

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 990 062**

51 Int. Cl.:

G06F 18/2115 (2013.01)
G06F 18/213 (2013.01)
C12Q 1/6886 (2008.01)
C12Q 1/68 (2008.01)
C12Q 1/6883 (2008.01)
G16B 30/00 (2009.01)
G16B 40/00 (2009.01)
G16H 50/20 (2008.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

- 86 Fecha de presentación y número de la solicitud internacional: **22.05.2020** **PCT/US2020/034317**
87 Fecha y número de publicación internacional: **26.11.2020** **WO20237184**
96 Fecha de presentación y número de la solicitud europea: **22.05.2020** **E 20810838 (1)**
97 Fecha y número de publicación de la concesión europea: **24.07.2024** **EP 3973080**

54 Título: **Sistemas y métodos para determinar si un sujeto tiene una condición de cáncer utilizando aprendizaje por transferencia**

30 Prioridad:

22.05.2019 US 201962851486 P

45 Fecha de publicación y mención en BOPI de la traducción de la patente:
28.11.2024

73 Titular/es:

GRAIL, INC. (100.0%)
1525 O'Brien Drive
Menlo Park, California 94025, US

72 Inventor/es:

MAHER, M., CYRUS

74 Agente/Representante:

VALLEJO LÓPEZ, Juan Pedro

ES 2 990 062 T3

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

DESCRIPCIÓN

Sistemas y métodos para determinar si un sujeto tiene una condición de cáncer utilizando aprendizaje por transferencia

5 **Campo técnico**

Esta memoria descriptiva describe el uso del aprendizaje por transferencia para determinar si un sujeto tiene una condición de cáncer.

10 **Antecedentes**

El creciente conocimiento de la base molecular del cáncer y el rápido desarrollo de las técnicas de secuenciación de próxima generación están impulsando el estudio de las alteraciones moleculares tempranas implicadas en el desarrollo del cáncer en los fluidos corporales. Las tecnologías de secuenciación a gran escala, tales como la secuenciación de próxima generación (NGS, por sus siglas en inglés), han brindado la oportunidad de lograr la secuenciación a un coste inferior a un dólar estadounidense por millón de bases y, de hecho, se han logrado costes de menos de diez centavos de dólar estadounidense por millón de bases. Las alteraciones genéticas y epigenéticas específicas asociadas con dicho desarrollo de cáncer se encuentran en el ADN libre de células (ADNlc) en plasma, suero y orina. Tales alteraciones podrían potencialmente usarse como biomarcadores de diagnóstico para varias clases de cánceres. Ver, Salvi y col., 2016, *Onco Targets Ther.* 9:6549-6559.

El ADN libre de células (ADNlc) puede encontrarse en el suero, el plasma, la orina y otros fluidos corporales (Chan y col., 2003, *Ann Clin Biochem.* 40(Pt 2):122-130), lo que representa una “biopsia líquida”, que es una imagen circulante de una enfermedad específica. Ver, De Mattos-Arruda y Caldas, 2016, *Mol Oncol.* 10(3):464-474. Esto representa un método potencial no invasivo para detectar una variedad de tipos de cáncer.

La existencia del ADNlc fue demostrada hace décadas por Mandel y Metais (Mandel y Metais, 1948, *C R Seances Soc Biol Fil.* 142(3-4):241-243). El ADNlc se origina en células necróticas o apoptóticas, y generalmente lo liberan todos los tipos de células. Stroun y col. demostraron además que se podían encontrar alteraciones específicas del cáncer en el ADNlc de los pacientes. Ver, Stroun y col., 1989 *Oncology* 1989 46(5):318-322. Varios artículos posteriores confirmaron que el ADNlc contiene alteraciones específicas relacionadas con los tumores, tales como mutaciones, metilación y variaciones en el número de copias (CNV, por sus siglas en inglés), confirmando así la existencia de ADN tumoral circulante (ADNtc). Ver, Goessl y col., 2000 *Cancer Res.* 60(21):5941-5945 y Frenel y col., 2015, *Clin Cancer Res.* 21(20):4586-4596.

El ADNlc en plasma o suero se ha caracterizado bien, mientras que el ADNlc urinario (ADNlcu) tradicionalmente se ha caracterizado menos. Sin embargo, estudios recientes demostraron que el ADNlcu también podría ser una fuente prometedora de biomarcadores (p. ej., Casadio y col., 2013, *Urol Oncol.* 31(8):1744-1750).

En la sangre, la apoptosis es un evento frecuente que determina la cantidad de ADNlc. Sin embargo, en pacientes con cáncer, la cantidad de ADNlc también parece estar influenciada por la necrosis. Ver, Hao y col., 2014, *Br J Cancer* 111(8): 1482-1489 y Zonta y col., 2015 *Adv Clin Chern.* 70:197-246. Dado que la apoptosis parece ser el principal mecanismo de liberación, el ADNlc circulante tiene una distribución de tamaños que revela un enriquecimiento en fragmentos cortos de aproximadamente 167 pb correspondientes a los nucleosomas generados por las células apoptóticas. Ver, Heitzer y col., 2015, *Colin Chern.* 61(1):112-123 y Lo y col., 2010, *Sci Transl Med.* (261):61ra91.

La cantidad de ADNlc circulante en el suero y el plasma parece ser significativamente mayor en los pacientes con tumores que en los controles sanos, especialmente en aquellos con tumores en etapa avanzada que en aquellos con tumores en etapa temprana. Ver Sozzi y col., 2003, *J. Colin Oncol.* 21(21):3902-3908, Kim y col., 2014, *Ann Surg Treat Res.* 86(3):136-142; y Shao y col., 2015, *Oncol Lett.* 10(6):3478-3482. La variabilidad de la cantidad de ADNlc circulante es mayor en los pacientes con cáncer que en los individuos sanos (ver, Heitzer y col., 2013, *Int J Cancer.* 133(2):346-356), y la cantidad de ADNlc circulante está influenciada por varios desórdenes fisiológicos y patologías, incluidas las enfermedades proinflamatorias (ver, Raptis y Menard, 1980, *J Clin Invest.* 66(6):1391-1399, y Shapiro y col., 1983, *Cancer* 51(11):2116-2120).

