD) que incluyen un fragmento de ADN anómalamente metilado que se superpone a CpG, y se determina un segundo recuento que es el número total de muestras que contienen fragmentos que se superponen a CpG (total) en el conjunto. Las regiones genómicas pueden seleccionarse en base a los números, por ejemplo, en base a criterios correlacionados positivamente con el número de muestras HD (recuento_HD) que incluyen un fragmento de ADN que se superpone a CpG, e inversamente se correlaciona con el número total de muestras que contienen fragmentos que se superponen a CpG (total) en el conjunto.

El sistema de análisis puede calcular además las relaciones de probabilidad logarítmica (“R”) para un fragmento que indica una probabilidad de que el fragmento sea indicativo de HD considerando los diversos tipos de HD con los modelos probabilísticos ajustados para cada tipo de HD y tipo sin HD. Las dos probabilidades pueden tomarse de modelos probabilísticos ajustados para cada uno de los tipos de HD y el tipo sin HD, los modelos probabilísticos definidos para calcular una probabilidad de observar un patrón de metilación en un fragmento dado cada uno de los tipos de HD y el tipo sin HD. Por ejemplo, los modelos probabilísticos pueden definirse ajustados para cada uno de los tipos de HD y el tipo sin HD.

Selección de regiones genómicas indicativas del HD

El sistema de análisis identifica 460 regiones genómicas indicativas de HD. Para identificar estas regiones informativas, el sistema de análisis calcula una ganancia de información para cada región genómica o más específicamente cada sitio CpG que describe una capacidad para distinguir entre diversos resultados.

Un método para identificar regiones genómicas capaces de distinguir entre el tipo de HD y el tipo sin HD utiliza un modelo de clasificación entrenado que puede aplicarse en el conjunto de moléculas de ADN anómalamente metiladas o fragmentos correspondientes o derivados de un grupo HD o sin HD. El modelo de clasificación entrenado puede entrenarse para identificar cualquier condición de interés que pueda identificarse a partir de los vectores de estado de metilación.

En una disposición, el modelo de clasificación entrenado es un clasificador binario entrenado basándose en estados de metilación para fragmentos de ADNlc o secuencias genómicas obtenidas de una cohorte de sujeto con HD o un tipo específico de HD, y una cohorte de sujeto sano sin HD, y después se usa para clasificar una probabilidad de que un sujeto de prueba tenga HD, o no tenga HD, basándose en vectores de estado de metilación anómalo. En otras disposiciones, se pueden entrenar diferentes clasificadores usando cohortes de sujetos que se sabe que tienen un HD particular (p. ej., CHIP, leucemia, etc.); o se sabe que tiene diferentes estadios de un HD particular (p. ej., CHIP, cáncer en estadio I, II, III o IV). En estas disposiciones, se pueden entrenar diferentes clasificadores usando lecturas de secuencias obtenidas de muestras enriquecidas para células tumorales de cohortes de sujetos que se sabe que padecen un cáncer de la sangre particular (p. ej., leucemia, neoplasias linfoides [p. ej., linfoma], mieloma múltiple y neoplasia mieloide, etc.). Cada capacidad de la región genómica para distinguir entre el tipo de HD y el tipo de sin HD en el modelo de clasificación se usa para clasificar las regiones genómicas de la más informativa al menos informativa en el rendimiento de clasificación. El sistema de análisis puede identificar regiones genómicas de la clasificación según la ganancia de información en la clasificación entre el tipo sin HD y el tipo de HD.

Calcular información de información de fragmentos hipometilados y hipermetilados indicativos del HD

Con fragmentos indicativos del HD, el sistema de análisis puede entrenar un clasificador según un proceso 600 ilustrado en la figura 6A, según una disposición. El proceso 600 accede a dos grupos de entrenamiento de muestras - un grupo sin HD y un grupo de HD - y obtiene 605 un conjunto sin HD de vectores de estado de metilación y un conjunto de HD de vectores de estado de metilación que comprenden fragmentos débilmente metilados, p. ej., a través de la etapa 440 del proceso 400.

El sistema de análisis determina 610, para cada vector de estado de metilación, si el vector de estado de metilación es indicativo de HD. Aquí, los fragmentos indicativos de HD pueden definirse como fragmentos hipermetilados o hipometilados determinados si al menos algún número de sitios CpG tiene un estado particular (metilado o no metilado, respectivamente) y/o tener un porcentaje umbral de sitios que son el estado particular (nuevamente, metilado o no metilado, respectivamente). En un ejemplo, los fragmentos de ADNlc se identifican como hipometilados o hipermetilados, respectivamente, si el fragmento se superpone al menos 5 sitios CpG, y al menos el 80 %, el 90 % o el 100 % de sus sitios CpG están metilados o al menos el 80 %, el 90 % o el 100 % no están metilados.

En una disposición alternativa, el sistema de análisis considera porciones del vector de estado de metilación y determina si la porción está hipometilada o hipermetilada, y puede distinguir esa porción que va a hipometilarse o hipermetilarse. Esta alternativa resuelve los vectores de estado de metilación ausentes que son grandes de tamaño pero contienen al menos una región de hipometilación densa o hipermetilación. Este proceso de definir hipometilación e hipermetilación puede aplicarse en la etapa 450 de la **figura 4.** En otra realización, los fragmentos indicativos del HD pueden definirse según las probabilidades emitidas desde modelos probabilísticos entrenados.

En una disposición, el proceso genera 620 una puntuación de hipometilación (P_{hipo}) y una puntuación de hipermetilación (P_{hiper}) por sitio CpG en el genoma. Para generar una puntuación en un sitio CpG dado, el clasificador toma cuatro recuentos en ese recuento de sitio CpG- (1) de los vectores (estado de metilación) del conjunto de HD marcado hipometilado que se superpone al sitio CpG; (2) recuento de vectores del conjunto de HD marcado hipermetilado que se solapan con el sitio CpG; (3) recuento de vectores del conjunto sin HD marcado hipometilado que se superpone con el sitio CpG; y (4) recuento de vectores del conjunto sin HD marcado hipermetilado que se superpone con el sitio CpG. Además, el proceso puede normalizar estos recuentos para cada grupo para tener en cuenta la varianza en el tamaño del grupo entre el grupo sin HD y el grupo HD. En disposiciones alternativas en donde se usan más generalmente fragmentos indicativos del HD, las puntuaciones pueden definirse más ampliamente como recuentos de fragmentos indicativos del HD en cada región genómica y/o sitio CpG.

En una disposición, para generar 620 la puntuación de hipometilación en un sitio CpG dado, el proceso toma una relación de (1) sobre (1) sumado con (3). De manera similar, la puntuación de hipermetilación se calcula tomando una relación de (2) en (2) y (4). Además, estas razones pueden calcularse con una técnica de suavizado adicional como se discutió anteriormente. La puntuación de hipometilación y la puntuación de hipermetilación se relacionan con una estimación de la probabilidad del HD dada la presencia de hipometilación o hipermetilación de fragmentos del conjunto de HD.

El sistema de análisis genera 630 una puntuación de hipometilación agregada y una puntuación de hipermetilación agregada para cada vector de estado de metilación anómalo. Las puntuaciones agregadas de hiper e hipometilación se determinan basándose en las puntuaciones de hipermetilación de los sitios CpG en el vector de estado de metilación. En una disposición, las puntuaciones agregadas de hiper e hipometilación se asignan como las puntuaciones más grandes de hipermetilación de la hipermetilación de los sitios en cada vector de estado, respectivamente. Sin embargo, en disposiciones alternativas, las puntuaciones agregadas podrían basarse en medias, medianas u otros cálculos que usan las puntuaciones de hiper/hipo metilación de los sitios en cada vector.

El sistema de análisis clasifica 640 todos los vectores del estado de metilación del sujeto por su puntuación de hipometilación agregada y por su puntuación de hipermetilación agregada, lo que da como resultado dos clasificaciones por sujeto. El proceso selecciona puntuaciones de hipometilación agregadas a partir de la clasificación de hipometilación y las puntuaciones de hipermetilación agregada de la clasificación de hipermetilación. Con las puntuaciones seleccionadas, el clasificador genera 650 un único vector de características para cada sujeto. En una disposición, las puntuaciones seleccionadas de cualquier clasificación se seleccionan con un orden fijo que es el mismo para cada vector de característica generado para cada sujeto en cada uno de los grupos de entrenamiento. Como un ejemplo, en una disposición, el clasificador selecciona el primer, el segundo, el cuarto y el octavo agregado de hipermetilación agregado, y de modo similar para cada puntuación de hipometilación agregada, de cada clasificación y escribe esas puntuaciones en el vector de características para ese sujeto.

El sistema de análisis entrena 660 un clasificador binario para distinguir vectores de características entre los grupos de entrenamiento HD y sin HD. Generalmente, puede usarse una cualquiera de una serie de técnicas de clasificación. En una disposición, el clasificador es un clasificador no lineal. En una disposición específica, el clasificador es un clasificador no lineal que utiliza una regresión logística de núcleo regularizado de L2 con un núcleo de función base radial (RBF) gaussiana.

Específicamente, en una disposición, el número de muestras sin HD o tipo(s) de HD diferentes (n_{otro}) y el número de muestras de HD o tipo(s) de HD(n_{HD}) que tiene un fragmento anómalamente metilado que se superpone a un sitio

CpG. A continuación, la probabilidad de que una muestra sea HD se estima mediante una puntuación ("S") que se correlaciona positivamente con n_{HD} e inversamente con n_{otro}. La puntuación se puede calcular usando la ecuación: $(n_{HD} + 1) / (n_{HD} + n_{otro} + 2)$ o $(n_{HD}) / (n_{HD} + n_{otro})$. El sistema de análisis calcula 670 una ganancia de información para cada tipo de HD y para cada región genómica o sitio CpG para determinar si la región genómica o el sitio CpG es indicativo de HD. La ganancia de información se calcula para las muestras de entrenamiento con un determinado tipo de HD en comparación con todas las demás muestras. Por ejemplo, se usan dos variables aleatorias 'fragmentos anómalos' ('AF') y 'tipo de HD' ('CT'). En una realización, AF es una variable binaria que indica si hay un fragmento anómalo que se superpone a un sitio CpG dado en una muestra dada según se determina para la puntuación de anomalía/vector de características anterior. El CT es una variable aleatoria que indica si el HD es de un tipo particular. El sistema de análisis calcula la información mutua con respecto a CT dada AF. Es decir, cuántos bits de información sobre el tipo de HD se obtienen si se sabe si hay un fragmento anómalo que se superpone a un sitio CpG particular.

Para un tipo de HD dado, el sistema de análisis usa esta información para clasificar sitios CpG basándose en cómo son específicos de HD. Este procedimiento se repite para todos los tipos de HD en consideración. Si una región particular se ha metilado comúnmente de forma anómala en muestras de entrenamiento de un HD dado, pero no en muestras de entrenamiento de otros tipos de HD o en muestras de entrenamiento saludable, entonces los sitios CpG superpuestos por esos fragmentos anómalos tenderán a tener altas ganancias de información para el tipo de HD dado. Los sitios CpG clasificados para cada tipo de HD se añaden (seleccionan) con avidez a un conjunto seleccionado de sitios CpG en función de su clasificación para su uso en el clasificador de HD.

Calcular la ganancia de información por pares a partir de fragmentos indicativos de HD identificado de modelos probabilísticos

Con fragmentos indicativos del HD identificado según el método descrito en la presente memoria, la analítica puede identificar regiones genómicas según el proceso 680 en la figura 6B. El sistema de análisis define 690 un vector de característica para cada muestra, para cada región, para cada tipo de HD por un recuento de fragmentos de ADN que tienen una relación de probabilidad logarítmica calculada que el fragmento es indicativo de HD por encima de una pluralidad de umbrales, en donde cada recuento es un valor en el vector de características. En una disposición, el sistema de análisis cuenta el número de fragmentos presente en una muestra en una región para cada tipo de HD con relaciones de probabilidad logarítmica por encima de uno o una pluralidad de posibles valores umbral. El sistema de análisis define un vector de características para cada muestra, por un recuento de fragmentos de ADN para cada región genómica para cada tipo de HD que proporciona una relación logarítmica de probabilidad calculada para el fragmento por encima de una pluralidad de umbrales, en donde cada recuento es un valor en el vector de características. El sistema de análisis usa los vectores de características definidos para calcular una puntuación informativa para cada región genómica que describe la capacidad de la región genómica para distinguir entre cada par de tipos de HD. Para cada par de tipos de HD, el sistema de análisis clasifica las regiones en base a las puntuaciones informativas. El sistema de análisis puede seleccionar regiones en base a la clasificación según las puntuaciones informativas.

El sistema de análisis calcula 695 una puntuación informativa para cada región que describe la capacidad de esa región para distinguir entre cada par de tipos de HD. Para cada par de tipos de HD distintos, el sistema de análisis puede especificar un tipo como un tipo positivo y el otro como un tipo negativo. En una disposición, la capacidad de una región para distinguir entre el tipo positivo y el tipo negativo se basa en información mutua, calculada usando la fracción estimada de muestras de ADNlc del tipo positivo y del tipo negativo para el que se esperaría que la característica sea distinta de cero en el ensayo final, es decir, al menos un fragmento de ese nivel que se secuenciaría en un ensayo de metilación dirigido. Esas fracciones se estiman usando las tasas observadas a las que se produce la característica en muestras de ADNlc sano y en muestras de ADNlc y/o tumorales de alta señal de cada tipo de HD. Por ejemplo, si una característica se produce con frecuencia en ADNlc sano, entonces también se estima que se produce con frecuencia en ADNlc de cualquier tipo de HD y probablemente daría como resultado una puntuación informativa baja. El sistema de análisis puede elegir un cierto número de regiones para cada par de tipos de HD de la clasificación, p. ej., 1024.

En disposiciones adicionales, el sistema de análisis identifica además regiones predominantemente hipermetiladas o hipometiladas a partir de la clasificación de regiones. El sistema de análisis puede cargar el conjunto de fragmentos en el tipo o tipos positivos para una región que se identificó como informativa. El sistema de análisis, de los fragmentos cargados, evalúa si los fragmentos cargados están predominantemente hipermetilados o hipometilados. Si los fragmentos cargados están predominantemente hipermetilados o hipometilados, el sistema de análisis puede seleccionar sondas correspondientes al patrón de metilación predominante. Si los fragmentos cargados no están predominantemente hipermetilados o hipometilados, el sistema de análisis puede usar una mezcla de sondas para dirigir tanto la hipermetilación como la hipometilación. El sistema de análisis puede identificar además un conjunto mínimo de sitios CpG que se superponen más que algún porcentaje de los fragmentos.

En otras disposiciones, el sistema de análisis, después de clasificar las regiones basadas en puntuaciones informativas, marca cada región con la clasificación informativa más baja en todos los pares de tipos de HD. Por ejemplo, si una región fuera la décima más informativa para distinguir entre mama y pulmón, y la quinta más informativa para distinguir entre mama y colorrectal, se le daría una etiqueta global de "5". El sistema de análisis puede diseñar

sondas que comienzan con las regiones con etiqueta más baja mientras se añaden regiones al panel, p. ej., hasta que se ha agotado el presupuesto del tamaño del panel.

Regiones genómicas fuera de la diana

En algunas disposiciones, las sondas que se dirigen a regiones genómicas seleccionadas se filtran adicionalmente 475 en base al número de regiones fuera de diana. Esto es para sondas de cribado que extraen demasiados fragmentos de ADNlc correspondientes a, o derivados de, regiones genómicas fuera de diana. La exclusión de sondas que tienen muchas regiones fuera de diana puede ser valiosa disminuyendo las tasas fuera de diana y aumentando la cobertura objetivo para una cantidad dada de secuenciación.

Una región genómica fuera de diana es una región genómica que tiene una homología suficiente con respecto a una región genómica diana, de manera que las moléculas de ADN o fragmentos derivados de regiones genómicas fuera de diana hibridan y arrastran por una sonda diseñada para hibridarse con una región genómica diana. Una región genómica fuera de diana puede ser una región genómica (o una secuencia convertida de esa misma región) que se alinea con una sonda a lo largo de al menos 35 pb, 40 pb, 45 pb, 50 pb, 60 pb, 70 pb o 80 pb con al menos el 80 %, el 85 %, el 90 %, el 95 % o el 97 % de tasa de coincidencia. En una disposición, una región genómica fuera de la diana es una región genómica (o una secuencia convertida de esa misma región) que se alinea con una sonda a lo largo de al menos 45 pb con al menos una tasa de coincidencia del 90 %. Pueden adoptarse diversos métodos conocidos en la técnica para cribar regiones genómicas fuera de diana.

La búsqueda rápida del genoma para encontrar todas las regiones genómicas fuera de la diana puede ser computacionalmente desafiante. En una disposición, una estrategia de siembra de k-mero (que puede permitir una o más faltas de coincidencia) se combina con la alineación local en las ubicaciones de las semillas. En este caso, se puede garantizar una búsqueda exhaustiva de buenas alineaciones en base a la longitud de k-mero, se permite el número de faltas de coincidencia y el número de aciertos de semilla de k-mero en una ubicación particular. Esto requiere la alineación local de programación dinámica en un gran número de ubicaciones, por lo que este enfoque está altamente optimizado para usar las instrucciones de la CPU de vectores (por ejemplo, AVX2, AVX512) y también puede paralelizarse en muchos núcleos dentro de una máquina y también en muchas máquinas conectadas por una red. Un experto en la técnica reconocerá que las modificaciones y variaciones de este enfoque pueden implementarse con el fin de identificar regiones genómicas fuera de diana.

En algunas disposiciones, se excluyen (o filtran) del panel las sondas que tienen homología de secuencia con regiones genómicas fuera de diana, o moléculas de ADN correspondientes a, o derivadas de regiones genómicas fuera de diana que comprenden más de un número umbral. Por ejemplo, las sondas que tienen homología de secuencia con regiones genómicas fuera de diana, o moléculas de ADN correspondientes a, o derivadas de regiones genómicas fuera de diana de más de 30, más de 25, más de 20, más de 18, más de 15, más de 12, más de 10 o más de 5 regiones fuera de diana se excluyen.

En algunas disposiciones, las sondas se dividen en 2, 3, 4, 5, 6 o más grupos separados dependiendo de los números de regiones fuera de la diana. Por ejemplo, las sondas que tienen homología de secuencia sin regiones fuera de diana o moléculas de ADN correspondientes a, o derivadas de regiones fuera de diana se asignan al grupo de alta calidad, las sondas que tienen homología de secuencia con las regiones fuera de diana de 1-18 o las moléculas de ADN correspondientes a, o derivadas de, las regiones fuera de diana 1-18, se asignan al grupo de baja calidad, y las sondas que tienen homología de secuencia con más de 19 regiones fuera de diana o moléculas de ADN correspondientes, o derivadas de 19 regiones fuera de diana, se asignan a un grupo de calidad deficiente. Pueden usarse otros valores de corte para el agrupamiento.

En algunas disposiciones, se excluyen las sondas en el grupo de calidad más baja. En algunas disposiciones, se excluyen las sondas en grupos distintos del grupo de más alta calidad. En algunas disposiciones, se hacen paneles separados para las sondas en cada grupo. En algunas disposiciones, todas las sondas se colocan en el mismo panel, pero el análisis separado se realiza en base a los grupos asignados.

En algunas disposiciones, un panel comprende un mayor número de sondas de alta calidad que el número de sondas en grupos inferiores. En algunas disposiciones, un panel comprende un número menor de sondas de baja calidad que el número de sondas en otro grupo. En algunas disposiciones, más del 95 %, el 90 %, el 85 %, el 80 %, el 75 % o el 70 % de las sondas en un panel son sondas de alta calidad. En algunas disposiciones, menos del 35 %, el 30 %, el 20 %, el 10 %, el 5 %, el 4 %, el 3 %, el 2 % o el 1 % de las sondas en un panel son sondas de baja calidad. En algunas disposiciones, menos del 5 %, el 4 %, el 3 %, el 2 % o el 1 % de las sondas en un panel son sondas de baja calidad. En algunas disposiciones, no se incluyen sondas de baja calidad en un panel.

En algunas disposiciones, se excluyen las sondas que tienen por debajo del 50 %, por debajo del 40 %, por debajo del 30 %, por debajo del 20 %, por debajo del 10 % o por debajo del 5 %. En algunas modalidades, las sondas que tienen por encima del 30 %, por encima del 40 %, por encima del 50 %, por encima del 60 %, por encima del 70 %, por encima del 80 %, o por encima del 90 % se incluyen selectivamente en un panel.

Métodos de uso del panel de ensayo de HD

En otra disposición más, se proporcionan métodos para usar un panel de ensayo de HD. Los métodos pueden comprender las etapas de tratar moléculas o fragmentos de ADN para convertir citosinas no metiladas en uracilos (p. ej., usando tratamiento con bisulfito), aplicar un panel de HD (como se describe en la presente memoria) a las moléculas o fragmentos de ADN convertido, enriquecer un subconjunto de moléculas o fragmentos de ADN convertido que se hibridan (o unen) a las sondas en el panel y detectar la secuencia de ácido nucleico y determinar el estado de metilación de esta, por ejemplo, secuenciando los fragmentos de ADNlc enriquecidos. En algunas disposiciones, las lecturas de secuencia pueden compararse con un genoma de referencia (p. ej., un genoma de referencia humana), lo que permite la identificación de estados de metilación en una pluralidad de sitios CpG dentro de las moléculas o fragmentos de ADN y, por lo tanto, proporciona información relevante para la detección de un trastorno hematológico (HD). Si bien la presente descripción presta especial atención a los enfoques basados en la secuenciación para detectar ácidos nucleicos y determinar su estado de metilación (mediante lecturas de secuencia), la descripción es lo suficientemente amplia como para abarcar otros métodos para detectar ácidos nucleicos y determinar su estado de metilación (tal como otros enfoques de secuenciación compatible con la metilación (p. ej., como se describe en WO 2014/043763), micromatrices de ADN (p. ej., con sondas etiquetadas adheridas o conjugadas a una superficie sólida o un chip de matriz de ADN). etc.

Análisis de lecturas de secuencia

En algunas disposiciones, las lecturas de secuencia pueden alinearse con un genoma de referencia usando métodos conocidos en la técnica para determinar la información de posición de alineación. La información de la posición de alineación puede indicar una posición inicial y una posición final de una región en el genoma de referencia que corresponde a una base de nucleótidos inicial y una base de nucleótidos final de una secuencia determinada leída. La información de la posición de alineación también puede incluir una longitud de lectura de secuencia, que puede determinarse desde la posición inicial y la posición final. Una región en el genoma de referencia puede asociarse con un gen o un segmento de un gen.

En diversas disposiciones, una lectura de secuencia comprende un par de lectura indicado como R_1 y R_2. Por ejemplo, la primera lectura R_1 puede secuenciarse desde un primer extremo de un fragmento de ácido nucleico mientras que la segunda lectura R_2 puede secuenciarse desde el segundo extremo del fragmento de ácido nucleico. Por tanto, los pares de bases de nucleótidos de la primera lectura R_1 y la segunda lectura R_2 pueden alinearse de modo consistente (p. ej., en orientaciones opuestas) con bases de nucleótidos del genoma de referencia. La información de posición de alineación derivada del par de lectura R_1 y R_2 puede incluir una posición inicial en el genoma de referencia que corresponde a un extremo de una primera lectura (p. ej., R_1) y una posición final en el genoma de referencia que corresponde a un extremo de una segunda lectura (p. ej., R_2). En otras palabras, la posición inicial y la posición final en el genoma de referencia representan la ubicación probable dentro del genoma de referencia al que corresponde el fragmento de ácido nucleico. Se puede generar un archivo de salida que tiene un formato SAM (mapa de alineación de secuencia) o un formato BAM (mapa de alineación binaria) y se emite para un análisis adicional.

A partir de las lecturas de secuencia, la ubicación y el estado de metilación para cada sitio CpG se pueden determinar en función de la alineación con un genoma de referencia. Además, puede generarse un vector de estado de metilación para cada fragmento que especifica una ubicación del fragmento en el genoma de referencia (por ejemplo, como se especifica por la posición del primer sitio CpG en cada fragmento, U otra métrica similar), un número de sitios CpG en el fragmento y el estado de metilación de cada sitio CpG en el fragmento ya sea metilado (por ejemplo, indicado como M), no metilado (por ejemplo, indicado como U), o indeterminado (por ejemplo, indicado como I). Los vectores de estado de metilación pueden almacenarse en memoria informática temporal o persistente para su uso y procesamiento posterior. Además, pueden eliminarse lecturas duplicadas o vectores de estado de metilación duplicados de un solo sujeto. En una disposición adicional, se puede determinar que un cierto fragmento tiene uno o más sitios CpG que tienen un estado de metilación indeterminado. Dichos fragmentos pueden excluirse del procesamiento posterior o incluirse selectivamente donde el modelo de datos aguas abajo representa dichos estados de metilación indeterminados.

La **Figura 7B** es una ilustración del proceso 100 de la **Figura 7A** de secuenciación de un fragmento de ADNlc para obtener un vector de estado de metilación, según una disposición. Como ejemplo, el sistema de análisis toma un fragmento de ADNlc 112. En este ejemplo, el fragmento de ADNlc 112 contiene tres sitios CpG. Como se muestra, el primer y tercer sitios de CpG del fragmento de ADNlc 112 están metilados 114. Durante la etapa de tratamiento 120, el fragmento de ADNlc 112 se convierte para generar un fragmento de ADNlc convertido 122. Durante el tratamiento 120, el segundo sitio CpG que no estaba metilado tenía su citosina convertida en uracilo. Sin embargo, el primer y tercer sitios CpG no se convierten.

Después de la conversión, se prepara y secuenció una biblioteca 130 de secuenciación generando una lectura de secuencia 142. El sistema de análisis alinea 150 la lectura de secuencia 142 con un genoma de referencia 144. El genoma de referencia 144 proporciona el contexto en cuanto a qué posición en un genoma humano se origina el fragmento ADNlc En este ejemplo simplificado, el sistema de análisis alinea 150 la secuencia leída de manera que los tres sitios CpG se correlacionan con los sitios CpG 23, 24 y 25 (identificadores de referencia arbitrarios usados por

conveniencia de la descripción). Por tanto, el sistema de análisis genera información tanto en el estado de metilación de todos los sitios CpG en el fragmento de ADNlc 112 como en la posición en el genoma humano, el mapa de sitios CpG. Como se muestra, los sitios CpG en la secuencia de lectura 142 que estaban metilados se leen como citosinas. En este ejemplo, las citosina aparecen en la lectura de secuencia 142 solo en el primer y tercer sitios CpG que permite inferir que el primer y tercer sitios CpG en el fragmento de ADNlc original estaban metilados. El segundo sitio CpG se lee como una timina (U se convierte en T durante el proceso de secuenciación) y, por lo tanto, se puede inferir que el segundo sitio CpG no estaba metilado en el fragmento de ADNlc original. Con estos dos fragmentos de información, el estado y ubicación de metilación, el sistema de análisis genera 160 un vector de estado de metilación 152 para el fragmento ADNlc 112. En este ejemplo, el vector de estado de metilación 152 resultante es $< M_{23}, U_{24}, M_{25} >$, en donde M corresponde a un sitio CpG metilado, U corresponde a un sitio CpG no metilado, y los números de subíndice corresponden a las posiciones de cada sitio CpG en el genoma de referencia.

Detección de HD

Las lecturas de secuencia obtenidas por los métodos proporcionados en la presente memoria se procesan adicionalmente mediante algoritmos automatizados. Por ejemplo, el sistema de análisis se utiliza para recibir datos de secuenciación de un secuenciador y realizar diversos aspectos del procesamiento como se describe en la presente memoria. El sistema de análisis puede ser uno de un ordenador personal (PC), un ordenador de sobremesa, un ordenador portátil, un cuaderno, una tablet, un dispositivo móvil. Un dispositivo informático puede acoplarse comunicativamente al secuenciador a través de una combinación inalámbrica, cableada o de comunicación por cable. Generalmente, el dispositivo informático está configurado con un procesador y memoria que almacena instrucciones informáticas que, cuando son ejecutadas por el procesador, hacen que el procesador realice las etapas como se describe en el resto de este documento. Generalmente, la cantidad de datos genéticos y datos derivados de los mismos es lo suficientemente grande, y la cantidad de potencia computacional requerida es tan grande, por lo que debe ser imposible realizarse en papel o por la mente humana.

La interpretación clínica del estado de metilación de las regiones genómicas específicas es un proceso que incluye clasificar el efecto clínico de cada uno o una combinación del estado de metilación e informar los resultados de formas que son significativas para un profesional médico. La interpretación clínica puede basarse en la comparación de las lecturas de secuencia con base de datos específica para sujetos con HD o sin HD, y/o basarse en números y tipos de los fragmentos de ADNlc que tienen patrones de metilación específicos del HD identificados a partir de una muestra.

En algunas disposiciones, las regiones genómicas dirigidas se clasifican o clasifican basándose en su similitud para metilarse diferencialmente en muestras de HD, y los rangos o clasificaciones se usan en el proceso de interpretación. Los montones y clasificaciones pueden incluir (1) el tipo de efecto clínico, (2) la resistencia de evidencia del efecto y (3) el tamaño del efecto. Pueden adoptarse diversos métodos para el análisis clínico e interpretación de los datos del genoma para el análisis de las lecturas de secuencia. En algunas otras disposiciones, la interpretación clínica de los estados de metilación de tales regiones metiladas diferencialmente puede basarse en enfoques de aprendizaje automático que interpretan una muestra actual basándose en un método de clasificación o regresión que se entrenó usando los estados de metilación de tales regiones metiladas diferencialmente a partir de muestras de pacientes con HD y sin HD con estado de HD conocido, tipo de HD, estadio de HD, etc.

La información con significado clínico puede incluir la presencia o ausencia de HD en general, la presencia o ausencia de determinados tipos de HD, el estadio del HD, o la presencia o ausencia de otros tipos de enfermedades. En algunas disposiciones, la información se refiere a la presencia o ausencia de uno o más trastornos hematológicos, seleccionados del grupo que consiste en CHIP, leucemia, neoplasias linfoides (p. ej., linfoma), mieloma múltiple y neoplasia mieloide. En algunas disposiciones, la información se refiere a la presencia o ausencia de uno o más trastornos hematológicos, seleccionados del grupo que consiste en neoplasia linfoide, mieloma múltiple y neoplasia mieloide. En algunas disposiciones, las muestras no son cancerosas y son de sujetos que tienen expansión clonal de glóbulos blancos o ningún trastorno hematológico.

Clasificador de HD

Para entrenar un clasificador de tipo HD, el sistema de análisis obtiene una pluralidad de muestras de prueba, cada una de las cuales tiene un conjunto de fragmentos hipometilados e hipermetilados indicativos de HD, p. ej., identificados mediante la etapa 450 en el proceso 400, y una etiqueta del tipo de HD de la muestra de prueba. El sistema de análisis determina, para cada muestra de entrenamiento, un vector característico basado en el conjunto de fragmentos hipometilados e hipermetilados indicativos del HD. El sistema de análisis calcula una puntuación de anomalía para cada sitio CpG en las regiones genómicas diana. En una disposición, el sistema de análisis define la puntuación de anomalía para el vector de características como una puntuación binaria basándose en si hay un fragmento hipometilado o hipermetilado del conjunto que abarca el sitio CpG. Una vez que se determinan todas las puntuaciones de anomalías para una muestra de entrenamiento, el sistema de análisis determina el vector de características como un vector de elementos que incluye, para cada elemento, una de las puntuaciones de anomalías asociadas a uno de los sitios CpG. El sistema de análisis puede normalizar las puntuaciones de anomalías del vector de características basándose en la cobertura de la muestra, es decir, una profundidad de secuenciación mediana o promedio en todos los sitios CpG.

Con los vectores de características de las muestras de entrenamiento, el sistema de análisis puede entrenar el clasificador de HD. En una disposición, el sistema de análisis entrena un clasificador de HD binario para distinguir entre las etiquetas, el HD y el sin HD, basándose en los vectores de características de las muestras de entrenamiento. En esta disposición, el clasificador emite una puntuación de predicción que indica la probabilidad de la presencia o ausencia de HD. En otra disposición, el sistema de análisis entrena un clasificador multiclase de HD para distinguir entre muchos tipos de HD. En este clasificador multiclase de HD, el clasificador de HD se entrena para determinar una predicción de HD que comprende un valor de predicción para cada uno de los tipos de HD que se clasifican. Los valores de predicción pueden corresponder a una probabilidad de que una muestra dada tenga cada uno de los tipos de HD. Por ejemplo, el clasificador de HD devuelve una predicción de HD que incluye un valor de predicción para CHIP, leucemia, neoplasias linfoides (p. ej., linfoma), mieloma múltiple, neoplasia mieloide o cualquier combinación de estos. Por ejemplo, el clasificador de HD puede devolver una predicción de HD para una muestra de prueba que incluye una puntuación de predicción para el CHIP, la leucemia, las neoplasias linfoides (p. ej., linfoma), el mieloma múltiple, la neoplasia mieloide o cualquier combinación de estas. En cualquier disposición, el sistema de análisis entrena el clasificador de HD introduciendo conjuntos de muestras de entrenamiento con sus vectores de características en el clasificador de HD y ajustando los parámetros de clasificación para que una función del clasificador relacione con precisión los vectores de características de entrenamiento con su etiqueta correspondiente. El sistema de análisis puede agrupar las muestras de entrenamiento en conjuntos de una o más muestras de entrenamiento para el entrenamiento por lotes iterativo del clasificador de HD. Tras introducir todos los conjuntos de muestras de entrenamiento, que incluyen sus vectores de características de entrenamiento, y ajustar los parámetros de clasificación, el clasificador de HD está lo suficientemente entrenado para etiquetar las muestras de prueba según su vector de características dentro de cierto margen de error. El sistema de análisis puede entrenar el clasificador de HD según cualquiera de una serie de métodos. Por ejemplo, el clasificador de HD binario puede ser un clasificador de regresión logística L2-regularizado que se entrena usando una función de pérdida logarítmica. Como otro ejemplo, el clasificador multi-HD puede ser una regresión logística multinomial. En la práctica, cualquier tipo de clasificador de HD puede entrenarse usando otras técnicas. Estas técnicas son numerosas incluyendo uso potencial de métodos de núcleo, algoritmos de aprendizaje automático tales como redes neuronales multicapa, etc. En particular, métodos como se describe en el documento PCT/US2019/022122 y la solicitud de patente n.º US-16/352.602 (publicado como US-2019-0287652 A1).

Durante la implementación, el sistema de análisis obtiene una muestra de prueba de un sujeto de tipo HD desconocido. El sistema de análisis procesa la muestra de prueba para lograr un conjunto de fragmentos hipometilados e hipermetilados indicativos de HD. El sistema de análisis define un vector de características de prueba en un proceso similar al descrito para las muestras de entrenamiento. A continuación, el sistema de análisis introduce el vector de características de prueba en el clasificador de HD entrenado para producir una predicción de HD, p. ej., una predicción binaria (HD o sin HD) o una predicción de HD multiclase (puntuación de predicción para cada uno de una pluralidad de tipos de HD).

Clasificador del trastorno hematológico

En algunos ejemplos, el panel de ensayo descrito en la presente memoria puede usarse con un clasificador de trastorno hematológico que predice un estado de enfermedad para una muestra, tal como una predicción de trastorno hematológico o trastorno no hematológico, y/o una predicción indeterminada. En algunos ejemplos, el clasificador de trastorno hematológico puede generar características basándose en lecturas de secuencia teniendo en cuenta fragmentos de ADN metilados o no metilados en ciertas áreas genómicas de interés. Por ejemplo, si el clasificador de trastorno hematológico determina que un patrón de metilación en un fragmento se asemeja al de un cierto trastorno hematológico, entonces el clasificador de trastorno hematológico puede establecer una característica para ese fragmento como 1, y de lo contrario si no hay tal fragmento presente, entonces la característica puede establecerse como 0. De este modo, el clasificador de trastorno hematológico puede producir un conjunto de características binarias (simplemente a modo de ejemplo, 30.000 características) para cada muestra. Además, en algunos ejemplos, toda o una parte del conjunto de características binarias para una muestra puede introducirse en el clasificador de trastorno hematológico para proporcionar un conjunto de puntuaciones de probabilidad, tal como una puntuación de probabilidad por clase de trastorno hematológico y para una clase de trastorno no hematológico. Además, en algunos ejemplos, el clasificador de trastorno hematológico puede incorporar o usarse de otro modo junto con el umbral para determinar si una muestra debe denominarse trastorno hematológico o no trastorno hematológico, y/o un umbral indeterminado para reflejar la confianza en una llamada de trastorno hematológico específica. Dichos métodos se describen adicionalmente a continuación.

Para entrenar el clasificador de trastorno hematológico, el sistema de análisis (p. ej., el sistema 800 de análisis) puede obtener un conjunto de muestras de entrenamiento. En algunos ejemplos, cada muestra de entrenamiento incluye archivo(s) de fragmento (p. ej., datos de lectura de secuencia que contienen archivo), una etiqueta correspondiente a un tipo de trastorno hematológico o estado de trastorno no hematológico de la muestra y/o sexo del individuo de la muestra. El sistema de análisis puede utilizar el conjunto de entrenamiento para entrenar el clasificador de trastorno hematológico para predecir el estado de enfermedad de la muestra.

En algunos ejemplos, para el entrenamiento, el sistema de análisis divide el genoma (por ejemplo, el genoma completo) o un subconjunto del genoma (por ejemplo, regiones de metilación específicas) en regiones. Simplemente a modo de ejemplo, las porciones del genoma pueden separarse en "bloques" de CpG, por lo que un nuevo bloque comienza siempre que haya una separación entre los CpG más cercanos a los vecinos es al menos una distancia de separación mínima (p. ej., al menos 500 pb). Además, en algunos ejemplos, cada bloque puede dividirse en regiones de 1000 pb y colocarse de manera que las regiones vecinas tengan una cierta cantidad (por ejemplo, 50 % o 500 pb) de superposición.

Además, en algunos ejemplos, el sistema de análisis puede dividir el conjunto de entrenamiento en K subconjuntos o pliegues que se utilizarán en una validación cruzada de K pliegues. En algunos ejemplos, los pliegues pueden equilibrarse en función del estado del trastorno hematológico/no hematológico, el estadio del cáncer, la edad (p. ej., agrupados en grupos de 10 años) y/o el hábito de fumar. En algunos ejemplos, el conjunto de entrenamiento se divide en 5 pliegues, por lo que 5 clasificadores separados son entrenados, en cada caso entrenamiento en 4/5 de las muestras de entrenamiento y usando el 1/5 restante para la validación.