Se sabe que el estado de metilación y otras modificaciones epigenéticas se correlacionan con la presencia de algunas enfermedades tales como el cáncer (ver Jones, 2002, *Oncogene* 21:5358-5360). Y se ha determinado que patrones de metilación específicos se asocian con determinadas condiciones de cáncer. Ver Paska y Hudler, 2015, *Biochemia Medica* 25(2):161-176.

Warton y Samimi han demostrado que se pueden observar patrones de metilación incluso en el ADN libre de células (Warton y Samimi, 2015, *Front Mol Biosci.* 2(13) doi:10.3389/fmolb.2015.00013).

El estado de la técnica pertinente adicional incluye los documentos WO2019/084559A1 y US2019/130065A1.

Dada la promesa del ADNlc circulante, así como de otras formas de datos de genotipo, como indicador de diagnóstico, en la técnica se necesitan mejores formas de evaluar dichos datos para identificar una condición de cáncer en los sujetos.

5 Resumen

Un aspecto de la presente descripción proporciona un sistema informático para entrenar a uno o más clasificadores en la discriminación de cada condición de cáncer de un conjunto de condiciones de cáncer. El conjunto de condiciones de cáncer comprende dos o más condiciones de cáncer. El sistema informático comprende al menos un procesador y una memoria que almacena al menos un programa para la ejecución por parte del al menos un procesador.

El al menos un programa comprende instrucciones para proporcionar una primera pluralidad de funciones de extracción de características basándose en un primer conjunto de datos. El primer conjunto de datos comprende, para cada sujeto respectivo de una primera pluralidad de sujetos de una especie, la primera información de genotipo correspondiente que comprende (i) una primera pluralidad correspondiente de valores de bin, cada valor de bin respectivo de la primera pluralidad correspondiente de valores de bin para un bin correspondiente de una pluralidad de bins y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer.

Cada bin de la pluralidad de bins representa una porción de un genoma de referencia de la especie.

La primera pluralidad correspondiente de valores de bin de cada sujeto respectivo de la primera pluralidad de sujetos se obtiene de una muestra biológica correspondiente del sujeto respectivo, que comprende un primer tipo de tejido, utilizando una primera pluralidad correspondiente de lecturas de secuencia determinadas mediante un primer método de secuenciación de ácido nucleico.

La primera pluralidad de funciones de extracción de características se obtiene aplicando una primera técnica de extracción de característica a los valores de bin respectivos de los sujetos respectivos del primer conjunto de datos, identificando de este modo la primera pluralidad de funciones de extracción de característica. Cada función de extracción de características de la primera pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal de los valores de bin de la totalidad o un subconjunto de la pluralidad de bins. La primera pluralidad de funciones de extracción de características discrimina colectivamente a los sujetos respectivos de la primera pluralidad de sujetos según tengan una condición de cáncer dentro del conjunto de condiciones de cáncer basándose en los valores de bin respectivos para los sujetos respectivos.

El al menos un programa comprende además instrucciones para obtener un segundo conjunto de datos que comprende, para cada sujeto respectivo de una segunda pluralidad de sujetos de la especie, la segunda información de genotipo correspondiente que comprende (i) una segunda pluralidad correspondiente de valores de bin, representando cada valor de bin respectivo de la segunda pluralidad correspondiente de valores de bin un bin correspondiente en la pluralidad de bins y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer. La segunda pluralidad correspondiente de valores de bin de cada sujeto respectivo de la segunda pluralidad de sujetos se obtiene de una muestra biológica correspondiente del sujeto respectivo, que comprende un segundo tipo de tejido, utilizando una segunda pluralidad correspondiente de lecturas de secuencia determinadas mediante un segundo método de secuenciación de ácido nucleico.

Al menos el primer método de secuenciación de ácido nucleico difiere del segundo método de secuenciación de ácido nucleico o el primer tipo de tejido difiere del segundo tipo de tejido.

El al menos un programa comprende además instrucciones para transformar, basándose en al menos cada función de extracción de características respectiva de la primera pluralidad de funciones de extracción de características, la segunda pluralidad respectiva de valores de bin de cada sujeto correspondiente de la segunda pluralidad de sujetos con respecto a la función de extracción de características respectiva, contribuyendo de este modo a un segundo conjunto de datos transformado con una pluralidad respectiva de valores de características para cada sujeto correspondiente.

El al menos un programa comprende además instrucciones para utilizar el segundo conjunto de datos transformado y la indicación de la condición de cáncer de los sujetos respectivos de la segunda pluralidad de sujetos para entrenar a un primer clasificador del uno o más clasificadores en al menos un par de condiciones de cáncer del conjunto de condiciones de cáncer.

En algunas realizaciones, cada valor de bin respectivo de la primera pluralidad de valores de bin o la segunda pluralidad de valores de bin representa un número de lecturas de secuencia medidas a partir de ácidos nucleicos libres de células en la muestra biológica correspondiente que se asocian con el bin respectivo.

En algunas realizaciones, el primer o el segundo método de secuenciación de ácido nucleico es la secuenciación dirigida o la secuenciación del genoma completo.

En algunas realizaciones, el primer método de secuenciación de ácido nucleico es la secuenciación dirigida utilizando una pluralidad de sondas de ácido nucleico, y el segundo método de secuenciación de ácido nucleico es la secuenciación del genoma completo. En algunas realizaciones alternativas, el primer método de secuenciación de ácido nucleico es la secuenciación del genoma completo, y el segundo método de secuenciación de ácido nucleico es la secuenciación dirigida utilizando una pluralidad de sondas de ácido nucleico.