Durante el entrenamiento con el conjunto de entrenamiento, el sistema de análisis puede, para cada tipo de trastorno hematológico (y para el ADNlc sano), ajustar un modelo probabilístico a los fragmentos derivados de las muestras de ese tipo. Como se usa en la presente memoria, un "modelo probabilístico" es cualquier modelo matemático capaz de asignar una probabilidad a una secuencia leída basándose en el estado de metilación en uno o más sitios en la lectura. Durante el entrenamiento, el sistema de análisis se ajusta a lecturas de secuencia derivadas de una o más muestras de sujetos que tienen una enfermedad conocida y pueden usarse para determinar las probabilidades de lecturas de secuencia indicativas de un estado de enfermedad que utiliza información de metilación o vectores de estado de metilación. En particular, en algunos casos, el sistema de análisis determina las tasas observadas de metilación para cada sitio CpG dentro de una secuencia leída. La tasa de metilación representa una fracción o porcentaje de pares de bases que están metilados dentro de un sitio CpG. El modelo probabilístico entrenado puede parametrizarse mediante productos de las tasas de metilación. En general, puede usarse cualquier modelo probabilístico conocido para asignar probabilidades a lecturas de secuencia de una muestra. Por ejemplo, el modelo probabilístico puede ser un modelo binomial, en donde cada sitio (p. ej., sitio CpG) en un fragmento de ácido nucleico se asigna una probabilidad de metilación, o un modelo de sitios independientes, en donde cada metilación de CpG se especifica por una probabilidad de metilación distinta con metilación en un sitio que se supone que es independiente de la metilación en uno o más sitios diferentes en el fragmento de ácido nucleico.

En algunos ejemplos, el modelo probabilístico es un modelo de Markov, en donde la probabilidad de metilación en cada sitio CpG depende del estado de metilación en algún número de sitios CpG anteriores en la secuencia leída, o la molécula de ácido nucleico de la que se deriva la lectura de secuencia. Véase, p. ej., la solicitud de patente n.º US-16/352.602, titulada "Anomalous Fragment Detection and Classification," y presentada el 13 de marzo de 2019 (publicada como US-2019-0287652 A1)

En algunos ejemplos, el modelo probabilístico es un "modelo de mezcla" equipado con una mezcla de componentes de modelos subyacentes. Por ejemplo, en algunas disposiciones, los componentes de la mezcla pueden determinarse usando de múltiples modelos de sitios independientes, donde se supone que la metilación (p. ej., las tasas de metilación) en cada sitio CpG es independiente de la metilación en otros sitios CpG. Utilizando un modelo de sitios independientes, la probabilidad asignada a una lectura de secuencia, o la molécula de ácido nucleico de la que deriva, es el producto de la probabilidad de metilación en cada sitio CpG donde la lectura de secuencia está metilada y una menos la probabilidad de metilación en cada sitio CpG donde la lectura de secuencia no está metilada. Según este ejemplo, el sistema de análisis determina las tasas de metilación de cada uno de los componentes de la mezcla. El modelo de mezcla se parametriza mediante una suma de los componentes de la mezcla, cada uno asociado con un producto de las tasas de metilación. Un modelo probabilístico *Pr* de *n* componentes de mezcla pueden representarse como:

$$Pr(fragmento|\{\beta_{ki}, f_k\}) = \sum_{k=1}^{n} f_k \prod_i \beta_{ki}^{m_i}(1-\beta_{ki})^{1-m_i}$$

Para un fragmento de entrada, $m_i \in \{0,1\}$ representa el estado de metilación observado del fragmento en la posición *i* de un genoma de referencia, indicando 0 ninguna metilación e indicando 1 metilación. Una asignación fraccional A cada componente de mezcla *k* es f_k, donde $f_k \geq 0$ y $\sum_{k=1}^{n} f_k = 1$. La probabilidad de metilación en posición *i* en un sitio CpG del componente de mezcla *k* es β_{ki}. Por tanto, la probabilidad de no metilación es $1 - \beta_{ki}$. El número de componentes de mezcla *n* puede ser 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, etc.

En algunos ejemplos, el sistema de análisis se ajusta al modelo probabilístico usando una estimación máxima de la probabilidad para identificar un conjunto de parámetros $\{\beta_{ki}, f_k\}$ que maximiza la probabilidad logarítmica de todos los fragmentos derivados de un estado de enfermedad, sujeto a una penalización de regularización aplicada a cada probabilidad de metilación con resistencia de regularización *r*. La cantidad maximizada para N fragmentos totales puede representarse como:

$$\sum_{j}^{N} ln\left(Pr\left(fragmento_j|\{\beta_{ki}, f_k\}\right)\right) + r \cdot ln\left(\beta_{ki}(1-\beta_{ki})\right)$$

En algunos ejemplos, el sistema de análisis realiza ajustes por separado para cada trastorno hematológico y para el ADNlc sano. Como apreciaría un experto en la técnica, pueden usarse otros medios para ajustar los modelos probabilísticos o para identificar parámetros que maximizan la probabilidad logarítmica de todas las lecturas de secuencia derivadas de las muestras de referencia. Por ejemplo, en algunos ejemplos, se usa el ajuste bayesiano (usando, por ejemplo, la cadena Markov Monte Carlo), en donde cada parámetro no se asigna un valor único, sino que se asocia a una distribución. En algunos ejemplos, se usa la optimización basándose en gradiente, en la que el gradiente de la probabilidad (o probabilidad logarítmica) con respecto a los valores de parámetro se usa para paso a través del espacio de parámetro hacia un óptimo. En todavía algunos ejemplos, la maximización de expectativa, en la que se deriva un conjunto de parámetros latentes (tales como identidades del componente de mezcla de las cuales se deriva cada fragmento) se fija en sus valores esperados bajo los parámetros del modelo anteriores, y después los parámetros del modelo se asignan para maximizar la probabilidad condicional de los valores supuestos de esas variables latentes. El proceso de dos etapas se repite hasta la convergencia.

Además, en algunos ejemplos, el sistema de análisis puede generar características para cada muestra en el conjunto de entrenamiento. Por ejemplo, para cada muestra (independientemente de la etiqueta), en cada región, para cada tipo de trastorno hematológico, para cada fragmento, el sistema de análisis puede evaluar la relación log-probabilidad *R* con los modelos probabilísticos ajustados según:

$$R_{trastorno\ hematológico\ A}\left(fragmento\right) \equiv\ ln\left(\frac{Pr\left(fragmento|trastorno\ hematológico\ A\right)}{Pr\left(fragmento|cfDNA\ sano\right)}\right)$$

A continuación, para cada muestra, para cada región, para cada trastorno hematológico, para cada uno de un conjunto de valores de "nivel", el sistema de análisis puede contar el número de fragmentos con $R_{\text{trastorno hematológico}}$> nivel y asignar esos recuentos como características no negativas de valor entero. Por ejemplo, los niveles incluyen valores umbral de 1, 2, 3, 4, 5, 6, 7, 8 y 9, lo que da como resultado cada región que aloja 9 características por trastorno hematológico.

En algunos ejemplos, el sistema de análisis puede seleccionar ciertas características para su inclusión en un vector de características para cada muestra. Por ejemplo, para cada par de trastorno hematológico distinto, el sistema de análisis puede especificar un tipo como el "tipo positivo" y el otro como el "tipo negativo" y clasificar las características por su capacidad para distinguir esos tipos. En algunos casos, la clasificación se basa en información mutua calculada por el sistema de análisis. Por ejemplo, la información mutua puede calcularse usando la fracción estimada de muestras del tipo positivo y tipo negativo (p. ej., trastornos hematológicos A y B) para los cuales se espera que la característica sea distinta de cero en un ensayo resultante. Por ejemplo, si una característica se produce con frecuencia en ADNlc sano, el sistema de análisis determina que la característica es poco probable que ocurra con frecuencia en ADNlc asociado con diversos tipos de trastorno hematológico. Por consiguiente, la característica puede ser una medida débil para distinguir entre estados de enfermedad. Al calcular la información mutua *I*, la variable *X* es una determinada característica (p. ej., binaria) y variable *Y* representa un estado de enfermedad, p. ej., trastornos hematológicos A o B:

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\ log\ log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

$$I \approx \frac{1}{2}\left(p(1|A)\cdot log\left(\frac{p(1|A)}{\frac{1}{2}\left(p(1|A)+p(1|B)\right)}\right) + p(1|B)\cdot log\left(\frac{p(1|B)}{\frac{1}{2}\left(p(1|A)p(1|B)\right)}\right)\right)$$

$$p(1|A) = f_A + f_H - f_H f_A$$

La función de masa de probabilidad conjunta de X y *Y* es $p(x,y)$ y las funciones de masa de probabilidad marginal son $p(x)$ y $p(y)$. El sistema de análisis puede asumir que la ausencia de características no es informativa y el estado de la enfermedad es igualmente probable *a priori,* por ejemplo, $p(Y=A) = p(Y=B) = 0.5$. La probabilidad de observar (p. ej., en ADNlc) una característica binaria dada de trastorno hematológico A está representada por $p(1|A)$, donde f_A es la probabilidad de observar la característica en muestras de ADNtc del tumor (o muestras de ADNlc de alta señal) asociadas con el trastorno hematológico A, y f_H es la probabilidad de observar la característica en una muestra de ADNlc sana o de trastorno no hematológico.

En algunos ejemplos, solo las características correspondientes al tipo positivo se incluyen en la clasificación, y solo cuando esas características de la tasa de ocurrencia prevista son mayores en el tipo positivo que en el tipo negativo. Por ejemplo, si "hígado" es el tipo positivo y "mama" es el tipo negativo, entonces solo se consideran características

"hígado_x", y solo si su aparición estimada en el ADNlc hepático es mayor que su aparición estimada en ADNlc de mama. Además, en algunos ejemplos, para cada región, para cada par de trastorno hematológico (que incluye trastorno no hematológico como un tipo negativo), el sistema de análisis mantiene solo el nivel de mejor rendimiento. Además, en algunos ejemplos, el sistema de análisis transforma los valores característicos mediante binarización, por lo que cualquier valor característico mayor que 0 se establece en 1, de modo que todas las características sean 0 o 1.

En algunos ejemplos, el sistema de análisis trenes de un clasificador de regresión logística multinomial en los datos de entrenamiento para un pliegue, y genera predicciones para los datos de salida. Por ejemplo, para cada uno de los K pliegues, se puede entrenar una regresión logística para cada combinación de hiperparámetros. Tales hiperparámetros pueden incluir penalización de L2 y/o topK (p. ej., el número de regiones de alto rango para mantenerse por par de tipos de tejido (incluido trastorno no hematológico), como se clasifica por el procedimiento de información mutua descrito anteriormente). Para cada conjunto de hiperparámetros, el rendimiento se evalúa en las predicciones validadas cruzadas del conjunto de entrenamiento completo, y el conjunto de hiperparámetros con el mejor rendimiento se selecciona para reentrenar en el conjunto de entrenamiento completo. En algunos ejemplos, el sistema de análisis usa la pérdida de registro como una métrica de rendimiento, por lo que la pérdida de registro se calcula tomando el logaritmo negativo de la predicción para la etiqueta correcta para cada muestra, y luego sumando sobre muestras (es decir, una predicción perfecta de 1,0 para la etiqueta correcta daría una pérdida logarítmica de 0).

Para generar predicciones para una nueva muestra, los valores de las características se calculan utilizando el mismo método descrito anteriormente, pero restringido a las características (combinaciones región/clase positiva) seleccionadas bajo el valor topK elegido. Las características generadas se usan entonces para crear una predicción usando el modelo de regresión logística entrenado anteriormente.

En algunos ejemplos, los trenes de analíticas un clasificador de dos etapas. Por ejemplo, el sistema de análisis entrena un clasificador de trastorno hematológico binario para distinguir entre las etiquetas, el trastorno hematológico y el trastorno no hematológico, basándose en los vectores de características de las muestras de entrenamiento. En este caso, el clasificador binario emite una puntuación de predicción que indica la probabilidad de la presencia o ausencia del trastorno hematológico. En otro ejemplo, el sistema de análisis entrena un clasificador multiclase de trastorno hematológico para distinguir entre muchos trastornos hematológicos. En este clasificador multiclase de trastorno hematológico, el clasificador de trastorno hematológico se entrena para determinar una predicción de trastorno hematológico que comprende un valor de predicción para cada uno de los trastornos hematológicos que se clasifican. Los valores de predicción pueden corresponder a una probabilidad de que una muestra dada tenga cada uno de los trastornos hematológicos. Por ejemplo, el clasificador de trastornos hematológicos devuelve una predicción de trastorno hematológico que incluye un valor de predicción para el CHIP, la leucemia, las neoplasias linfoides (p. ej., linfoma), el mieloma múltiple, una neoplasia mieloide y un trastorno no hematológico. Por ejemplo, el clasificador de trastorno hematológico puede devolver una predicción de trastorno hematológico para una muestra de prueba que incluye una puntuación de predicción para el CHIP, la leucemia, las neoplasias linfoides (p. ej., linfoma), el mieloma múltiple, una neoplasia mieloide y/o un trastorno no hematológico.

El sistema de análisis puede entrenar el clasificador de trastorno hematológico según cualquiera de una serie de métodos. Por ejemplo, el clasificador binario de trastorno hematológico puede ser un clasificador de regresión logística L2-regularizado que se entrena utilizando una función de pérdida logarítmica. Como otro ejemplo, el clasificador de múltiples trastornos hematológicos puede ser una regresión logística multinomial. En la práctica, cualquier tipo de clasificador de trastorno hematológico puede entrenarse usando otras técnicas. Estas técnicas son numerosas incluyendo uso potencial de métodos de núcleo, algoritmos de aprendizaje automático tales como redes neuronales multicapa, etc. En particular, métodos como se describe en el documento PCT/US2019/022122 y la solicitud de patente n.º US-16/352.602 (publicado como US-2019-0287652 A1).

Secuenciador y sistema de análisis ilustrativos

La Figura **9A** es un diagrama de flujo de sistemas y dispositivos para secuenciar muestras de ácido nucleico según una disposición. Este diagrama de flujo ilustrativo incluye dispositivos tales como un secuenciador 820 y un sistema 800 de análisis. El secuenciador 820 y el sistema 800 de analítica pueden funcionar en tándem para realizar una o más etapas en los procesos descritos en la presente memoria.

En diversas disposiciones, el secuenciador 820 recibe una muestra 810 de ácido nucleico enriquecida. Como se muestra en la Figura **9A,** el secuenciador 820 puede incluir una interfaz gráfica de usuario 825 que permite interacciones de usuario con tareas particulares (p. ej., iniciar secuenciación o terminar secuenciación) así como una más estaciones 830 de carga para cargar un cartucho de secuenciación que incluye las muestras de fragmentos enriquecidas y/o para cargar los tampones necesarios para realizar los ensayos de secuenciación. Por tanto, una vez que un usuario del secuenciador 820 ha proporcionado los reactivos necesarios y el cartucho de secuenciación a la estación 830 de carga del secuenciador 820, el usuario puede iniciar la secuenciación interactuando con la interfaz gráfica de usuario 825 del secuenciador 820. Una vez iniciada, el secuenciador 820 realiza la secuenciación y emite las lecturas de secuencia de los fragmentos enriquecidos de la muestra de ácido nucleico 810.

En algunas disposiciones, el secuenciador 820 está acoplado comunicativamente con el sistema 800 de análisis. El sistema 800 de análisis incluye algún número de dispositivos informáticos utilizados para procesar las lecturas de secuencia para diversas aplicaciones, tales como evaluar el estado de metilación en uno o más sitios CpG, llamada variante o control de calidad. El secuenciador 820 puede proporcionar las lecturas de secuencia en un formato de archivo BAM al sistema 800 de analítica. El sistema 800 de analítica se puede acoplar comunicativamente al secuenciador 820 a través de una red inalámbrica, cableada o de comunicación inalámbrica. Generalmente, el sistema 800 de analítica está configurado con un procesador y un medio de almacenamiento legible por ordenador no transitorio que almacena instrucciones informáticas que, cuando son ejecutadas por el procesador, hacen que el procesador procese las lecturas de secuencia o realice una o más etapas de cualquiera de los métodos o procesos descritos en la presente memoria.

En algunas disposiciones, las lecturas de secuencia pueden alinearse con un genoma de referencia usando métodos conocidos en la técnica para determinar la información de posición de alineación. La posición de alineación generalmente puede describir una posición inicial y una posición final de una región en el genoma de referencia que corresponde a una base de nucleótidos inicial y una base de nucleótidos final de una secuencia de lectura dada. Correspondientes a la secuenciación de metilación, la información de la posición de alineación puede generalizarse para indicar un primer sitio CpG y un último sitio CpG incluido en la secuencia leída según la alineación con el genoma de referencia. La información de la posición de alineación puede indicar además estados de metilación y ubicaciones de todos los sitios CpG en una secuencia dada. Una región en el genoma de referencia puede asociarse con un gen o un segmento de un gen; como tal, el sistema de análisis 800 puede etiquetar una secuencia leída con uno o más genes que se alinean con la secuencia leída. En una realización, la longitud del fragmento (o tamaño) se determina desde las posiciones inicial y final.

En diversas disposiciones, por ejemplo cuando se usa un proceso de secuenciación de extremos emparejados, una lectura de secuencia comprende un par de lectura indicado como R_1 y R_2. Por ejemplo, la primera lectura R_1 se puede secuenciar desde un primer extremo de una molécula de ADN bicatenario (ADNbc) mientras que la segunda lectura R_2 se puede secuenciar desde el segundo extremo del ADN bicatenario (ADNbc). Por tanto, los pares de bases de nucleótidos de la primera lectura R_1 y la segunda lectura R_2 pueden alinearse de manera consistente (por ejemplo, en orientaciones opuestas) con bases de nucleótidos del genoma de referencia. La información de posición de alineación derivada del par de lectura R_1 y R_2 puede incluir una posición inicial en el genoma de referencia que corresponde a un extremo de una primera lectura (por ejemplo, R_1) y una posición final en el genoma de referencia que corresponde a un extremo de una segunda lectura (por ejemplo, R_2). En otras palabras, la posición inicial y la posición final en el genoma de referencia representan la ubicación probable dentro del genoma de referencia al que corresponde el fragmento de ácido nucleico. En una disposición, el par de lectura R_1 y R_2 se pueden ensamblar en un fragmento, y el fragmento usado para el análisis y/o clasificación posterior. Se puede generar un archivo de salida que tenga formato SAM (mapa de alineación de secuencia) o formato BAM (binario) y se emita para su posterior análisis.

Con referencia ahora a la **Figura 9B, Figura 9B** es un diagrama de bloques de un sistema 800 de análisis para procesar muestras de ADN según una disposición. El sistema de análisis implementa uno o más dispositivos informáticos para su uso en el análisis de muestras de ADN. El sistema de análisis 800 incluye un procesador de secuencia 840, base de datos de secuencias 845, base de datos de modelos 855, modelos 850, base de datos de parámetros 865 y motor de puntuación 860. En algunas disposiciones, el sistema 800 de análisis realiza una o más etapas en los procesos 300 de la **Figura 3A,** 340 de la **Figura 3B,** 400 de la **Figura 4,** 500 de la **Figura 5,** 600 de la **Figura 6A,** o 680 de la **Figura 6B** y otro proceso descrito en la presente memoria.

El procesador de secuencia 840 genera vectores de estado de metilación para fragmentos de una muestra. En cada sitio CpG en un fragmento, el procesador 840 de secuencia genera un vector de estado de metilación para cada fragmento que especifica una ubicación del fragmento en el genoma de referencia, un número de sitios CpG en el fragmento y el estado de metilación de cada sitio CpG en el fragmento ya sea metilado, no metilado o indeterminado a través del proceso 300 de la **Figura 3A.** El procesador de secuencia 840 puede almacenar vectores de estado de metilación para fragmentos en la base de datos de secuencias 845. Los datos en la base de datos de secuencias 845 pueden organizarse de manera que los vectores de estado de metilación de una muestra están asociados entre sí.

Además, pueden almacenarse múltiples modelos 850 diferentes en la base de datos 855 de modelo o recuperarse para su uso con muestras de prueba. En un ejemplo, un modelo es un clasificador de trastorno hematológico entrenado para determinar una predicción de trastorno hematológico para una muestra de prueba usando un vector de características derivado de fragmentos anómalos. El entrenamiento y uso del clasificador de trastorno hematológico se describe en otra parte de la presente memoria. El sistema de análisis 800 puede entrenar uno o más modelos 850 y almacenar diversos parámetros entrenados en la base de datos de parámetros 865. El sistema de análisis 800 almacena los modelos 850 junto con funciones en la base de datos de modelos 855.

Durante la inferencia, el motor 860 de puntuación usa los uno o más modelos 850 para devolver salidas. El motor de puntuación 860 accede a los modelos 850 en la base de datos de modelos 855 junto con parámetros entrenados de la base de datos de parámetros 865. Según cada modelo, el motor de puntuación recibe una entrada adecuada para el modelo y calcula una salida basándose en la entrada recibida, los parámetros y una función de cada modelo

relacionan la entrada y la salida. En algunos casos de uso, el motor 860 de puntuación calcula además métricas que se correlacionan con una confianza en las salidas calculadas del modelo. En otros casos de uso, el motor 860 de puntuación calcula otros valores intermediarios para su uso en el modelo.

Ejemplos

Los siguientes ejemplos se presentan para proporcionar a los expertos en la técnica una descripción completa y descripción de cómo hacer y usar la presente descripción, y no pretenden limitar el alcance de lo que los inventores consideran su descripción ni pretenden representar que los experimentos siguientes son todos o los únicos experimentos realizados. Se han realizado esfuerzos para garantizar la precisión con respecto a los números utilizados (por ejemplo, cantidades, temperatura, etc.) pero se deben tener en cuenta algunos errores experimentales y desviaciones.

Ejemplo 1 - Análisis de las cualidades de la sonda

Para comprobar cuánto solapamiento es necesario entre un fragmento de ANDlc y una sonda para lograr una cantidad no despreciable de extracción, se probaron varias longitudes de solapamientos utilizando paneles diseñados para incluir tres tipos diferentes de sondas (V1D3, V1D4, V1E2) con varios solapamientos con fragmentos de ADN diana de 175 pb específicos para cada sonda. Los solapamientos probados oscilaban entre 0 pb y 120 pb. Las muestras que contenían fragmentos de ADN diana de 175 pb se aplicaron al panel y se lavaron, y a continuación se recogieron los fragmentos de ADN unidos a las sondas. Se midieron las cantidades de los fragmentos de ADN recogidos y las cantidades se trazaron como densidades sobre los tamaños de superposiciones, tal como se indica en la **Figura 8.**

No hubo una unión significativa y la extracción de fragmentos de ADN diana cuando había menos de 45 pb de superposición. Estos resultados sugieren que generalmente se requiere un solapamiento de sonda-sonda de al menos 45 pb para lograr una cantidad no despreciable de extracción, aunque este número puede variar dependiendo de las condiciones del ensayo.

Además, se ha sugerido que más de una tasa de emparejamiento erróneo del 10 % entre la sonda y las secuencias de fragmentos en la región de solapamiento es suficiente para interrumpir en gran medida la unión y, por tanto, la eficiencia de la extracción. Por tanto, las secuencias que pueden alinearse con la sonda a lo largo de al menos 45 pb con al menos una tasa de coincidencia del 90 % son candidatas para la extracción fuera de la diana.

Por tanto, hemos realizado una búsqueda exhaustiva de todas las regiones genómicas que tienen alineaciones de 45 pb con una tasa de coincidencia de 90 % + (es decir, regiones fuera de diana) para cada sonda. Específicamente, se combinó una estrategia de siembra de k-mero (que puede permitir una o más faltas de coincidencia) con alineación local en las ubicaciones de las semillas. Esto garantizó que no falte ningún buen alineamiento basándose en la longitud de k-mero, se permitió el número de faltas de coincidencia y el número de aciertos de semilla de k-mero en una ubicación particular. Esto implica realizar una alineación local de programación dinámica en un gran número de ubicaciones, por lo que la implementación se optimizó para usar instrucciones de CPU de vectores (por ejemplo, AVX2, AVX512) y paralelizó a través de muchos núcleos dentro de una máquina y también a través de muchas máquinas conectadas por una red. Esto permite una búsqueda exhaustiva que es valiosa para diseñar un panel de alto rendimiento (es decir, velocidad baja fuera de la diana y alta cobertura diana para una cantidad dada de secuenciación).

Después de la búsqueda exhaustiva, cada sonda se calificó en base al número de regiones fuera de la diana. Las mejores sondas tienen una puntuación de 1, lo que significa que coinciden en un solo lugar (Q alta). Se aceptaron las sondas con una puntuación baja entre 2-19 aciertos (Q baja), pero se descartaron las sondas con una puntuación baja de más de 20 aciertos (Q baja). Pueden usarse otros valores de corte para muestras específicas.

A continuación, se contaron los números de sondas de alta, baja calidad y mala calidad entre las sondas dirigidas a regiones genómicas hipermetiladas o regiones genómicas hipometiladas.

Ejemplo 2 - Un panel de ensayo para detectar trastornos hematológicos

Trastornos hematológicos: Se diseñó un panel de HD para detectar diferentes tipos de trastornos hematológicos, que incluye el CHIP, la leucemia, el mieloma múltiple y el linfoma.

Muestras usadas para la selección de región genómica: Se usaron muestras de diferentes fuentes para la selección de las regiones genómicas diana. Incluyen (1) células tumorales diseminadas (DTC, por sus siglas en inglés) enriquecidas con células de diferentes tipos de cáncer, (2) muestras de células mononucleares de médula ósea (PBMC, por sus siglas en inglés) de pacientes con leucemia, linfoma o mieloma múltiple, (3) muestras de células mononucleares de sangre periférica (PBMC) de pacientes con leucemia, linfoma o mieloma múltiple, (4) ADN genómico de bloques de tejido FFPE de muestras de cáncer, o (5) ADN genómico de glóbulos blancos, o (7) muestras de ADNlc de más de 1800 individuos.

Selección de región (según el estado de metilación): Para la selección de la diana, se seleccionaron fragmentos que tienen patrones de metilación anormales en muestras de distintos trastornos hematológicos usando de uno o más métodos como se describe en la presente memoria. El uso de estos métodos permitió la identificación de regiones de bajo ruido como dianas putativas. Entre las regiones de bajo ruido, se clasificaron y seleccionaron fragmentos más informativos en los tipos de enfermedad discriminatorios.

Específicamente, en algunas disposiciones, cuando se usaron los datos WGBS, las secuencias de fragmentos en la base de datos se filtraron en función del valor de p usando una distribución en individuos de control sanos, y sólo se retuvieron los fragmentos con $p < 0,001$, como se describe en la presente memoria. En algunos casos, los ADNlc seleccionados se filtraron adicionalmente para retener solo aquellos que estaban al menos 90 % metilados o 90 % no metilados. A continuación, para cada sitio CpG en los fragmentos seleccionados, se contaron los números de muestras con un trastorno hematológico o muestras de control sanas que incluyen fragmentos que se superponen en el sitio CpG. En concreto, se calculó P (trastorno hematológico | fragmento superpuesto) para cada CpG y se seleccionaron los sitios genómicos con altos valores de P como dianas generales del trastorno. Por diseño, los fragmentos seleccionados tenían un ruido muy bajo (es decir, se superponen pocos fragmentos sanos).

Para encontrar dianas específicas para un trastorno hematológico, se realizaron procesos de selección similares. Los sitios CpG se clasificaron en base a su ganancia de información, comparando (i) entre los números de muestras de un trastorno hematológico específico y otras muestras, en donde otras muestras que incluyen las muestras de control sanas y las muestras de un trastorno hematológico diferente, (ii) entre los números de muestras de un trastorno hematológico específico y sanas, muestras de control y/o (iii) entre los números de muestras de un trastorno hematológico específico y un trastorno hematológico diferente que incluyen fragmentos que se superponen en el sitio CpG. El proceso se aplicó a cada uno de los trastornos hematológicos y la comparación se realizó para todas las combinaciones por pares para los trastornos hematológicos, tal como se ilustra en la **Figura 2**. Por ejemplo, se calculó P (un trastorno hematológico | fragmento superpuesto) y después se comparó con P (un trastorno hematológico diferente | fragmento superpuesto). Como diana para el trastorno hematológico, se seleccionó un fragmento atípico en cada trastorno hematológico que tenía una probabilidad mucho mayor bajo un trastorno hematológico que bajo un trastorno hematológico diferente. Por consiguiente, las regiones genómicas seleccionadas por las comparaciones por pares incluyeron regiones genómicas diferencialmente metiladas para separar un trastorno hematológico diana y un trastorno hematológico de contraste.

Las regiones genómicas diana seleccionadas como se describe en esta sección se listan en la **TABLA 1**. Las regiones genómicas diana de las listas 2-4 contienen subconjuntos de los sitios de metilación de las regiones genómicas diana de las lista 5-7, respectivamente. Del mismo modo, las regiones genómicas diana de la lista 8 contienen un subconjunto de los sitios de metilación de las regiones genómicas diana de la lista 1.

TABLA 1 - Id. de sec. n.° correspondientes a las listas 1-8. Para cada lista, la tabla identifica el número total de regiones genómicas diana de la lista, un intervalo de la Id. de sec. n.° correspondientes a todas las regiones genómicas diana de la lista que se encuentran en el listado de secuencias presentado con esta solicitud, y el total de las longitudes de todas las regiones genómicas diana de la lista. El listado de secuencias identifica la ubicación cromosómica de cada región genómica diana, si los fragmentos de ADNlc se enriquecen de la región están hipermetilados o hipometilados, y la secuencia de una cadena de ADN de la región genómica diana. Los números de cromosomas y las posiciones de inicio y parada se proporcionan en relación con un genoma de referencia humana conocido, hg19. La secuencia del genoma humano de referencia, hg19, está disponible en Genome Reference Consortium con un número de referencia, GRCh37/hg19, y también está disponible en Genome Browser proporcionado por Santa Cruz Genomics Institute.

Lista	Trastorno hematológico dirigido	Regiones genómicas diana	Id. de sec. n.°		Tamaño del panel (kb)
			Primera	Última	
1	Todos	28130	1	28130	1586
2	Neoplasia linfoide	1447	28131	29577	403
3	Mieloma múltiple	879	29578	30456	277
4	Neoplasia mieloide	1255	30457	31711	299
5	Neoplasia linfoide	1170	31712	32881	612
6	Mieloma múltiple	822	32882	33703	315
7	Neoplasia mieloide	1177	33704	34880	447
8	Todos	22456	34881	57336	1160

Ejemplo 3 - Generación de un clasificador de modelo de mezcla

Para maximizar el rendimiento, los modelos predictivos de cáncer descritos en este Ejemplo se entrenaron utilizando datos de secuencia obtenidos de una pluralidad de muestras de tipos de cáncer conocidos y no cancerosos de ambos subestudios CCGA (CCGA1 y CCGA22), una pluralidad de muestras de tejido para cánceres conocidos obtenidas de CCGA1, y una pluralidad de muestras no cancerosas del estudio STRIVE (véase Clinical Trail.gov Identifier: NCT03085888 (//clinicaltrials.gov/ct2/show/NCT03085888)). El estudio STRIVE es un estudio de cohortes prospectivo, multicéntrico y observacional para validar un ensayo para la detección precoz del cáncer de mama y otros cánceres invasivos, del que se obtuvieron muestras de entrenamiento no cancerosas adicionales para entrenar el clasificador descrito en la presente memoria. Los tipos de cáncer conocidos incluidos en el conjunto de muestras CCGA fueron los siguientes: mama, pulmón, próstata, colorrectal, renal, uterino, páncreas, esofágico, linfoma, cabeza y cuello, ovario, hepatobiliar, melanoma, cervical, mieloma múltiple, leucemia, tiroides, vejiga, gástrico y anorrectal. Como tal, un modelo puede ser un modelo multicáncer (o un clasificador multicáncer) para detectar uno o más, dos o más, tres o más, cuatro o más, cinco o más, diez o más, o 20 o más tipos diferentes de cáncer.

Los datos de rendimiento del clasificador que se muestran a continuación se obtuvieron para un clasificador bloqueado entrenado con muestras de cáncer y no cáncer obtenidas de CCGA2, un subestudio de CCGA, y con muestras no cancerosas de STRIVE. Los individuos del subestudio CCGA2 eran diferentes de los individuos del subestudio CCGA1 cuyo ADNlc se usó para seleccionar los genomas diana. En el estudio CCGA2 se recogieron muestras de sangre de individuos diagnosticados de cáncer sin tratar (incluidos 20 tipos de tumores y todos los estadios del cáncer) y de individuos sanos sin diagnóstico de cáncer (controles). Para STRIVE, se recogieron muestras de sangre de mujeres en los 28 días siguientes a su mamografía de cribado. El ADN libre de células (ADNlc) se extrajo de cada muestra y se trató con bisulfito para convertir citosinas no metiladas en uracilos. El ADNlc tratado con bisulfito se enriqueció para dar moléculas de ADNlc informativas usando sondas de hibridación diseñadas para enriquecer ácidos nucleicos convertidos por bisulfito derivados de cada una de una pluralidad de regiones genómicas dirigidas en un panel de ensayo que comprende todas las regiones genómicas de las listas 1-8. Las moléculas de ácido nucleico enriquecidas con bisulfito enriquecidas se secuenciaron mediante el uso de secuenciación de extremos emparejados en una plataforma Illumina (San Diego, CA) para obtener un conjunto de lecturas de secuencia para cada una de las muestras de entrenamiento, y los pares de lecturas resultantes se alinearon con el genoma de referencia, se ensamblaron en fragmentos y se identificaron sitios CpG metilados y no metilados.

Caracterización basándose en el modelo de mezcla

Para cada tipo de cáncer (incluido el no canceroso) se entrenó y utilizó un modelo de mezcla probabilística para asignar una probabilidad a cada fragmento de cada muestra cancerosa y no cancerosa en función de la probabilidad de que el fragmento se observara en un tipo de muestra determinado.

Análisis a nivel de fragmento

Brevemente, para cada tipo de muestra (muestras de cáncer y de no cáncer), para cada región (donde cada región se utilizó tal cual si era menor de 1 kb, o bien se subdividió en regiones de 1 kb de longitud con un solapamiento del 50 % (por ejemplo, solapamiento de 500 pares de bases) entre regiones adyacentes), se ajustó un modelo probabilístico a los fragmentos derivados de las muestras de entrenamiento para cada tipo de cáncer y de no cáncer. El modelo probabilístico entrenado para cada tipo de muestra fue un modelo de mezcla, en el que cada uno de los tres componentes de la mezcla era un modelo de sitios independientes en el que se supone que la metilación en cada CpG es independiente de la metilación en otros CpG. Los fragmentos se excluyeron del modelo si: tenían un valor de p (de un modelo de Markov no canceroso) superior a 0,01; se marcaron como fragmentos duplicados; los fragmentos tenían un tamaño de bolsa superior a 1 (sólo para las muestras de metilación dirigida); no cubrían al menos un sitio CpG; o si el fragmento tenía una longitud superior a 1000 bases. Los fragmentos de entrenamiento retenidos se asignaron a una región si se solaparon al menos un CpG de esa región. Si un fragmento solapaba CpG en varias regiones, se asignaba a todas ellas.

Modelos de fuentes locales

Cada modelo probabilístico se ajustó utilizando la estimación de máxima verosimilitud para identificar un conjunto de parámetros que maximizaran la probabilidad logarítmica de todos los fragmentos derivados de cada tipo de muestra, sujeto a una penalización de regularización.

Específicamente, en cada región de clasificación se entrenó un conjunto de modelos probabilísticos, uno para cada etiqueta de entrenamiento (es decir, uno para cada tipo de cáncer y otro para los no cancerosos). Cada modelo tomó la forma de un modelo de mezcla Bernoulli con tres componentes. Matemáticamente,

$$(1)\ Pr\,(fragmento|\{\beta_{ki}, f_k\}) = \sum_{k=1}^{n} f_k \prod_i \beta_{ki}^{m_i}(1-\beta_{ki})^{1-m_i}$$

cuando n es el número de componentes de mezcla, establecido en 3; $m_i \in \{0, 1\}$ es la metilación observada del fragmento en la posición i; f_k es la asignación fraccionaria al componente k (con $f_k \geq 0$ y $\sum f_k = 1$); y β_{ki} es la fracción de metilación en el componente k en CpG i. El producto sobre i incluyó solo aquellas posiciones para las que podría identificarse un estado de metilación a partir de la secuenciación. Se estimaron valores de probabilidad máxima de los

parámetros $\{f_k, \beta_{ki}\}$ de cada modelo usando el algoritmo rprop (p. ej., el algoritmo rprop tal como se describe en Riedmiller M, Braun H. RPROP-A Fast Adaptive Learning Algorithm. Proceedings of the International Symposium on Computer and Information Science VII, 1992) para maximizar la probabilidad logarítmica total de los fragmentos de una etiqueta de entrenamiento, sujeto a una penalización de regularización en β_{ki} que tomó la forma de un beta antes distribuido. Matemáticamente, la cantidad maximizada fue

$$(2)\ \sum_j \ ln\left(Pr\left(fragmento_j|\{\beta_{ki}, f_k\}\right)\right) + \sum_{k,i} \ r \ ln\left(\beta_{ki}(1-\beta_{ki})\right)$$

donde r es la resistencia de regularización, que se estableció en 1.

Caracterización

Una vez entrenados los modelos probabilísticos, se calculó un conjunto de características numéricas para cada muestra. Específicamente, se extrajeron características para cada fragmento de cada muestra de entrenamiento, para cada muestra de cáncer y muestra no cancerosa, en cada región. Las características extraídas fueron los recuentos de fragmentos atípicos (es decir, fragmentos anómalamente metilados), que se definieron como aquellos cuya probabilidad logarítmica bajo un primer modelo de cáncer superaba la probabilidad logarítmica bajo un segundo modelo de cáncer o modelo de no cáncer en al menos un valor de nivel umbral. Los fragmentos atípicos se calcularon por separado para cada región genómica, modelo de muestra (es decir, tipo de cáncer) y nivel (para los niveles 1, 2, 3, 4, 5, 6, 7, 8 y 9), produciendo 9 características por región para cada tipo de muestra. De esta manera, cada característica se definió por tres propiedades: una región genómica; un marcador de tipo cáncer "positivo" (que excluye el no cáncer); y el valor del nivel seleccionado del conjunto {1, 2, 3, 4, 5, 6, 7, 8, 9}. El valor numérico de cada característica se definió como el número de fragmentos en esa región de manera que

$$(3)\ ln\left(\frac{Pr(fragmento|tipo\ cáncer\ positivo)}{Pr(fragmento|no\ cáncer)}\right) > nivel$$

donde las probabilidades se definieron mediante la ecuación (1) usando los valores de parámetros estimados de probabilidad máxima correspondientes al tipo de cáncer "positivo" (en el numerador del logaritmo) o al no cáncer (en el denominador).