En algunas realizaciones, cada valor de bin respectivo de la primera pluralidad de valores de bin o la segunda pluralidad de valores de bin representa un número de lecturas de secuencia asociadas con el bin respectivo de la pluralidad de bins o un patrón de metilación respectivo medido en la muestra biológica correspondiente para el bin respectivo de la pluralidad de bins.

En algunas realizaciones, la muestra biológica correspondiente del sujeto respectivo de la primera pluralidad o la segunda pluralidad de sujetos comprende sangre, sangre completa, plasma, suero, orina, líquido cefalorraquídeo, materia fecal, saliva, sudor, lágrimas, líquido pleural, líquido pericárdico o líquido peritoneal del sujeto respectivo.

En algunas realizaciones, el primer tipo de tejido es sangre y la muestra biológica correspondiente para cada sujeto respectivo de la primera pluralidad de sujetos es sangre, y el segundo tipo de tejido es uno de tejido mamario, tejido pulmonar, tejido prostático, tejido colorrectal, tejido renal, tejido uterino, tejido pancreático, tejido esofágico, tejido de cabeza/cuello, tejido ovárico, tejido hepatobiliar, tejido de cuello uterino, tejido tiroideo, tejido vesical y la muestra biológica correspondiente para cada sujeto respectivo de la segunda pluralidad de sujetos consiste del segundo tipo de tejido.

En algunas realizaciones, el primer tipo de tejido es sangre y la muestra biológica correspondiente para cada sujeto respectivo de la primera pluralidad de sujetos es sangre, y el segundo tipo de tejido es uno de tejido mamario, tejido pulmonar, tejido prostático, tejido colorrectal, tejido renal, tejido uterino, tejido pancreático, tejido esofágico, tejido de cabeza/cuello, tejido ovárico, tejido hepatobiliar, tejido de cuello uterino, tejido tiroideo, tejido vesical y la muestra biológica correspondiente para cada sujeto respectivo de la segunda pluralidad de sujetos comprende un tumor sólido del segundo tipo de tejido.

En algunas realizaciones, el primer tipo de tejido es uno de tejido mamario, tejido pulmonar, tejido prostático, tejido colorrectal, tejido renal, tejido uterino, tejido pancreático, tejido esofágico, tejido de cabeza/cuello, tejido ovárico, tejido hepatobiliar, tejido de cuello uterino, tejido tiroideo, tejido vesical, y la muestra biológica correspondiente para cada sujeto respectivo de la primera pluralidad de sujetos consiste en el primer tipo de tejido, y el segundo tipo de tejido es sangre y la muestra biológica correspondiente para cada sujeto respectivo de la segunda pluralidad de sujetos es sangre.

En algunas realizaciones, el primer tipo de tejido es uno de tejido mamario, tejido pulmonar, tejido prostático, tejido colorrectal, tejido renal, tejido uterino, tejido pancreático, tejido esofágico, tejido de cabeza/cuello, tejido ovárico, tejido hepatobiliar, tejido de cuello uterino, tejido tiroideo, tejido vesical y la muestra biológica correspondiente para cada sujeto respectivo de la primera pluralidad de sujetos comprende un tumor sólido del primer tipo de tejido, y el segundo tipo de tejido es sangre y la muestra biológica correspondiente para cada sujeto respectivo de la segunda pluralidad de sujetos es sangre.

En algunas realizaciones, cada sujeto respectivo de la primera pluralidad de sujetos comprende un tumor sólido del primer tipo de tejido, el segundo tipo de tejido es sangre y la muestra biológica correspondiente para cada sujeto respectivo de la segunda pluralidad de sujetos es sangre.

En algunas realizaciones, el primer tipo de tejido es sangre, la muestra biológica correspondiente para cada sujeto respectivo de la primera pluralidad de sujetos es sangre, y cada sujeto respectivo de la segunda pluralidad de sujetos comprende un tumor sólido del segundo tipo de tejido.

En algunas realizaciones, las instrucciones para la transformación hacen uso de cada función de extracción de características respectiva de la primera pluralidad de funciones de extracción de características y una segunda pluralidad de funciones de extracción de características. En tales realizaciones, la segunda pluralidad de funciones de extracción de características se obtiene aplicando una segunda técnica de extracción de características a los valores de bin respectivos de los sujetos respectivos del segundo conjunto de datos, identificando de este modo la segunda pluralidad de funciones de extracción de características. Cada función de extracción de características de la segunda pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal de los valores de bin de la totalidad o un subconjunto de la pluralidad de bins. Además, la segunda pluralidad de funciones de extracción de características discrimina colectivamente a los sujetos respectivos de la segunda pluralidad de sujetos según tengan una condición de cáncer dentro del conjunto de condiciones de cáncer basándose en los valores de bin respectivos para los sujetos respectivos.

En algunas realizaciones, el conjunto de condiciones de cáncer comprende tres o más condiciones de cáncer y, para cada condición de cáncer respectiva del conjunto de condiciones de cáncer, hay dos o más sujetos de la primera pluralidad de sujetos que tienen la condición de cáncer respectiva. Además, para cada par respectivo de condiciones

de cáncer del conjunto de condiciones de cáncer, la aplicación de la primera técnica de extracción de características se realiza aplicando una instancia de la primera técnica de extracción de características a los valores de bin respectivos de los sujetos respectivos del primer conjunto de datos que tienen una primera condición de cáncer o una segunda condición de cáncer del par respectivo de condiciones de cáncer. Además, cada instancia de la primera técnica de extracción de características contribuye con un subconjunto correspondiente de la primera pluralidad de funciones de extracción de características a la primera pluralidad de funciones de extracción de características.