Clasificación de características

Para cada conjunto de características por pares, las características se clasificaron mediante el uso de información mutua basándose en su capacidad para distinguir el primer tipo de cáncer (que definió el modelo de probabilidad logarítmica a partir del cual se obtuvo la característica) del segundo tipo de cáncer o no cáncer. Específicamente, se compilaron dos listas clasificadas de características para cada par único de marcadores de clase: uno con el primer marcador asignado como "positivo" y el segundo como el "negativo", y el otro con la asignación positiva/negativa intercambiada (con la excepción del marcador "no canceroso", que solo se permitió como marcador negativo). Para cada una de estas listas clasificadas, solo las características cuyo marcador de tipo cáncer positivo (como en la ecuación (3)) coincidió con el marcador positivo en consideración se incluyeron en la clasificación. Para cada una de dichas características, la fracción de muestras de entrenamiento con valor de característica distinto de cero se calculó por separado para los marcadores positivos y negativos. Las características para las cuales esta fracción fue mayor en el marcador positivo se clasificaron por su información mutua con respecto a esa pareja de marcadores de clase.

Las 256 características superiores clasificadas de cada comparación por pares se identificaron y se añadieron al conjunto final de características para cada tipo de cáncer y no cáncer. Para evitar la redundancia, si se seleccionó más de una característica del mismo tipo positivo y región genómica (es decir, para múltiples tipos negativos), solo se retuvo el rango más bajo (más informativo) para su par de tipos de cáncer, rompiendo el valor del nivel más alto. Las características del conjunto final de características de cada muestra (tipo de cáncer y no cáncer) se binarizaron (cualquier valor de característica superior a 0 se fijó en 1, de modo que todas las características fueran 0 o 1).

Entrenamiento del clasificador

A continuación, las muestras de entrenamiento se dividieron en distintos conjuntos de entrenamiento de validación cruzada de 5 pliegues, y se entrenó un clasificador de dos etapas para cada pliegue, en cada caso entrenando en 4/5 de las muestras de entrenamiento y usando las 1/5 restantes para la validación.

En la primera etapa del entrenamiento, se entrenó un modelo de regresión logística binaria (de dos clases) para detectar la presencia de cáncer para discriminar las muestras de cáncer (independientemente del TOO) de no cáncer. Al entrenar este clasificador binario, se asignó un peso de muestra a las muestras macho sin cáncer para contrarrestar el desequilibrio del sexo en el conjunto de entrenamiento. Para cada muestra, el clasificador binario emite una puntuación de predicción que indica la probabilidad de una presencia o ausencia de cáncer.

En la segunda etapa del entrenamiento, un modelo paralelo de regresión logística de múltiples clases para determinar el tejido canceroso de origen fue entrenado con TOO como marcador de destino. Solo se incluyeron las muestras de

cáncer que recibieron una puntuación por encima del percentil 95 de las muestras no cancerosas en el clasificador de primera etapa en el entrenamiento de este clasificador de múltiples clases. Para cada muestra de cáncer usada en el entrenamiento del clasificador de múltiples clases, el clasificador de múltiples clases emite valores de predicción para los tipos de cáncer que se clasifican, donde cada valor de predicción es una probabilidad de que la muestra dada tenga un cierto tipo de cáncer. Por ejemplo, el clasificador de cáncer puede devolver una predicción de cáncer para una muestra de prueba que incluye una puntuación de predicción para el cáncer de mama, una puntuación de predicción para el cáncer de pulmón y/o una puntuación de predicción para ningún cáncer.

Tanto los clasificadores binarios como los multiclase se entrenaron mediante descenso de gradiente estocástico con minilotes y, en cada caso, el entrenamiento se detuvo antes de tiempo cuando el rendimiento en el pliegue de validación (evaluado mediante la pérdida de entropía cruzada) empezó a degradarse. Para predecir en muestras fuera del conjunto de entrenamiento, en cada etapa se promediaron las puntuaciones asignadas por los cinco clasificadores de validación cruzada. Las puntuaciones asignadas a los tipos de cáncer inapropiados para el sexo se fijaron en cero, y los valores restantes se renormalizaron para sumar uno.

Las puntuaciones asignadas a los pliegues de validación dentro del conjunto de entrenamiento se retuvieron para su uso en la asignación de valores de corte (umbrales) para apuntar a determinadas métricas de rendimiento. En particular, las puntuaciones de probabilidad asignadas a las muestras de conjuntos de entrenamiento no cancerosas se usaron para definir umbrales correspondientes a niveles de especificidad particulares. Por ejemplo, para una diana de especificidad deseada del 99,4 %, el umbral se estableció en el percentil 99,4º de las puntuaciones de probabilidad de detección del cáncer de validación cruzada asignadas a las muestras no cancerosas en el conjunto de entrenamiento. Las muestras de entrenamiento con una puntuación de probabilidad que excedió un umbral se denominaron positiva para el cáncer.

Posteriormente, para cada muestra de entrenamiento determinada para ser positiva para el cáncer, se realizó una evaluación de tipo TOO o cáncer a partir del clasificador de múltiples clases. Primero, el clasificador de regresión logística de múltiples clases asignó un conjunto de puntuaciones de probabilidad, uno para cada tipo de cáncer prospectivo, a cada muestra. A continuación, la confianza de estas puntuaciones se evaluó como la diferencia entre las puntuaciones más altas y segundas más altas asignadas por el clasificador de múltiples clases para cada muestra. Luego, las puntuaciones del conjunto de entrenamiento validado en cruz se usaron para identificar el valor umbral más bajo de manera que de las muestras de cáncer en el conjunto de entrenamiento con un diferencial de puntuación superior-dos que exceda el umbral, se le ha asignado el 90 % del marcador TOO correcto como su puntuación más alta. De esta manera, las puntuaciones asignadas a los pliegues de validación durante el entrenamiento se usaron además para determinar un segundo umbral para distinguir entre las llamadas TOO de confianza e indeterminada.

En el tiempo de predicción, se asignaron muestras que reciben una puntuación del clasificador binario (primera etapa) por debajo del umbral de especificidad predefinido un marcador "no canceroso". Para las muestras restantes, aquellas cuyo diferencial de puntuación de TOO superior del clasificador de segunda etapa estaba por debajo del segundo umbral predefinido se asignaron el marcador de "cáncer indeterminado". Se asignaron las muestras restantes a la etiqueta de cáncer a la que el clasificador TOO asignó la puntuación más alta.

Ejemplo 4 - Clasificación con las regiones genómicas diana de las listas 2-4

El valor discriminatorio de las regiones genómicas diana de las listas 2-4 se evaluó probando la capacidad de un clasificador de cáncer para detectar 3 trastornos hematológicos diferentes según el estado de metilación de estas regiones genómicas diana. El rendimiento se evaluó sobre un conjunto de 1.532 muestras de cáncer y 1.521 muestras no cancerosas que no se usaron para entrenar el clasificador, como se muestra en la **TABLA 2.** Para cada muestra, el ADNlc metilado diferencialmente se enriqueció usando un conjunto de cebos que comprendía todas las regiones genómicas diana de las listas 1-8. El clasificador se restringió entonces para proporcionar determinaciones de cáncer basándose solo en el estado de metilación de las regiones genómicas diana de la lista que se está evaluando.

Tabla 2

Diagnósticos de cáncer de los individuos cuyo ADNlc se usó para validar el clasificador						
Tipo de cáncer	Total	Estadio				
		I	II	III	IV	No informado
Sin cáncer	1521	-	-	-	-	-
Pulmón	261	60	23	72	106	0
Mama	247	102	110	27	8	0
Próstata	188	39	113	19	17	0
Neoplasia linfoide	147	15	27	27	39	39

Colorrectal	121	13	22	41	45	0
Páncreas y vesícula biliar	95	15	15	19	46	0
De útero	84	73	3	5	3	0
Tracto gastrointestinal superior	67	9	12	19	27	0
Cabeza y cuello	62	7	13	16	26	0
Renal	56	37	4	4	11	0
Ovario	37	4	2	25	6	0
Mieloma múltiple	34	10	13	11	0	0
No notificado	29	8	5	7	6	3
Conducto biliar hepático	29	5	7	7	10	0
Sarcoma	17	2	4	5	6	0
Vejiga y urotelial	16	6	7	3	1	0
Anorrectal	14	4	5	5	0	0
De cuello uterino	11	8	1	2	0	0
Melanoma	7	3	1	0	3	0
Neoplasia mieloide	4	2	1	0	1	0
Tiroides	4	0	0	0	0	4
Sólo predicción	2	0	0	0	2	0

Los resultados del análisis de rendimiento del clasificador para listas 2-4 se presentan en las **TABLAS 2-3.** La **TABLA 2** muestra la precisión de la determinación de un trastorno hematológico mediante un clasificador teniendo en cuenta el estado de metilación de las regiones genómicas diana de las listas 2, 3 o 4. La **TABLA 3** muestra la sensibilidad con una especificidad de 0,990 para detectar diferentes estadios de los tres trastornos hematológicos mediante un clasificador que utiliza solo los marcadores de metilación de la lista correspondiente.

Tabla 3

Precisión de clasificación del trastorno hematológico usando las regiones genómicas de las listas 2-4.						
	Lista 2 (Neoplasia linfoide)		Lista 3 (Mieloma múltiple)		Lista 4 (Neoplasia mieloide)	
	%	Fxn	%	Fxn	%	Fxn
Neoplasia linfoide	88	93/106	95	54/57	94	87/93
Mieloma múltiple	100	9/9	89	25/28	100	21/21
Neoplasia mieloide	n/a	0/0	n/a	0/0	100	2/2

Tabla 4

Sensibilidad de la clasificación de los trastornos hematológicos usando las regiones genómicas de las listas 2-4			
Estadio	Neoplasia linfoide (Lista 2)	Mieloma múltiple (Lista 3)	Neoplasia mieloide (Lista 4)
I	33,3 % [11,8-61,6] (5/15)	70 % [34,8-93,3] (7/10)	n.d.
II	92,6 % [75,7-99,1] (25/27)	84,6 % [54,6-98,1] (11/13)	n.d.
III	74,1 % [53,7-88,9] (20/27)	100 % [71,5-100] (11/11)	n.d.
IV	82,1 % [66,5-92,5] (32/39)	100 % [71,5-100] (11/11)	n.d.
Todos	71,4 % [63,4-78,6] (105/147)	85,3 % [68,9-95] (29/34)	75 % [19,4-99,4] (3/4)

Ejemplo 5 - Clasificación con las regiones genómicas diana de las listas 2-4 y 8

Los resultados del análisis del rendimiento del clasificador para la lista 8 y los resultados adicionales para las listas 2-4 y 8 se presentan en las **TABLAS 5-8**. En la **Figura 10** se muestra una curva del operador receptor (ROC) ilustrativa generada por un clasificador entrenado. La ROC muestra resultados positivos verdaderos y resultados falsos positivos para la determinación de la presencia de cáncer o ausencia de cáncer basándose en el estado de metilación de un

subconjunto del 50 % seleccionado al azar de las regiones genómicas diana de la lista 8. La forma asimétrica de la curva ROC ilustra que el clasificador estaba diseñado para minimizar resultados falsos positivos. Las áreas bajo la curva están estrechamente agrupadas entre 0,76 y 0,79, como se muestra en la **TABLA 5.** Estos resultados indican que se puede realizar una determinación del cáncer basándose únicamente en el estado de metilación de las regiones genómicas diana seleccionadas para la discriminación de trastornos hematológicos o incluso trastornos hematológicos individuales. Además, el rendimiento de paneles pequeños de <500 kb indica que paneles de este tamaño son suficientes para una detección precisa del cáncer.

Tabla 5

Detección del cáncer y determinación del tipo de cáncer usando datos para listas de regiones genómicas diana optimizadas para la detección de trastornos hematológicos.				
Regiones genómicas	AUC	Verdadero positivo	Falso positivo	Falso negativo
Lista 2	0,76	103	12	1
Lista 3	0,76	79	6	0
Lista 4	0,78	110	6	0
El 25 % aleatorio de la lista 8	0,78	101	7	0
El 50 % aleatorio de la lista 8	0,79	106	6	0

Una vez que se determina el cáncer, el clasificador asigna el cáncer a uno de los veinte tipos de cáncer distintos. La precisión de estas determinaciones con una especificidad de 0,990 se presenta en varios formatos. La **TABLA 5** muestra los verdaderos positivos, falsos positivos y falsos negativos clasificados basándose en el estado de metilación de las listas de regiones genómicas diana optimizadas para la detección de trastornos hematológicos específicos o subconjuntos aleatorios de una lista optimizada para la detección de todos los trastornos hematológicos. Un verdadero positivo se produce cuando se detecta la presencia de cáncer y el clasificador determina con precisión que la muestra proviene de un sujeto con un trastorno hematológico. Se produce un falso positivo para las muestras de individuos diagnosticados con un tumor sólido cuando se detecta la presencia de cáncer y se determina de forma no precisa que es un trastorno hematológico. Se produce un falso negativo cuando la muestra proviene de un individuo diagnosticado con un trastorno hematológico, pero el clasificador determina incorrectamente que la muestra proviene de un individuo con un tumor sólido. Los falsos negativos fueron muy raros en las listas 2-4 y 8. Aproximadamente el 5-10 % de las muestras fueron falsos negativos. Esto puede ocurrir porque las listas 2-4 y 8 no incluyen algunos marcadores que ayudarían a determinar con precisión si un cáncer fue un tumor sólido.

La precisión de la detección del cáncer basada en el estado de metilación de las regiones genómicas diana en las listas 2-4 y 8 se evalúa para varios estadios del cáncer en la **TABLA 6**. Cuando se detecta un cáncer, se asigna un tipo de cáncer de una de las veinte clases posibles de tipos de cáncer. La precisión de la determinación del tipo de cáncer se presenta en la **TABLA 7**. Los resultados de la determinación del tipo de cáncer son para la precisión de la determinación de los veinte tipos de cáncer, a pesar de que las listas de regiones genómicas diana se optimizaron para detectar trastornos hematológicos.

Los resultados de las **TABLAS 6-7** están segregados para varios estadios del cáncer. La detección del cáncer y la determinación del tipo de cáncer fueron más precisas en las muestras de individuos diagnosticados con estadios posteriores del cáncer. Esto era de esperarse porque los tumores en estadio tardío cambian más ADNlc. Sin embargo, la precisión de la detección del cáncer y la asignación de un tipo de cáncer para los cánceres en estadio temprano es notablemente alta. Además, la precisión de la clasificación fue razonablemente precisa con solo el 50 % o incluso el 25 % de las regiones genómicas diana de la lista 8 (todos los trastornos hematológicos).

La sensibilidad con una especificidad de 0,990 para detectar trastornos hematológicos en los estadios I - IV mediante un clasificador que actúa sobre el estado de metilación de las regiones genómicas diana de las listas 2-4 o subconjuntos aleatorios de la lista 8 se presenta en la **TABLA 8.** Por ejemplo, cuando la tasa de falsos positivos para detectar el cáncer se limita al 1 %, un clasificador que tenga en cuenta el estado de metilación de las regiones genómicas diana de la lista 3 (optimizadas para el mieloma múltiple) detectó mieloma múltiple en el 70 % (7 de cada 10) de las muestras recolectadas de individuos diagnosticados con mieloma múltiple en estadio I. Del mismo modo, cuando la tasa de falsos positivos para detectar el cáncer se limita al 1 %, un clasificador que tenga en cuenta el estado de metilación de las regiones genómicas diana de la lista 2 (optimizadas para la neoplasia linfoide) detectó neoplasia linfoide en el 93 % (25 de 27) de las muestras recolectadas de individuos diagnosticados con neoplasia linfoide en estadio II. Además, la sensibilidad para HD basándose en el estado de metilación de subconjuntos aleatorios del 50 % y el 25 % de las regiones genómicas diana de la lista 8 fue esencialmente idéntica (con la excepción de la neoplasia linfoide en estadio I), lo que indica que una fracción sustancial de las regiones genómicas diana de la lista 8 contribuyen a determinaciones de HD con precisión por el clasificador.

Tabla 6

Precisión de detección del cáncer con una especificidad del 99,0 % por un clasificador que tenga en cuenta el estado de metilación de las regiones genómicas diana de la lista indicada.										
	Lista 2 (Neoplasia linfoide)		Lista 3 (Mieloma múltiple)		Lista 4 (Neoplasia mieloide)		El 25 % aleatorio de la lista 8 (todo HD)		El 50 % aleatorio de la lista 8 (todo HD)	
Estadio	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	39	603/1532	36	551/1532	43	663/1532	41	622/1532	43	655/1532
I	7	30/422	8	34/422	11	45/422	8	35/422	9	39/422
II	28	107/388	25	98/388	31	119/388	27	105/388	29	111/388
III	53	167/313	51	159/313	60	188/313	58	182/313	60	187/313
I+II	17	137/810	16	132/810	20	164/810	17	140/810	19	150/810
II+II+III	27	304/1123	26	291/1123	31	352/1123	29	322/1123	30	337/1123
III+IV	65	442/676	61	409/676	71	477/676	68	460/676	71	481/676
IV	76	275/363	69	250/363	80	289/363	77	278/363	81	294/363

Tabla 7

Precisión de las determinaciones del tipo de cáncer con una especificidad del 99,0 % mediante un clasificador que tenga en cuenta el estado de metilación de las regiones genómicas diana de la lista indicada.										
	Lista 2 (Neoplasia linfoide)		Lista 3 (Mieloma múltiple)		Lista 4 (Neoplasia mieloide)		El 25 % aleatorio de la lista 8 (todo HD)		El 50 % aleatorio de la lista 8 (todo HD)	
Estadio	%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Todos	90	427/477	90	322/358	90	496/553	89	420/470	90	495/551
I	74	14/19	74	14/19	75	21/28	71	10/14	82	18/22
II	88	73/83	92	60/65	91	90/99	89	69/78	90	81/90
III	89	113/127	90	96/107	88	137/155	89	122/137	89	140/157
I+II	85	87/102	88	74/84	87	111/127	86	79/92	88	99/112
II+II+III	87	200/229	89	170/191	88	248/282	88	201/229	89	239/269
III+IV	90	319/354	90	245/271	90	366/407	90	321/358	90	374/417
IV	91	206/227	91	149/164	91	229/252	90	199/221	90	234/260

Tabla 8

Sensibilidad con una especificidad del 99,0 % para el trastorno hematológico indicado mediante un clasificador usando solo las regiones genómicas diana de la lista indicada.											
Trastorno hematológico	Estadio	Lista 2 (Neoplasia linfoide)		Lista 3 (Mieloma múltiple)		Lista 4 (Neoplasia mieloide)		El 25 % aleatorio de la lista 8 (todo HD		El 50 % aleatorio de la lista 8 (todo HD)	
		%	Fxn	%	Fxn	%	Fxn	%	Fxn	%	Fxn
Mieloma múltiple	Todos	38	13/34	85	29/34	68	23/34	56	19/34	56	19/34
	I	10	1/10	70	7/10	40	4/10	30	3/10	30	3/10
	II	23	3/13	85	11/13	62	8/13	39	5/13	39	5/13
	III	82	9/11	100	11/11	100	11/11	100	11/11	100	11/11
	I+II	17	4/23	78	18/23	52	12/23	35	8/23	35	8/23
	II+II+III	38	13/34	85	29/34	68	23/34	56	19/34	56	19/34
	Todos	71	105/147	46	67/147	65	96/147	61	89/147	65	96/147

Neoplasia linfoide	I	33	5/15	13	2/15	33	5/15	13	2/15	27	4/15
	II	93	25/27	52	14/27	78	21/27	67	18/27	67	18/27
	III	74	20/27	74	20/27	74	20/27	70	19/27	70	19/27
	I+II	71	30/42	38	16/42	62	26/42	48	20/42	52	22/42
	II+II+III	73	50/69	52	36/69	67	46/69	57	39/69	59	41/69
	III+IV	79	52/66	67	44/66	77	51/66	71	47/66	76	50/66
	IV	82	32/39	62	24/39	80	31/39	72	28/39	80	31/39
Neoplasia mieloide	Todos	0	0/4	75	3/4	75	3/4	0	0/4	0	0/4

Ejemplo 6 - Detección de trastornos hematológicos usando un panel de ensayo

Se recolectan muestras de sangre de un grupo de individuos previamente diagnosticados con un trastorno hematológico (“grupo de prueba”), y de otros grupos de individuos sin trastorno hematológico o diagnosticados de un tipo diferente de trastorno hematológico (“otro grupo”). Se extraen fragmentos de ADNlc de las muestras de sangre y se tratan con bisulfito para convertir las citosinas no metiladas en uracilos. A las muestras tratadas con bisulfito se les aplicó el panel de ensayos de cáncer descrito en la presente memoria. Los fragmentos de ADNlc no unidos se lavan y se recogen fragmentos de ADNlc unidos a las sondas. Los fragmentos de ADNlc recogidos se amplifican y secuencian. Las lecturas de secuencia confirman que las sondas enriquecen específicamente fragmentos de ADNlc que tienen patrones de metilación indicativos de un trastorno hematológico y muestras del grupo de prueba incluyen significativamente más de los fragmentos de ADNlc metilados diferencialmente en comparación con el otro g

REIVINDICACIONES

1. Un método para detectar un trastorno hematológico (HD) en un sujeto, el método comprende:

(a) obtener lecturas de secuenciación del ADN libre de células (ADNlc) de una muestra de un sujeto, o productos de amplificación de los mismos, enriquecidos por contacto con una composición que comprende una pluralidad de diferentes oligonucleótidos cebo, en donde el ADNlc se convierte antes de la secuenciación mediante el tratamiento del ADNlc para convertir las citosinas no metiladas en uracilos, en donde:

i) cada oligonucleótido cebo en la pluralidad de oligonucleótidos cebo diferentes tiene al menos 45 nucleótidos de longitud, opcionalmente 45 a 300 nucleótidos de longitud;

ii) la pluralidad de oligonucleótidos cebo diferentes se hibrida colectivamente con al menos 100 regiones genómicas diana;

iii) las al menos 100 regiones genómicas diana están metiladas diferencialmente en al menos un HD en relación con un HD diferente;

iv) el al menos un HD y el HD diferente se seleccionan entre hematopoyesis clonal de potencial indeterminado (CHIP), leucemia, neoplasias linfoides, mieloma múltiple y una neoplasia mieloide; y

(b) usar un ordenador para aplicar un clasificador entrenado a las lecturas de secuenciación para determinar una presencia o ausencia del al menos un HD, en donde el clasificador se entrena usando secuencias de ADN convertido, en donde el ADN convertido se refiere al ADN que se ha tratado para convertir las citosinas no metiladas en uracilos, en donde:

i) el clasificador entrenado distingue entre el CHIP y uno o más HD seleccionados entre leucemia, neoplasias linfoides, mieloma múltiple y una neoplasia mieloide basándose en lecturas de secuenciación; y

ii) la detección de una serie de lecturas de secuenciación para una pluralidad de las al menos 100 regiones genómicas diana por encima de un umbral identifica la presencia del al menos un HD.

2. El método de la reivindicación 1, en donde las al menos 100 regiones genómicas diana comprenden al menos el 20 % de las regiones genómicas diana de una cualquiera de las listas 1-8, o complementos de las mismas.

3. El método de la reivindicación 1, en donde las al menos 100 regiones genómicas diana comprenden al menos el 20 % de las regiones genómicas diana de las listas 1-8, o complementos de las mismas.

4. El método de la reivindicación 1, en donde:

a) las al menos 100 regiones genómicas diana comprenden al menos el 20 % de las regiones genómicas diana de las listas 1 u 8, o complementos de las mismas;

b) al menos 100 regiones genómicas diana comprenden al menos el 20 % de las regiones genómicas diana de las listas 2-4, o complementos de las mismas; o

c) al menos 100 regiones genómicas diana comprenden al menos el 20 % de las regiones genómicas diana de las listas 5-7, o complementos de las mismas

5. El método de la reivindicación 1, en donde el tamaño total de las al menos 100 regiones genómicas diana es menor que 2000 kb.

6. El método de la reivindicación 1, que comprende además hibridar fragmentos de ADN libre de células (ADNlc) convertido o amplicones de los mismos con la pluralidad de diferentes oligonucleótidos cebo.

7. El método de una cualquiera de la reivindicaciones 1-6, en donde la pluralidad de oligonucleótidos cebo diferentes comprende una pluralidad de conjuntos de dos o más oligonucleótidos cebo, en donde cada oligonucleótido cebo dentro de un conjunto de oligonucleótidos cebo está configurado para unirse a moléculas de ADN convertido derivado de la misma región genómica diana.

8. El método de la reivindicación 7, en donde:

a) la pluralidad de diferentes oligonucleótidos cebo comprende pares de oligonucleótidos cebo;

b) cada par de oligonucleótidos cebo comprende un primer oligonucleótido cebo y un segundo oligonucleótido cebo;

c) cada oligonucleótido cebo comprende un extremo 5' y un extremo 3';

d) para cada par de oligonucleótidos cebo, una secuencia de al menos X bases de nucleótido en el extremo 3' del primer oligonucleótido cebo es idéntica a una secuencia de X bases de nucleótido en el extremo 5' del segundo oligonucleótido cebo; y

e) X es al menos 25, 30, 35, 40, 45, 50, 60, 70, 75 o 100;

opcionalmente, en donde, para cada par de oligonucleótidos cebo, el primer oligonucleótido cebo comprende una secuencia de al menos 31, 40, 50 o 60 bases de nucleótidos que no se superpone a una secuencia del segundo oligonucleótido cebo.

9. El método de la reivindicación 1, que comprende además

a) capturar fragmentos de ADN libre de células (ADNlc) del sujeto o productos de amplificación de los mismos con una composición que comprende la pluralidad de diferentes oligonucleótidos cebo, opcionalmente en donde la captura comprende separar el ADN unido al cebo del ADN no unido al cebo; y

b) secuenciar los fragmentos de ADNlc capturados o productos de amplificación de los mismos para producir lecturas de secuenciación.

10. El método de la reivindicación 1 o 9, en donde el clasificador entrenado es un clasificador modelo de mezcla.

11. El método de la reivindicación 1 o 9, en donde el clasificador se entrenó en secuencias de ADN convertido derivadas de al menos 100 regiones genómicas diana.

12. El método de la reivindicación 11, en donde el al menos un HD y el HD diferente son tipos de cáncer, y en donde el clasificador entrenado detecta células del al menos un HD:

generando un conjunto de características para la muestra, en donde cada característica del conjunto de características comprende un valor numérico;

introduciendo el conjunto de características en el clasificador, en donde el clasificador comprende un clasificador multinomial;

basándose en el conjunto de características, determinando, en el clasificador, un conjunto de puntuaciones de probabilidad, en donde el conjunto de puntuaciones de probabilidad comprende una puntuación de probabilidad por clase de tipo de cáncer y por clase de tipo distinto de cáncer; y

estableciendo un umbral del conjunto de puntuaciones de probabilidad basándose en uno o más valores determinados durante el entrenamiento del clasificador para detectar el ADNlc en la muestra de células del al menos un HD.

13. El método de la reivindicación 12, en donde

a) el conjunto de características comprende un conjunto de características binarizadas;

b) el valor numérico comprende un único valor binario;

c) el clasificador multinomial comprende un conjunto de regresión logística multinomial entrenado para predecir un tejido fuente para el cáncer; o

d) el método comprende además determinar una clasificación final del cáncer basándose en un diferencial de puntuación de las dos probabilidades superiores en relación con un valor mínimo, en donde el valor mínimo corresponde a un porcentaje predefinido de muestras de cáncer de entrenamiento que habían sido asignadas al tipo de cáncer correcto como su puntuación más alta durante el entrenamiento del clasificador.

14. El método de la reivindicación 1 o 9, en donde el al menos un HD es un tipo de cáncer, y en donde las células del al menos un HD se detectan con una especificidad de al menos 0,990; opcionalmente en donde:

a) la relación entre la probabilidad de determinar con precisión un trastorno hematológico y la probabilidad de determinar de forma incorrecta un tumor sólido es de al menos 25:1 o al menos 50:1;

b) la relación entre la probabilidad de determinar con precisión un trastorno hematológico y la probabilidad de determinar de forma incorrecta un trastorno hematológico es de al menos 8:1, al menos 12:1, o al menos 16:1;

c) la probabilidad de determinar con precisión un tipo de cáncer es de al menos el 80 %, al menos el 85 % o al menos el 89 %;

d) el al menos un HD es un cáncer en estadio I y la probabilidad de determinar con precisión un tipo de cáncer es de al menos el 65 %;

e) el al menos un HD es un cáncer en estadio II y la probabilidad de determinar con precisión un tipo de cáncer es de al menos el 75 %; o

f) el al menos un HD es un cáncer en estadio III y la probabilidad de determinar con precisión un tipo de cáncer es de al menos el 85 %.

15. El método de la reivindicación 1 o 9, en donde

a) las al menos 100 regiones genómicas diana se seleccionan de regiones genómicas diana de la lista 3 o la lista 6, o complementos de las mismas;

b) las al menos 100 regiones genómicas diana se seleccionan de regiones genómicas diana de la lista 2 o la lista 5, o complementos de las mismas;

c) las al menos 100 regiones genómicas diana se seleccionan de las listas 1-8, o complementos de las mismas; o

d) las al menos 100 regiones genómicas comprenden al menos el 20 % de las regiones genómicas diana de una cualquiera de las listas 1-8, o complementos de las mismas.

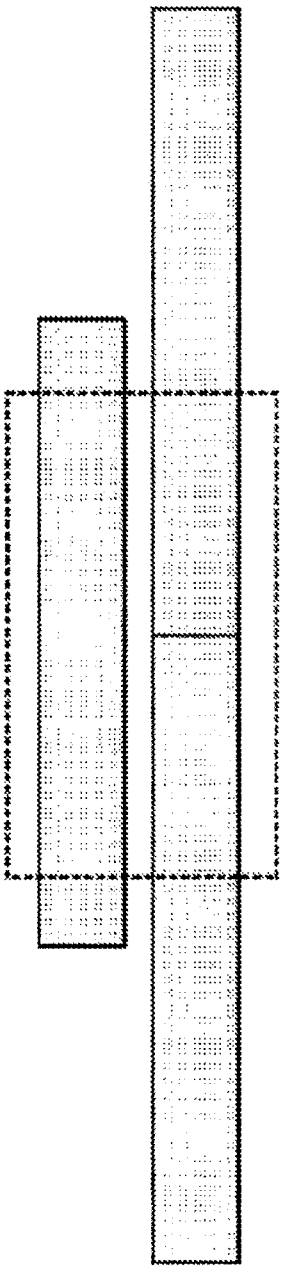

Figura 1A

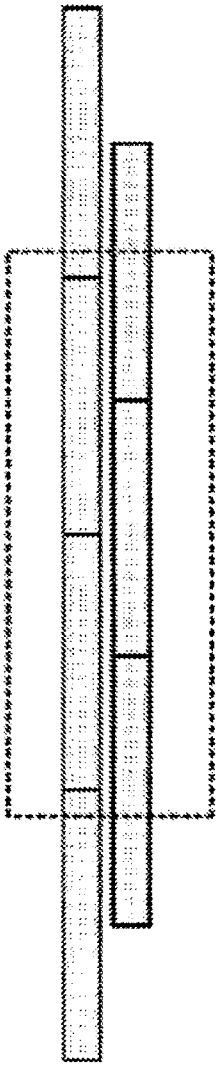

Figura 1B

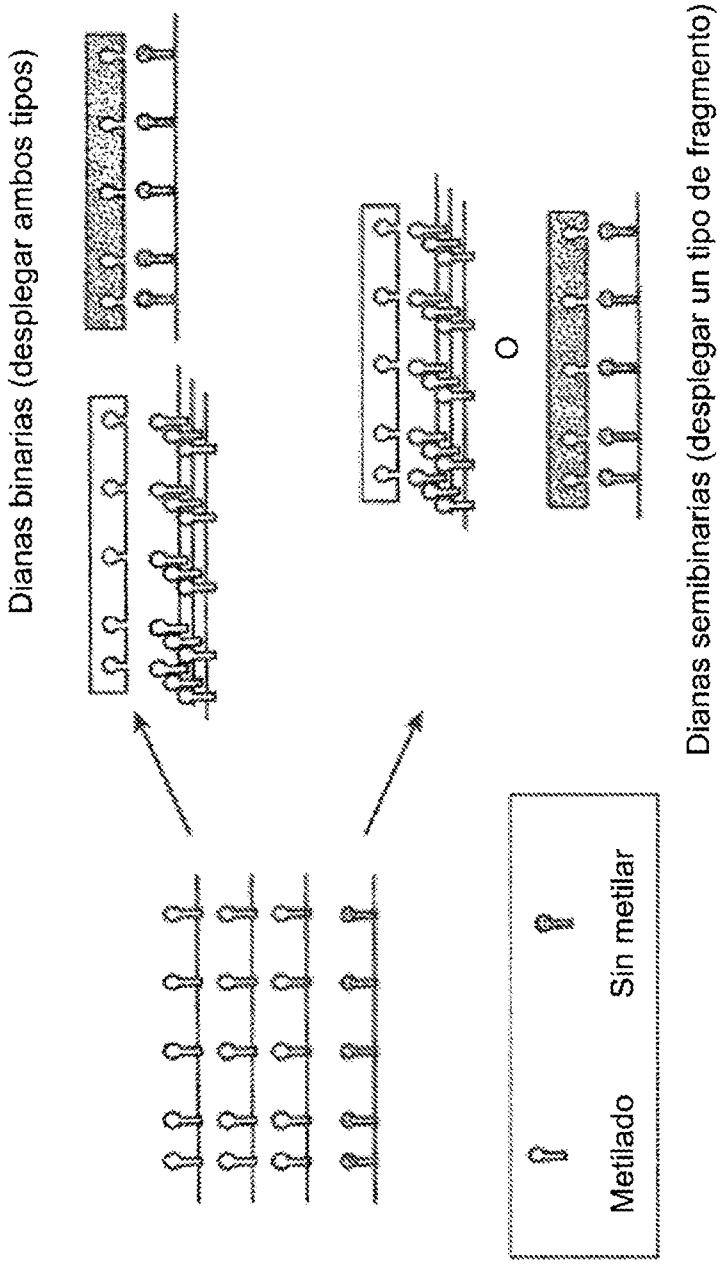
Dianas binarias (desplegar ambos tipos)
O
Dianas semibinarias (desplegar un tipo de fragmento)
Metilado
Sin metilar

Figura 1C

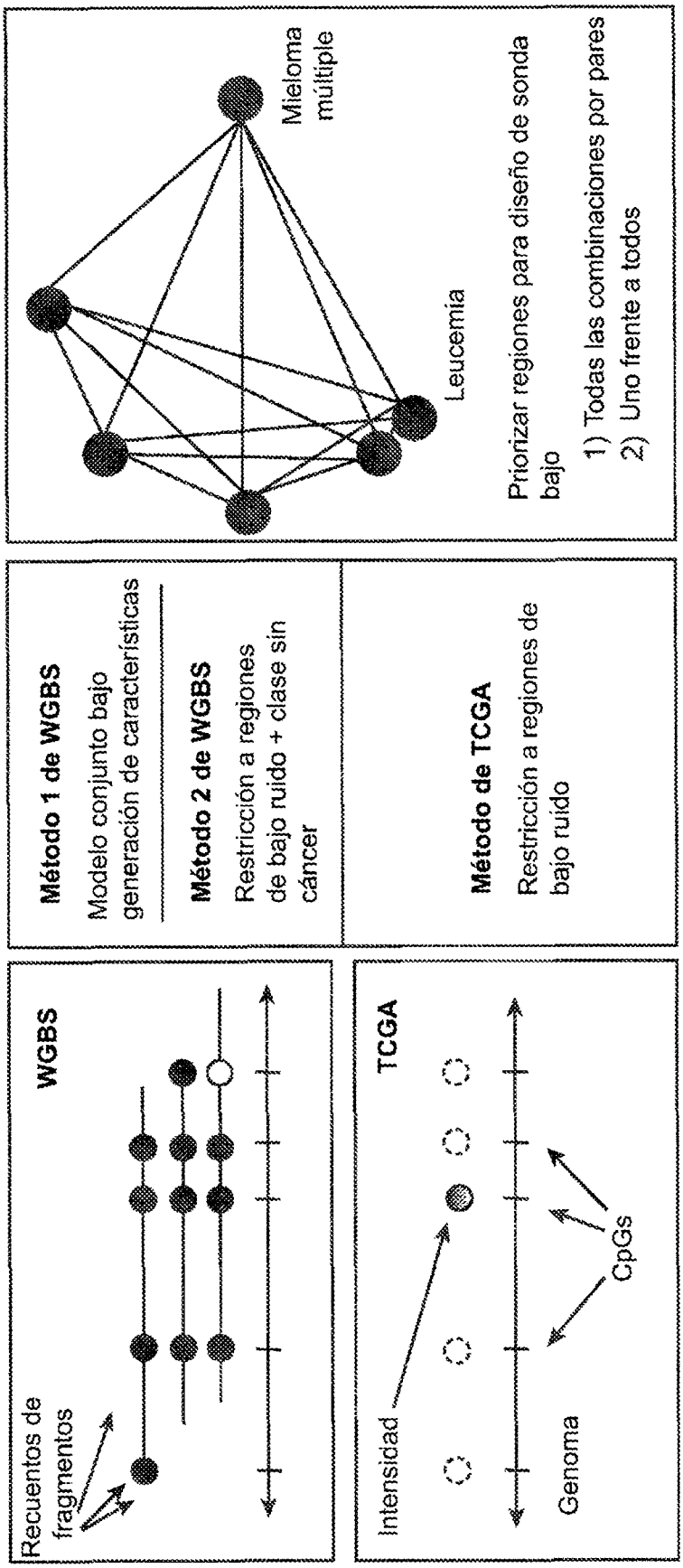
Características
Relacionar con ADNlc
Selección de región
WGBS
Recuentos de fragmentos
TCGA
Intensidad
Genoma
CpGs
Método 1 de WGBS
Modelo conjunto bajo generación de características
Método 2 de WGBS
Restricción a regiones de bajo ruido + clase sin cáncer
Método de TCGA
Restricción a regiones de bajo ruido
Mieloma múltiple
Leucemia
Priorizar regiones para diseño de sonda bajo
1) Todas las combinaciones por pares
2) Uno frente a todos