En algunas realizaciones, el al menos un programa comprende además instrucciones para aplicar una segunda técnica de extracción de características a los valores de bin respectivos de los sujetos respectivos del segundo conjunto de datos, identificando de este modo una segunda pluralidad de funciones de extracción de características. En tales realizaciones, cada función de extracción de características de la segunda pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal de los valores de bin de la totalidad o un subconjunto de la pluralidad de bins. Además, la segunda pluralidad de funciones de extracción de características discrimina colectivamente a los sujetos respectivos de la segunda pluralidad de sujetos según tengan una condición de cáncer dentro del conjunto de condiciones de cáncer basándose en los valores de bin respectivos para los sujetos respectivos. En tales realizaciones, la aplicación de la segunda pluralidad de funciones de extracción de características a los valores de bin respectivos de los sujetos respectivos del segundo conjunto de datos se utiliza para obtener una pluralidad respectiva de segundos valores de características para cada sujeto correspondiente del segundo conjunto de datos y el segundo conjunto de datos transformado comprende además la pluralidad respectiva de segundos valores de característica.

En algunas realizaciones, para cada par respectivo de condiciones de cáncer del conjunto de condiciones de cáncer, la aplicación de la segunda técnica de extracción de características se realiza aplicando una instancia de la segunda técnica de extracción de características a los valores de bin respectivos de los sujetos respectivos del segundo conjunto de datos que tienen una primera condición de cáncer o una segunda condición de cáncer del par respectivo de condiciones de cáncer, y cada instancia de la segunda técnica de extracción de características contribuye con un subconjunto correspondiente de la segunda pluralidad de funciones de extracción de características a la segunda pluralidad de funciones de extracción de características.

En algunas realizaciones, el conjunto de condiciones de cáncer consta de entre dos y cinco condiciones de cáncer únicas, al menos 10 condiciones de cáncer únicas, al menos 20 condiciones de cáncer únicas o al menos 22 condiciones de cáncer únicas.

En algunas realizaciones, cada subconjunto correspondiente de la primera pluralidad de funciones de extracción de características consta de entre cuatro y cien funciones de extracción de características.

En algunas realizaciones, cada subconjunto correspondiente de la primera pluralidad de funciones de extracción de características consta de entre cuatro y cien funciones de extracción de características, y cada subconjunto correspondiente de la segunda pluralidad de funciones de extracción de características consta de entre cuatro y cien funciones de extracción de características.

En algunas realizaciones, la aplicación de la primera técnica de extracción de características comprende aplicar un algoritmo de reducción de dimensiones a los sujetos del primer conjunto de datos que tienen una primera condición de cáncer o una segunda condición de cáncer del par respectivo de condiciones de cáncer, identificando de este modo el subconjunto correspondiente de la primera pluralidad de funciones de extracción de características.

En algunas realizaciones, la aplicación de la primera técnica de extracción de características comprende aplicar un algoritmo de reducción de dimensiones a los sujetos del primer conjunto de datos que tienen una primera condición de cáncer o una segunda condición de cáncer del par respectivo de condiciones de cáncer, identificando de este modo el subconjunto correspondiente de la primera pluralidad de funciones de extracción de características, y la aplicación de la segunda técnica de extracción de características comprende aplicar el algoritmo de reducción de dimensiones a los sujetos del segundo conjunto de datos que tienen una primera condición de cáncer o una segunda condición de cáncer del par respectivo de condiciones de cáncer, identificando de este modo el subconjunto correspondiente de la segunda pluralidad de funciones de extracción de características.

En algunas realizaciones, el conjunto de condiciones de cáncer comprende tres o más condiciones de cáncer y, para cada condición de cáncer respectiva del conjunto de condiciones de cáncer, hay dos o más sujetos de la primera pluralidad de sujetos que tienen la condición de cáncer respectiva, y la aplicación de la primera técnica de extracción de características comprende además aplicar un algoritmo de reducción de dimensiones a los sujetos del primer conjunto de datos que tienen una condición de cáncer de las tres o más condiciones de cáncer, identificando de este modo la primera pluralidad de funciones de extracción de características.

En algunas realizaciones, el al menos un programa comprende además instrucciones para aplicar una segunda técnica de extracción de características a los valores de bin respectivos de los sujetos respectivos del segundo conjunto de datos, identificando de este modo una segunda pluralidad de funciones de extracción de características, y para aplicar los valores de bin respectivos de los sujetos respectivos del segundo conjunto de datos a la segunda pluralidad de

funciones de extracción de características para obtener una pluralidad respectiva de segundos valores de características para cada sujeto correspondiente del segundo conjunto de datos. En tales realizaciones, cada función de extracción de características de la segunda pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal de los valores de bin de la totalidad o un subconjunto de la pluralidad de bins. Además, la segunda pluralidad de funciones de extracción de características discrimina colectivamente a los sujetos respectivos de la segunda pluralidad de sujetos según tengan una condición de cáncer dentro del conjunto de condiciones de cáncer basándose en los valores de bin respectivos para los sujetos respectivos. Además, En tales realizaciones, el uso del segundo conjunto de datos transformado y la indicación de la condición de cáncer de los sujetos respectivos de la segunda pluralidad de sujetos para entrenar un primer clasificador del uno o más clasificadores comprende utilizar la pluralidad de segundos valores de características del segundo conjunto de datos transformado para entrenar el primer clasificador. Además, la aplicación de la segunda técnica de extracción de características comprende además aplicar el algoritmo de reducción de dimensiones a los sujetos del segundo conjunto de datos que tienen una condición de cáncer de las tres o más condiciones de cáncer, identificando de este modo la segunda pluralidad de funciones de extracción de características. En algunas de tales realizaciones, el algoritmo de reducción de dimensiones es un algoritmo de reducción de dimensiones lineales o un algoritmo de reducción de dimensiones no lineales. En algunas de tales realizaciones, el algoritmo de reducción de dimensiones se selecciona del bin que consiste en un algoritmo de análisis de componentes principales, un algoritmo de análisis factorial, un mapeo de Sammon, un análisis de componentes curvilíneos, un algoritmo de incrustación de vecinos estocásticos (SNE), un algoritmo de Isomap, un algoritmo de despliegue de varianza máxima, un algoritmo de incrustación localmente lineal, un algoritmo de t-SNE, un algoritmo de factorización matricial no negativa, un algoritmo de análisis de componentes principales de kernel, un algoritmo de análisis de componentes principales de kernel basado en gráficos, un algoritmo de análisis discriminante no lineal, un algoritmo de análisis discriminante generalizado, un algoritmo de aproximación y proyección de variedad uniforme (UMAP), un algoritmo de Large Vis, un algoritmo Laplacian Eigenmap y un algoritmo de análisis discriminante lineal de Fisher.