Figura 2

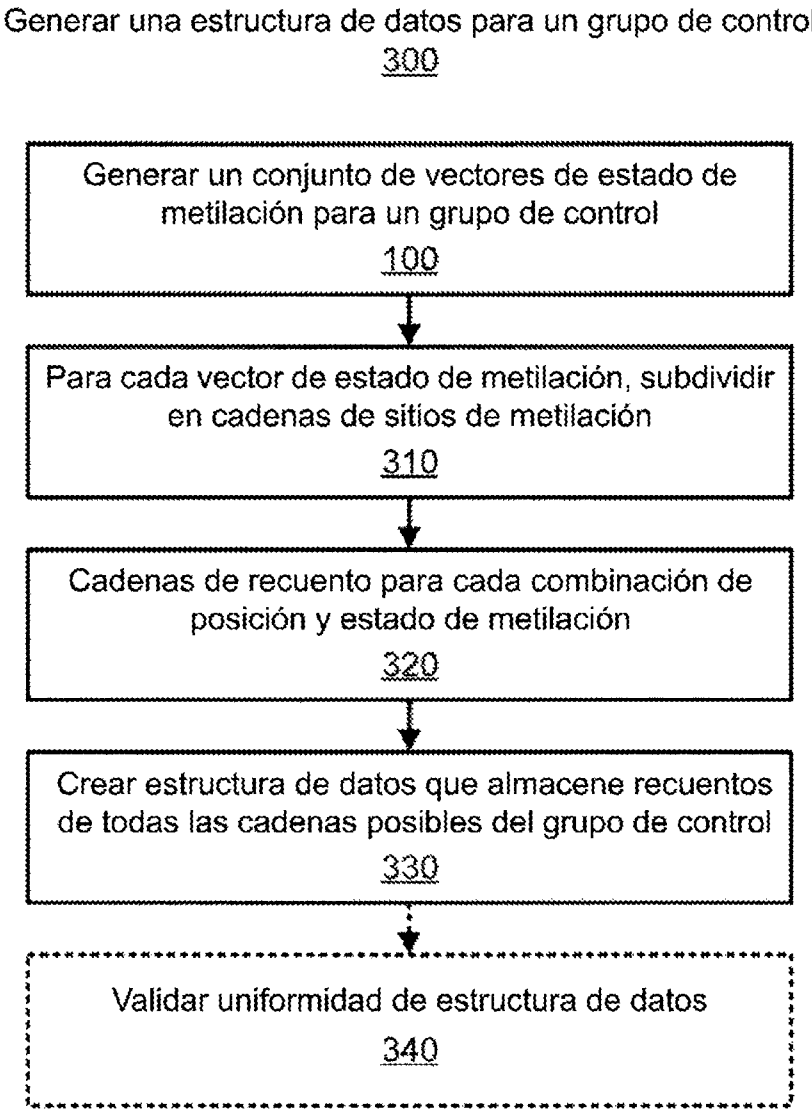
Generar una estructura de datos para un grupo de control
300
Generar un conjunto de vectores de estado de metilación para un grupo de control
100
Para cada vector de estado de metilación, subdividir en cadenas de sitios de metilación
310
Cadenas de recuento para cada combinación de posición y estado de metilación
320
Crear estructura de datos que almacene recuentos de todas las cadenas posibles del grupo de control
330
Validar uniformidad de estructura de datos
340

Figura 3A

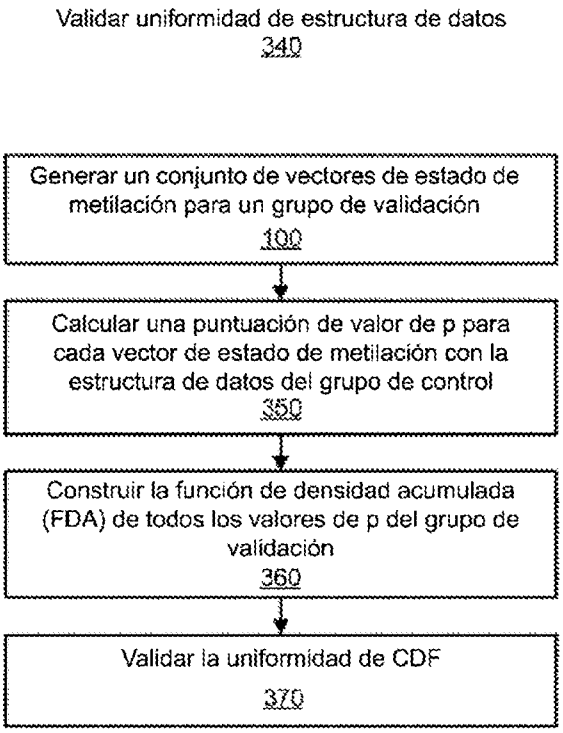
Validar uniformidad de estructura de datos
340
Generar un conjunto de vectores de estado de metilación para un grupo de validación
100
Calcular una puntuación de valor de p para cada vector de estado de metilación con la estructura de datos del grupo de control
350
Construir la función de densidad acumulada (FDA) de todos los valores de p del grupo de validación
360
Validar la uniformidad de CDF
370

Figura 3B

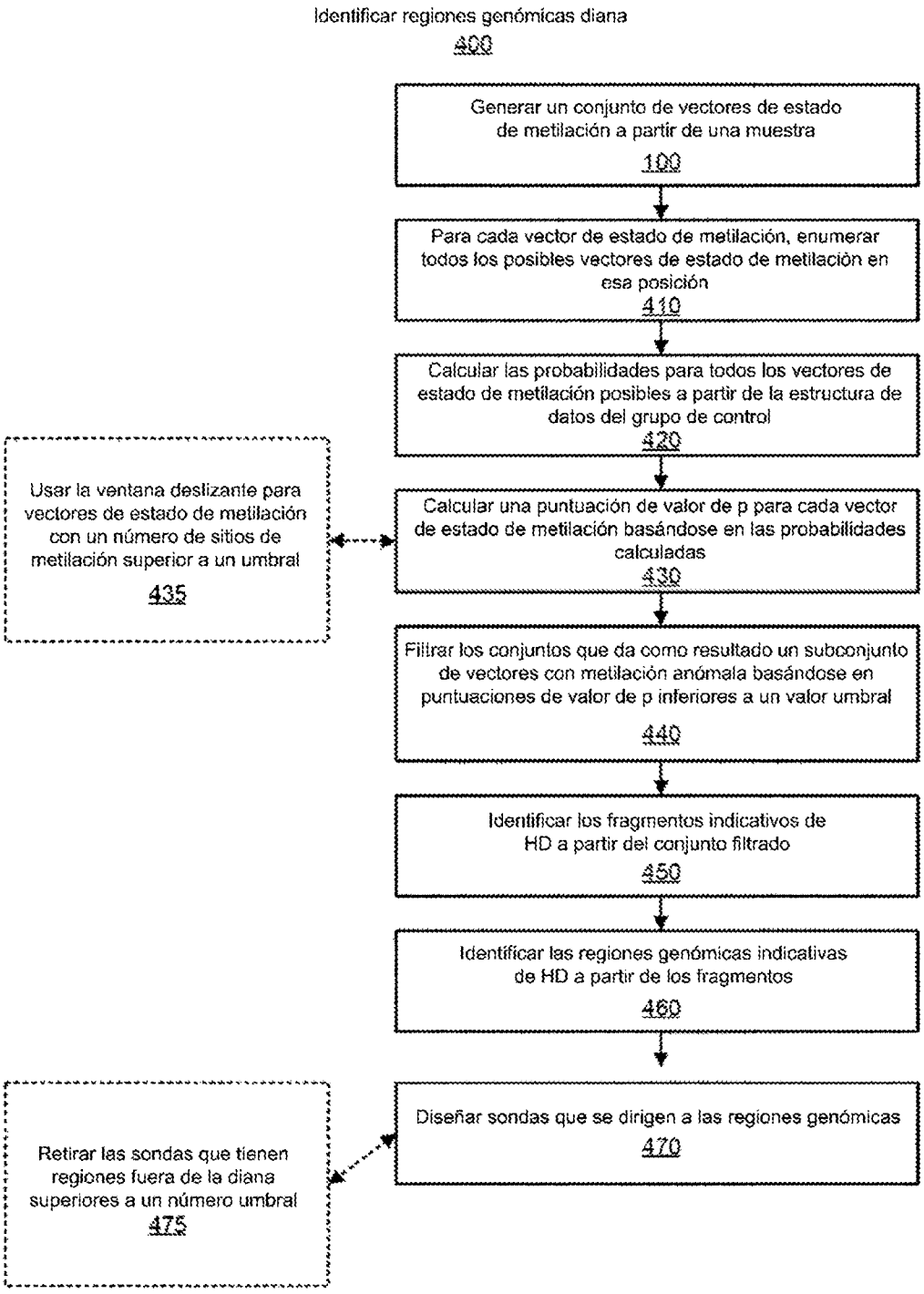
Identificar regiones genómicas diana
400
Generar un conjunto de vectores de estado de metilación a partir de una muestra
100
Para cada vector de estado de metilación, enumerar todos los posibles vectores de estado de metilación en esa posición
410
Calcular las probabilidades para todos los vectores de estado de metilación posibles a partir de la estructura de datos del grupo de control
420
Usar la ventana deslizante para vectores de estado de metilación con un número de sitios de metilación superior a un umbral
435
Calcular una puntuación de valor de p para cada vector de estado de metilación basándose en las probabilidades calculadas
430
Filtrar los conjuntos que da como resultado un subconjunto de vectores con metilación anómala basándose en puntuaciones de valor de p inferiores a un valor umbral
440
Identificar los fragmentos indicativos de HD a partir del conjunto filtrado
450
Identificar las regiones genómicas indicativas de HD a partir de los fragmentos
460
Diseñar sondas que se dirigen a las regiones genómicas
470
Retirar las sondas que tienen regiones fuera de la diana superiores a un número umbral
475

Figura 4

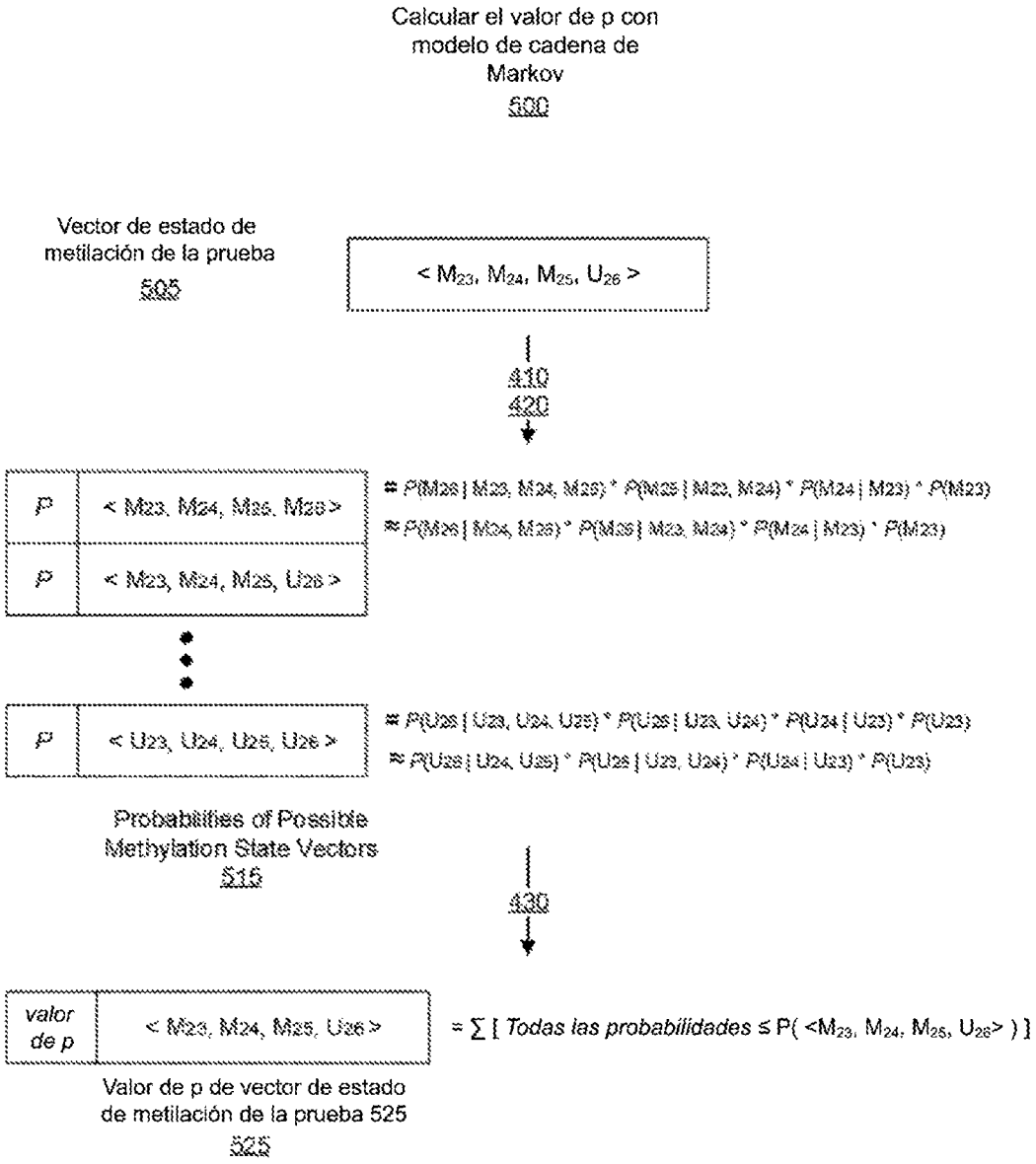
Calcular el valor de p con modelo de cadena de Markov
530
Vector de estado de metilación de la prueba
525
< M23, M24, M25, U26 >
410
420
p
< M23, M24, M25, M26 >
= P(M26 | M23, M24, M25) * P(M25 | M23, M24) * P(M24 | M23) * P(M23)
≈ P(M26 | M24, M25) * P(M25 | M23, M24) * P(M24 | M23) * P(M23)
p
< M23, M24, M25, U26 >
p
< U23, U24, U25, U26 >
= P(U26 | U23, U24, U25) * P(U25 | U23, U24) * P(U24 | U23) * P(U23)
≈ P(U26 | U24, U25) * P(U25 | U23, U24) * P(U24 | U23) * P(U23)
Probabilities of Possible Methylation State Vectors
515
430
valor de p
< M23, M24, M25, U26 >
= ∑ [Todas las probabilidades ≤ P(<M23, M24, M25, U26>)]
Valor de p de vector de estado de metilación de la prueba 525
525

Figura 5

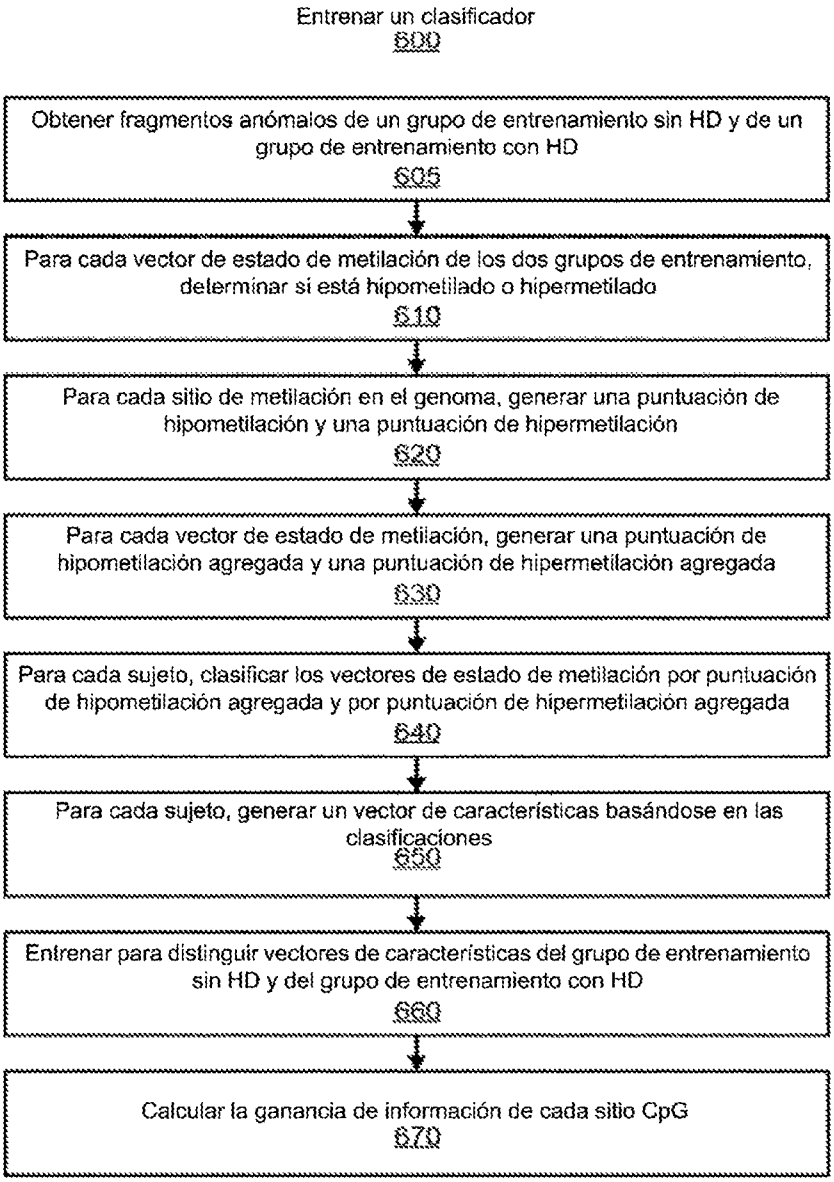
Entrenar un clasificador
600
Obtener fragmentos anómalos de un grupo de entrenamiento sin HD y de un grupo de entrenamiento con HD
605
Para cada vector de estado de metilación de los dos grupos de entrenamiento, determinar si está hipometilado o hipermetilado
610
Para cada sitio de metilación en el genoma, generar una puntuación de hipometilación y una puntuación de hipermetilación
620
Para cada vector de estado de metilación, generar una puntuación de hipometilación agregada y una puntuación de hipermetilación agregada
630
Para cada sujeto, clasificar los vectores de estado de metilación por puntuación de hipometilación agregada y por puntuación de hipermetilación agregada
640
Para cada sujeto, generar un vector de características basándose en las clasificaciones
650
Entrenar para distinguir vectores de características del grupo de entrenamiento sin HD y del grupo de entrenamiento con HD
660
Calcular la ganancia de información de cada sitio CpG
670