En otras realizaciones adicionales, el algoritmo de reducción de dimensiones es un algoritmo de análisis de componentes principales, y cada función de extracción de características respectiva del subconjunto correspondiente de la primera pluralidad de funciones de extracción de características comprende un componente principal respectivo derivado del algoritmo de reducción de dimensiones.

En algunas realizaciones, cada función de extracción de características respectiva del subconjunto correspondiente de la segunda pluralidad de funciones de extracción de características comprende un componente principal respectivo derivado del algoritmo de reducción de dimensiones.

En algunas realizaciones, el subconjunto correspondiente de la primera pluralidad de funciones de extracción de características se limita a un número umbral de componentes principales calculado por el algoritmo de análisis de componentes principales.

En algunas realizaciones, el subconjunto correspondiente de la primera pluralidad de funciones de extracción de características o el subconjunto correspondiente de la segunda pluralidad de funciones de extracción de características se limita a un número umbral de componentes principales calculado por el algoritmo de análisis de componentes principales.

En algunas realizaciones, a cada componente principal calculado por el algoritmo de análisis de componentes principales se le asigna un valor propio mediante el algoritmo de componentes principales, y el subconjunto correspondiente de la primera pluralidad de funciones de extracción de características se limita al número umbral de componentes principales a los que se asignan los valores de propios más altos.

En algunas realizaciones, a cada componente principal calculado por el algoritmo de análisis de componentes principales se le asigna un valor propio mediante el algoritmo de componentes principales, y el subconjunto correspondiente de la primera pluralidad de funciones de extracción de características o el subconjunto correspondiente de la segunda pluralidad de funciones de extracción de características se limita al número umbral de componentes principales a los que se asignan los valores de propios más altos.

En algunas realizaciones, el al menos un programa comprende además instrucciones para podar la primera pluralidad de funciones de extracción de características utilizando la primera pluralidad de funciones de extracción de características y las indicaciones respectivas de la condición de cáncer de los sujetos respectivos de la primera pluralidad de sujetos, eliminando de este modo una serie de funciones de extracción de características de la primera pluralidad de funciones de extracción de características.

En algunas realizaciones, el al menos un programa comprende instrucciones para podar la segunda pluralidad de funciones de extracción de características utilizando la segunda pluralidad de funciones de extracción de características y las indicaciones respectivas de la condición de cáncer de los sujetos respectivos de la segunda pluralidad de sujetos, eliminando de este modo una serie de funciones de extracción de características de la segunda pluralidad de funciones de extracción de características.

En algunas realizaciones, la poda de la primera pluralidad de funciones de extracción de características provoca una reducción de al menos cuatro veces en el número de funciones de extracción de características de la primera pluralidad de funciones de extracción de características.

- 5 En algunas realizaciones, la poda de la segunda pluralidad de funciones de extracción de características provoca una reducción de al menos cuatro veces en el número de funciones de extracción de características de la segunda pluralidad de funciones de extracción de características.

- 10 En algunas realizaciones, la poda de la primera pluralidad de funciones de extracción de características comprende hacer una regresión del primer conjunto de datos basándose en la totalidad o un subconjunto de la primera pluralidad de funciones de extracción de características utilizando un algoritmo de regresión para asignar de este modo un coeficiente de regresión a cada función de extracción de características respectiva de la primera pluralidad de funciones de extracción de características, conduciendo la regresión a la clasificación de una condición de cáncer del conjunto de condiciones de cáncer para cada sujeto de la primera pluralidad de sujetos, y a la eliminación de las funciones de extracción de características de la primera pluralidad de funciones de extracción de características a las que el algoritmo de regresión asigna un coeficiente negativo.

- 15 En algunas realizaciones, la poda de la segunda pluralidad de funciones de extracción de características comprende hacer una regresión del segundo conjunto de datos basándose en la totalidad o un subconjunto de la segunda pluralidad de funciones de extracción de características utilizando un algoritmo de regresión para asignar de este modo un coeficiente de regresión a cada función de extracción de características respectiva de la segunda pluralidad de funciones de extracción de características, conduciendo la regresión a la clasificación de una condición de cáncer del conjunto de condiciones de cáncer para cada sujeto de la segunda pluralidad de sujetos, y a la eliminación de las funciones de extracción de características de la segunda pluralidad de funciones de extracción de características a las que el algoritmo de regresión asigna un coeficiente negativo.

- 20 En algunas realizaciones, La poda comprende hacer una regresión del primer conjunto de datos basándose en la totalidad o un subconjunto de la primera pluralidad de funciones de extracción de características utilizando un algoritmo de regresión para asignar de este modo un coeficiente de regresión a cada función de extracción de características respectiva de la primera pluralidad de funciones de extracción de características, conduciendo la regresión a la clasificación de una condición de cáncer del conjunto de condiciones de cáncer para cada sujeto de la primera pluralidad de sujetos y a la eliminación de las funciones de extracción de características de la primera pluralidad de funciones de extracción de características a las que el algoritmo de regresión asigna un coeficiente que no cumple con un umbral de coeficiente.