Figura 6A

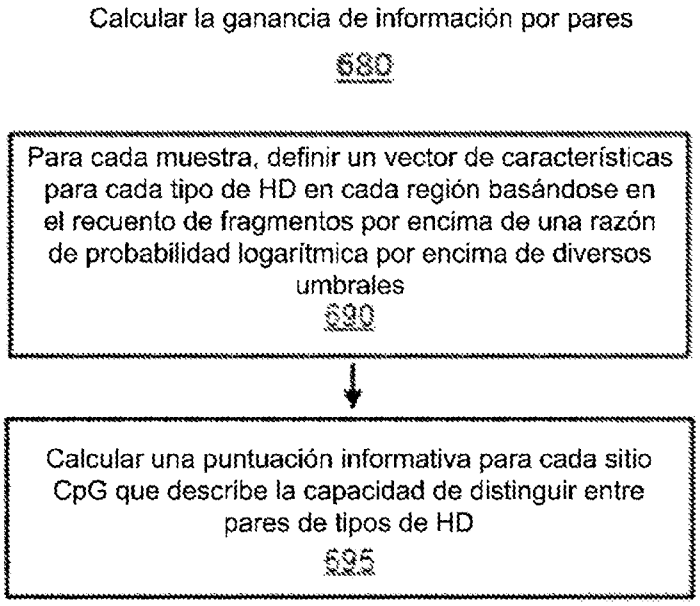
Calcular la ganancia de información por pares
680
Para cada muestra, definir un vector de características para cada tipo de HD en cada región basándose en el recuento de fragmentos por encima de una razón de probabilidad logarítmica por encima de diversos umbrales
690
Calcular una puntuación informativa para cada sitio CpG que describe la capacidad de distinguir entre pares de tipos de HD
695

Figura 6B

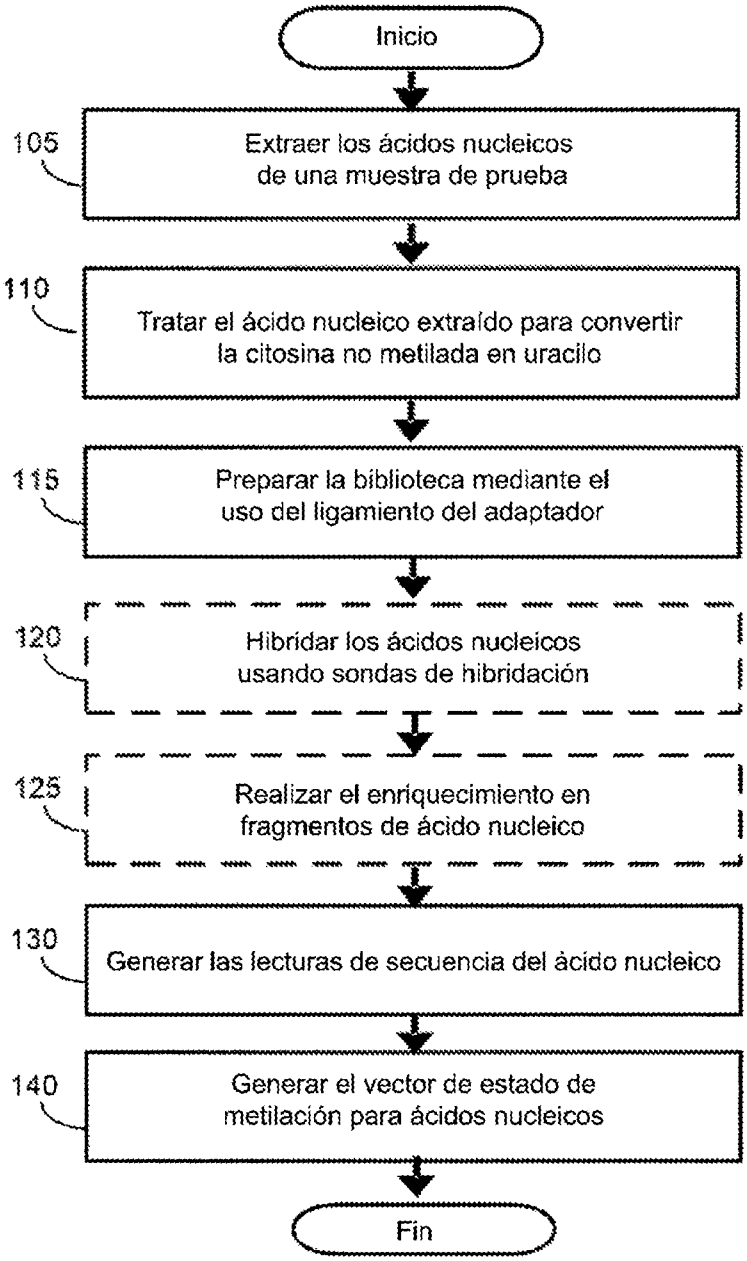
Inicio
105
Extraer los ácidos nucleicos de una muestra de prueba
110
Tratar el ácido nucleico extraído para convertir la citosina no metilada en uracilo
115
Preparar la biblioteca mediante el uso del ligamiento del adaptador
120
Hibridar los ácidos nucleicos usando sondas de hibridación
125
Realizar el enriquecimiento en fragmentos de ácido nucleico
130
Generar las lecturas de secuencia del ácido nucleico
140
Generar el vector de estado de metilación para ácidos nucleicos
Fin

Figura 7A

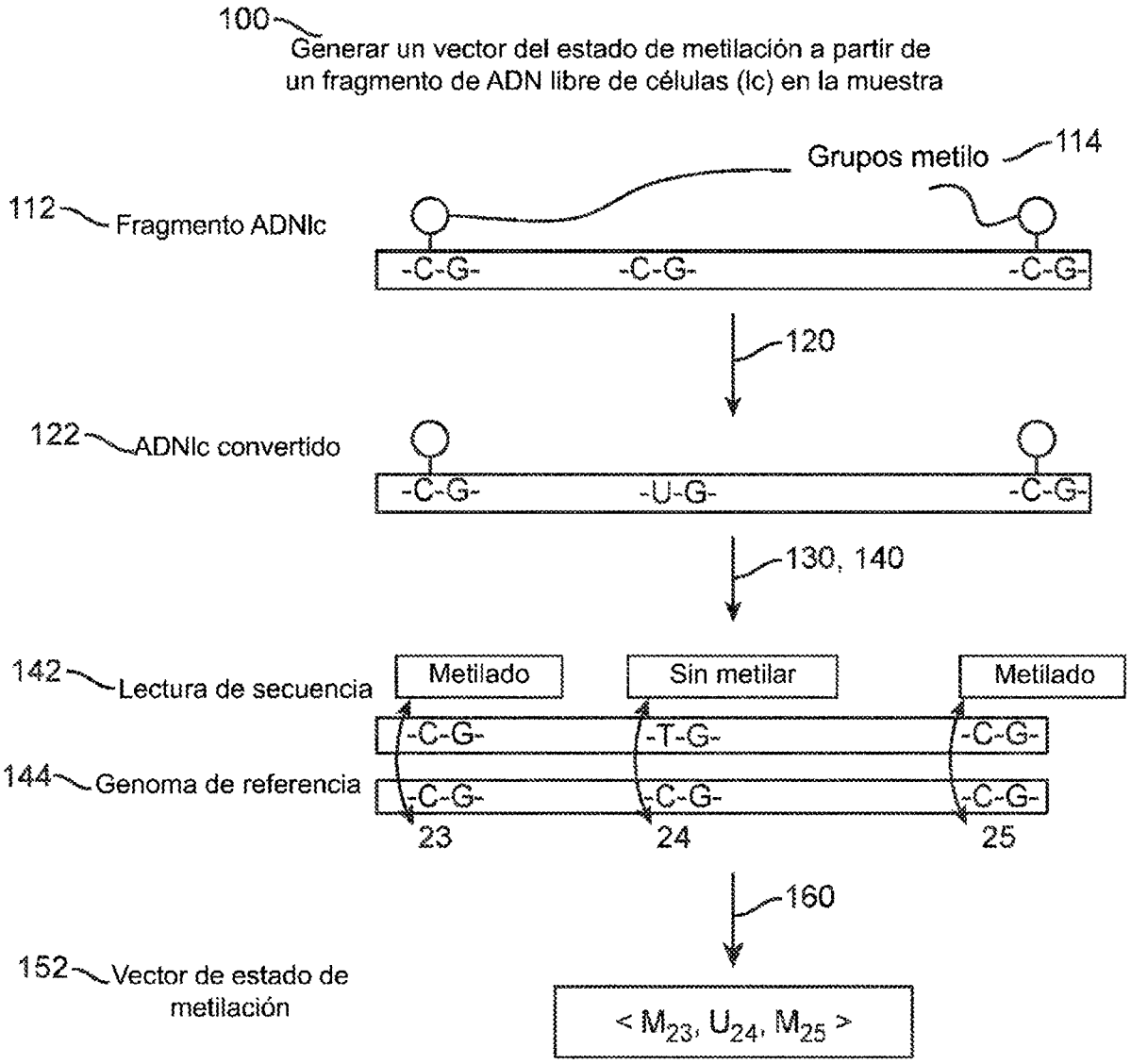
100
Generar un vector del estado de metilación a partir de un fragmento de ADN libre de células (lc) en la muestra
Grupos metilo
114
112
Fragmento ADNlc
-C-G-
-C-G-
-C-G-
120
122
ADNlc convertido
-C-G-
-U-G-
-C-G-
130, 140
142
Lectura de secuencia
Metilado
Sin metilar
Metilado
-C-G-
-T-G-
-C-G-
144
Genoma de referencia
-C-G-
-C-G-
-C-G-
23
24
25
160
152
Vector de estado de metilación
< M23, U24, M25 >

Figura 7B

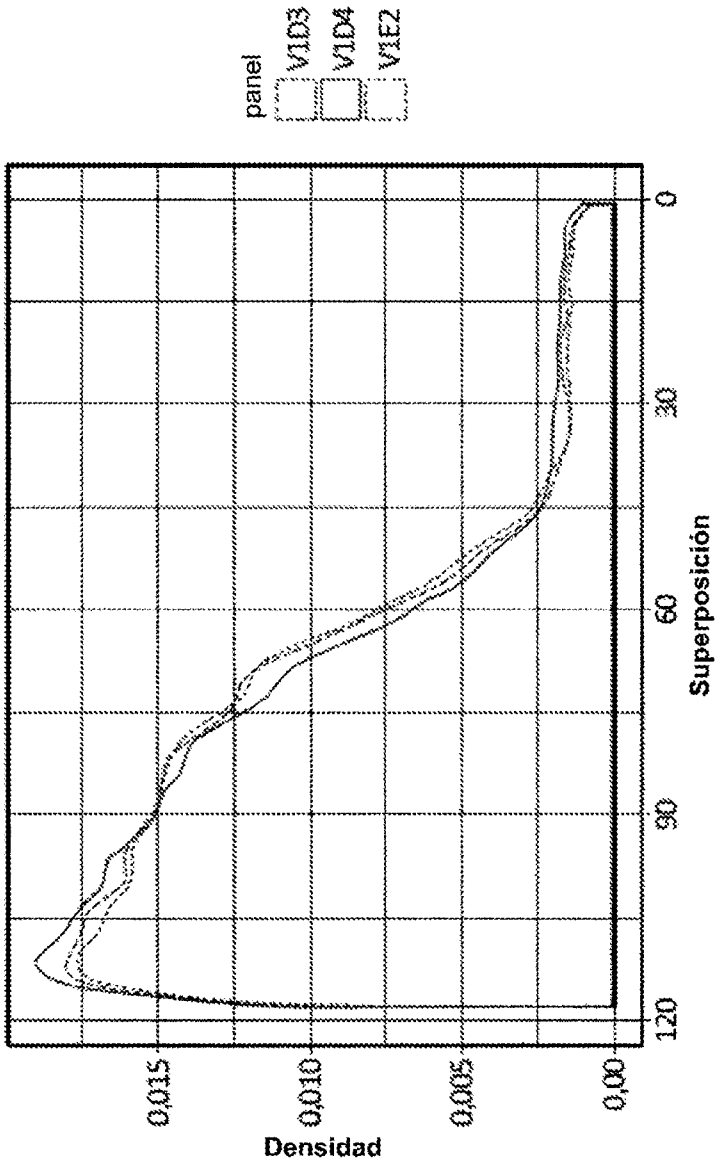
panel
V1D3
V1D4
V1E2
0,015
0,010
0,005
0,00
Densidad
0
30
60
90
120
Superposición

Figura 8

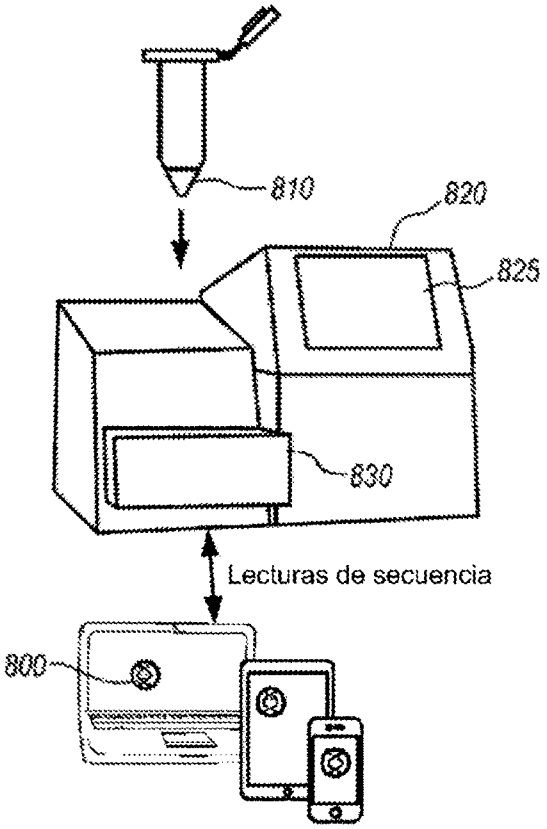

Figura 9A

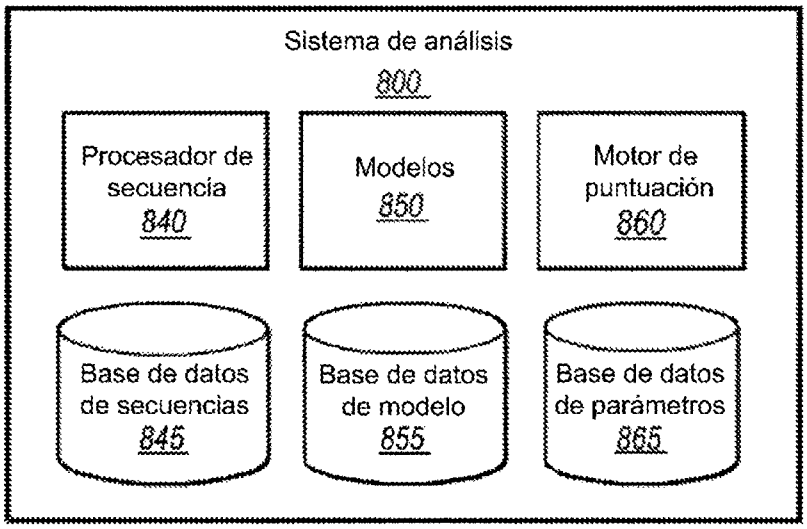

Figura 9B

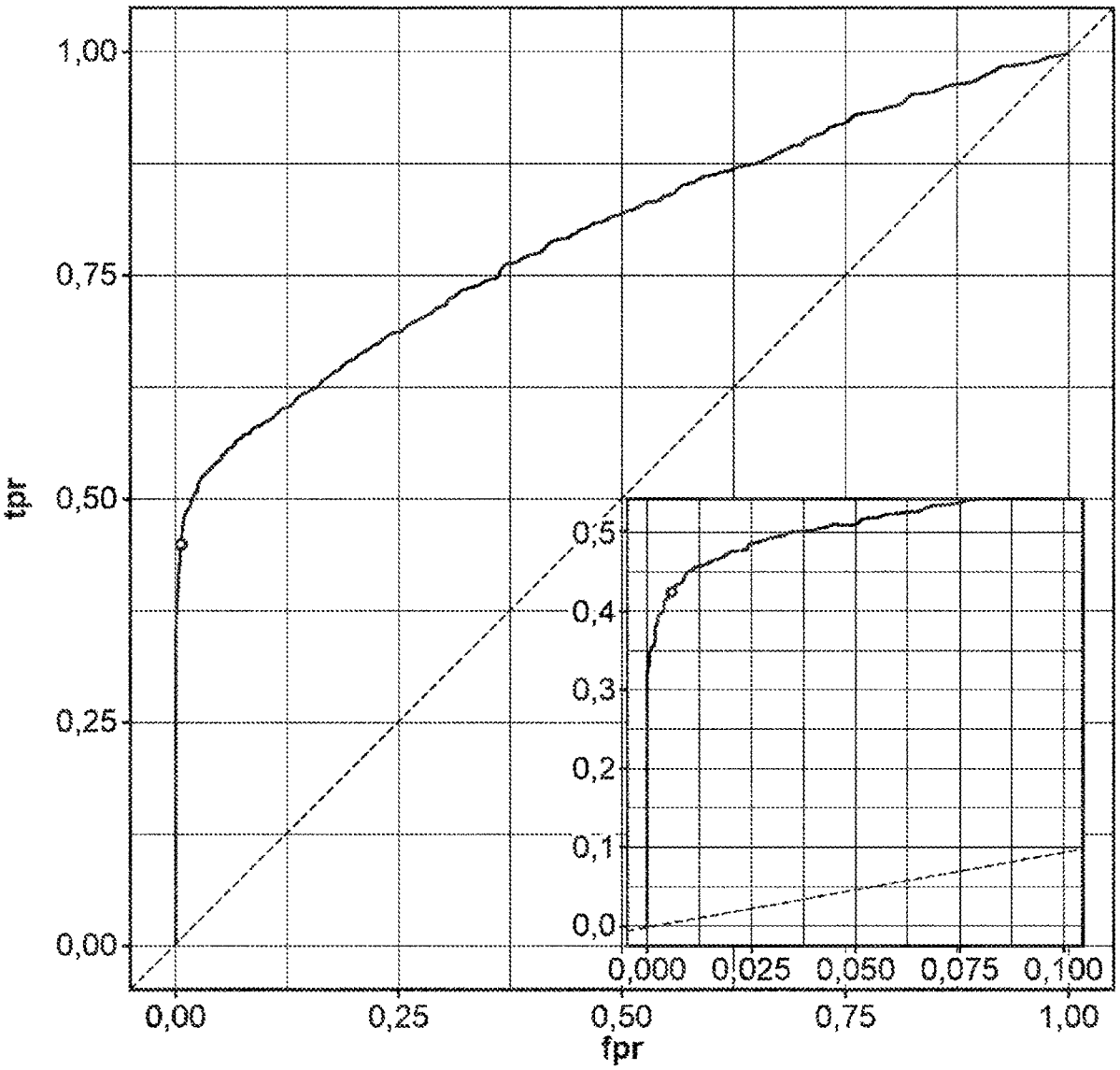
1,00
0,75
0,50
0,25
0,00
tpr
0,00
0,25
0,50
0,75
1,00
fpr
0,5
0,4
0,3
0,2
0,1
0,0
0,000
0,025
0,050
0,075
0,100

Figura 10