- 25 En algunas realizaciones, la poda comprende hacer una regresión del segundo conjunto de datos basándose en la totalidad o un subconjunto de la segunda pluralidad de funciones de extracción de características utilizando un algoritmo de regresión para asignar de este modo un coeficiente de regresión a cada función de extracción de características respectiva de la segunda pluralidad de funciones de extracción de características, conduciendo la regresión a la clasificación de una condición de cáncer del conjunto de condiciones de cáncer para cada sujeto de la segunda pluralidad de sujetos, y a la eliminación de las funciones de extracción de características de la segunda pluralidad de funciones de extracción de características a las que el algoritmo de regresión asigna un coeficiente que no cumple con un umbral de coeficiente.

- 30 En algunas realizaciones, el algoritmo de regresión es una regresión logística.

- 35 En algunas realizaciones, cada función de extracción de características de la primera pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal del valor de bin de la totalidad o un subconjunto de la pluralidad de bins en forma de un peso independiente para cada bin respectivo de la pluralidad de bins o el subconjunto de la pluralidad de bins.

- 40 En algunas realizaciones, cada función de extracción de características de la segunda pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal del valor de bin de la totalidad o un subconjunto de la pluralidad de bins en forma de un peso independiente para cada bin respectivo de la pluralidad de bins o el subconjunto de la pluralidad de bins.

- 45 En algunas realizaciones, la primera técnica de extracción de características es una primera red neuronal convolucional que comprende una primera pluralidad de capas convolucionales, cada capa convolucional respectiva en la primera pluralidad de capas convolucionales se asocia con un vector de peso aprendido que se obtiene mediante la retropropagación en la primera red neuronal convolucional utilizando los valores de bin respectivos y las indicaciones respectivas de la condición de cáncer de los sujetos respectivos del primer conjunto de datos, y el vector de peso aprendido respectivo de cada capa opcional de un subconjunto de la primera pluralidad de capas convolucionales representa colectivamente la primera pluralidad de funciones de extracción de características, y la transformación comprende introducir la segunda pluralidad correspondiente de valores de bin de un sujeto respectivo de la segunda pluralidad de sujetos en una segunda red convolucional que comprende el subconjunto de la primera pluralidad de capas convolucionales, en donde un vector de peso de cada capa convolucional respectiva de la segunda red neuronal

convolucional se inicializa y, en algunas realizaciones, se congela en valores del vector de peso aprendido de la capa convolucional correspondiente en la primera red neuronal convolucional.

5 En algunas realizaciones, la primera pluralidad de capas convolucionales comprende tres, cuatro, cinco, seis, siete, ocho, nueve, diez, entre 5 y 50, o entre 3 y 100 capas convolucionales.

10 En algunas realizaciones, la primera pluralidad de capas convolucionales comprende cinco capas convolucionales y el subconjunto de la primera pluralidad de capas convolucionales consiste en las tres primeras capas convolucionales en la primera red neuronal convolucional.

15 En algunas realizaciones, el al menos un programa comprende además instrucciones para escalar un primer valor de bin respectivo para cada bin respectivo de la pluralidad de bins para cada sujeto respectivo de la primera pluralidad de sujetos realizando una transformación logarítmica del primer valor de bin respectivo, formando de este modo un primer valor de bin transformado logarítmicamente para el bin respectivo, restando un valor medio del primer valor de bin respectivo transformado logarítmicamente en la primera pluralidad de sujetos del primer valor de bin transformado logarítmicamente del bin respectivo, formando de este modo un primer valor de bin normalizado para el bin respectivo y, posteriormente, dividiendo el primer valor de bin normalizado respectivo para el bin respectivo por una desviación estándar del primer valor de bin normalizado en la primera pluralidad de sujetos, escalando de este modo el primer valor de bin para cada bin respectivo de la pluralidad de bins para cada sujeto respectivo de la primera pluralidad de sujetos.

20 En algunas realizaciones, el al menos un programa comprende además instrucciones para escalar un segundo valor de bin respectivo para cada bin respectivo de la pluralidad de bins para cada sujeto respectivo de la segunda pluralidad de sujetos realizando una transformación logarítmica del segundo valor de bin respectivo, formando de este modo un segundo valor de bin transformado logarítmicamente para el bin respectivo, restando un valor medio del segundo valor de bin respectivo transformado logarítmicamente en la segunda pluralidad de sujetos del segundo valor de bin transformado logarítmicamente del bin respectivo, formando de este modo un segundo valor de bin normalizado para el bin respectivo y, posteriormente, dividiendo el segundo valor de bin normalizado respectivo para el bin respectivo por una desviación estándar del segundo valor de bin normalizado en la segunda pluralidad de sujetos, escalando de este modo el segundo valor de bin para cada bin respectivo de la pluralidad de bins para cada sujeto respectivo de la segunda pluralidad de sujetos.

25 En algunas realizaciones, cada condición de cáncer respectiva del conjunto de condiciones de cáncer se selecciona del grupo que consiste en no cáncer, cáncer de mama, cáncer de pulmón, cáncer de próstata, cáncer colorrectal, cáncer renal, cáncer de útero, cáncer de páncreas, cáncer de esófago, linfoma, cáncer de cabeza/cuello, cáncer de ovario, cáncer hepatobiliar, melanoma, cáncer de cuello uterino, mieloma múltiple, leucemia, cáncer de tiroides, cáncer de vejiga y cáncer gástrico.

30 En algunas realizaciones, cada condición de cáncer del conjunto de condiciones de cáncer es no cáncer, una etapa predeterminada de un cáncer de mama, una etapa predeterminada de un cáncer de pulmón, una etapa predeterminada de un cáncer de próstata, una etapa predeterminada de un cáncer colorrectal, una etapa predeterminada de un cáncer renal, una etapa predeterminada de un cáncer de útero, una etapa predeterminada de un cáncer de páncreas, una etapa predeterminada de un cáncer de esófago, una etapa predeterminada de un linfoma, una etapa predeterminada de un cáncer de cabeza/cuello, una etapa predeterminada de un cáncer de ovario, una etapa predeterminada de un cáncer hepatobiliar, una etapa predeterminada de un melanoma, una etapa predeterminada de un cáncer de cuello uterino, una etapa predeterminada de un mieloma múltiple, una etapa predeterminada de leucemia, una etapa predeterminada de un cáncer de tiroides, una etapa predeterminada de un cáncer de vejiga o una etapa predeterminada de un cáncer gástrico.

35 En algunas realizaciones, una condición de cáncer del conjunto de condiciones de cáncer es una métrica de supervivencia (p. ej., probabilidad predeterminada de supervivencia durante un período de tiempo predeterminado).

40 En algunas realizaciones, el primer tipo de tejido y el segundo tipo de tejido se seleccionan cada uno del grupo que consiste en tejido de mama, hígado, vejiga, pulmón, recto, tiroides, próstata, útero, ovario, esófago y vascular.

45 En algunas realizaciones, el primer tipo de tejido o el segundo tipo de tejido es sangre, sangre completa, plasma, suero, orina, líquido cefalorraquídeo, materia fecal, saliva, sudor, lágrimas, líquido pleural, líquido pericárdico o líquido peritoneal.

50 En algunas realizaciones, el primer tipo de tejido y el segundo tipo de tejido son los mismos.

55 En algunas realizaciones, el primer método de secuenciación o el segundo método de secuenciación generan lecturas de secuencia de ADN libre de células.

60 En algunas realizaciones, la especie es humana.

En algunas realizaciones, la primera pluralidad correspondiente de valores de bin de un sujeto respectivo de la primera pluralidad de sujetos se determina utilizando más de 20.000 lecturas de secuencia que se toman colectivamente de la muestra biológica correspondiente del sujeto respectivo según el primer método de secuenciación.

5 En algunas realizaciones, la primera pluralidad correspondiente de valores de bin de un sujeto respectivo de la primera pluralidad de sujetos se determina utilizando más de 30.000 lecturas de secuencia, más de 40.000 lecturas de secuencia, más de 50.000 lecturas de secuencia o más de 100.000 lecturas de secuencia que se toman colectivamente de la muestra biológica correspondiente del sujeto respectivo según el primer método de secuenciación.

10 En algunas realizaciones, la tasa de cobertura promedio de las lecturas de secuencia utilizadas para formar la primera pluralidad de valores de bin de un sujeto respectivo de la primera pluralidad de sujetos en todo el genoma de referencia de la especie es de al menos 10 veces, al menos 20 veces o al menos 40 veces.

15 En algunas realizaciones, cada lectura de secuencia respectiva utilizada para formar la primera pluralidad de valores de bin de un sujeto respectivo de la primera pluralidad de sujetos incluye (i) una primera porción que se puede mapear en el genoma de la especie y (ii) una segunda porción, y las lecturas de secuencia utilizadas para formar la primera pluralidad de valores de bin de un sujeto respectivo de la primera pluralidad de sujetos se filtran de modo que solo se usen las lecturas de secuencia cuya primera porción es inferior a 160 nucleótidos para formar el los valores de bin.

20 En algunas realizaciones, el primer método de secuenciación de ácido nucleico o el segundo método de secuenciación de ácido nucleico es la secuenciación del genoma completo, la secuenciación por panel dirigida o la secuenciación con bisulfito del genoma completo.

En algunas realizaciones, el primer clasificador es un clasificador multivariante.

25 En algunas realizaciones, el primer clasificador es un único clasificador multivariante que discrimina cada condición de cáncer del conjunto de condiciones de cáncer.

30 En algunas realizaciones, el primer clasificador es una regresión logística multivariante, una red neuronal o una red neuronal convolucional.

En algunas realizaciones, el primer clasificador es una máquina de vectores de soporte, una red neuronal, un árbol de decisiones, un agrupamiento o un algoritmo de regresión.

35 En algunas realizaciones, el uso del segundo conjunto de datos transformado y la indicación de la condición de cáncer de los sujetos respectivos de la segunda pluralidad de sujetos para entrenar a un primer clasificador se repite para cada par único de condiciones de cáncer del conjunto de condiciones de cáncer, entrenando de este modo a una pluralidad de clasificadores, en donde cada clasificador respectivo de la pluralidad de clasificadores se entrena en sujetos del segundo conjunto de datos transformado que tienen una de las primeras condiciones de cáncer o el segundo cáncer en el respectivo par de condiciones de cáncer representado por el clasificador respectivo.

40 En algunas realizaciones, el uno o más clasificadores consisten en el primer clasificador y el primer clasificador se entrena en una condición de cáncer distinta de una condición de cáncer del conjunto de condiciones de cáncer.

45 En algunas realizaciones, el al menos un programa comprende además instrucciones para utilizar el primer clasificador para clasificar a un sujeto de prueba con una primera condición de cáncer utilizando información de genotipo, en formato electrónico, basándose en datos de lectura de secuencia en una muestra biológica de prueba obtenida del sujeto de prueba como entrada al primer clasificador.

50 En algunas realizaciones, la muestra biológica de prueba comprende sangre, sangre completa, plasma, suero, orina, líquido cefalorraquídeo, materia fecal, saliva, sudor, lágrimas, líquido pleural, líquido pericárdico o líquido peritoneal del sujeto de prueba.

En algunas realizaciones, la primera condición de cáncer está en el conjunto de condiciones de cáncer.

55 En algunas realizaciones, el al menos un programa comprende además instrucciones para utilizar el uno o más clasificadores para determinar la probabilidad de que un sujeto de prueba tenga cada condición de cáncer del conjunto de condiciones de cáncer utilizando información de genotipo, en formato electrónico, basándose en los datos de lectura de secuencia en una muestra biológica de prueba obtenida del sujeto de prueba como entrada para cada clasificador del uno o más clasificadores. En algunas de tales realizaciones, la muestra biológica de prueba comprende sangre, sangre completa, plasma, suero, orina, líquido cefalorraquídeo, materia fecal, saliva, sudor, lágrimas, líquido pleural, líquido pericárdico o líquido peritoneal del sujeto de prueba.

60 En algunas realizaciones, el al menos un programa comprende además instrucciones para utilizar el primer clasificador para determinar la probabilidad de que un sujeto de prueba tenga cada condición de cáncer del conjunto de

condiciones de cáncer utilizando información de genotipo, en formato electrónico, basándose en los datos de lectura de secuencia en una muestra biológica de prueba obtenida del sujeto de prueba como entrada al primer clasificador.

5 En algunas realizaciones, la muestra biológica de prueba comprende sangre, sangre completa, plasma, suero, orina, líquido cefalorraquídeo, materia fecal, saliva, sudor, lágrimas, líquido pleural, líquido pericárdico o líquido peritoneal del sujeto de prueba.

10 En algunas realizaciones, la primera pluralidad correspondiente de lecturas de secuencia comprende más de 10.000 lecturas de secuencia, la primera pluralidad de sujetos comprende veinticinco o más sujetos, y la pluralidad de bins comprende veinte o más bins.

En algunas realizaciones, la primera pluralidad de funciones de extracción de características comprende cuatro o más funciones de extracción de características.

15 En algunas realizaciones, la segunda pluralidad correspondiente de lecturas de secuencia comprende más de 10.000 lecturas de secuencia.

20 En algunas realizaciones, el primer método de secuenciación de ácido nucleico es la secuenciación por metilación y en donde cada valor de bin de la primera pluralidad correspondiente de valores de bin es un número de fragmentos representados por la correspondiente primera pluralidad de lecturas de secuencia.

25 En algunas realizaciones, la primera pluralidad correspondiente de valores de bin es un número de fragmentos representados por la primera pluralidad correspondiente de lecturas de secuencia después de la aplicación de una o más condiciones de filtro.

30 En algunas realizaciones, la secuenciación por metilación produce un patrón de metilación correspondiente para cada fragmento respectivo de la serie de fragmentos, y una condición de filtro de la una o más condiciones de filtro es la aplicación de un umbral de valor de p al patrón de metilación correspondiente, en donde el umbral de valor de p representa la frecuencia con la que se observa un patrón de metilación en una cohorte de sujetos sin cáncer. En algunas de tales realizaciones, el umbral de valor de p está entre 0,001 y 0,20.

35 En algunas realizaciones, la secuenciación por metilación produce un patrón de metilación correspondiente para cada fragmento respectivo de la serie de fragmentos, y una condición de filtro de la una o más condiciones de filtro es la aplicación del requisito de que el fragmento respectivo esté representado por un número umbral de lecturas de secuencia de la primera pluralidad correspondiente de lecturas de secuencia. En algunas de tales realizaciones, el número umbral es 2, 3, 4, 5, 6, 7, 8, 9, 10 o un entero entre 10 y 100.

40 En algunas realizaciones, la secuenciación por metilación produce un patrón de metilación correspondiente para cada fragmento respectivo de la serie de fragmentos, y una condición de filtro de la una o más condiciones de filtro es la aplicación del requisito de que el fragmento respectivo tenga un número umbral de sitios CpG (p. ej., al menos 1, 2, 3, 4, 5, 6, 7, 8, 9 o 10 sitios CpG).

45 En algunas realizaciones, la secuenciación por metilación produce un patrón de metilación correspondiente para cada fragmento respectivo de la serie de fragmentos, y una condición de filtro de la una o más condiciones de filtro es un requisito de que el fragmento respectivo tenga una longitud inferior a un número umbral de pares de bases (p. ej., 1 mil, 2 mil, 3 mil o 4 mil pares de bases contiguos de longitud).

50 Otro aspecto de la presente descripción proporciona un método para entrenar a uno o más clasificadores para discriminar entre cada condición de cáncer de un conjunto de condiciones de cáncer, comprendiendo el conjunto de condiciones de cáncer dos o más condiciones de cáncer. El método comprende proporcionar una primera pluralidad de funciones de extracción de características basadas en un primer conjunto de datos. El primer conjunto de datos comprende, para cada sujeto respectivo de una primera pluralidad de sujetos de una especie, la primera información de genotipo correspondiente que comprende (i) una primera pluralidad correspondiente de valores de bin, cada valor de bin respectivo de la primera pluralidad correspondiente de valores de bin para un bin correspondiente de una pluralidad de bins y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer. Cada bin de la pluralidad de bins representa una porción de un genoma de referencia de la especie. La primera pluralidad correspondiente de valores de bin de cada sujeto respectivo de la primera pluralidad de sujetos se obtiene de una muestra biológica correspondiente del sujeto respectivo, que comprende un primer tipo de tejido, utilizando una primera pluralidad correspondiente de lecturas de secuencia determinadas mediante un primer método de secuenciación de ácido nucleico. La primera pluralidad de funciones de extracción de características se obtiene aplicando una primera técnica de extracción de característica a los valores de bin respectivos de los sujetos respectivos del primer conjunto de datos, identificando de este modo la primera pluralidad de funciones de extracción de característica. Cada función de extracción de características de la primera pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal de los valores de bin de la totalidad o un subconjunto de la pluralidad de bins. La primera pluralidad de funciones de extracción de características discrimina

colectivamente a los sujetos respectivos de la primera pluralidad de sujetos según tengan una condición de cáncer dentro del conjunto de condiciones de cáncer basándose en los valores de bin respectivos para los sujetos respectivos.

El método comprende además obtener un segundo conjunto de datos que comprende, para cada sujeto respectivo de una segunda pluralidad de sujetos de la especie, la segunda información de genotipo correspondiente que comprende (i) una segunda pluralidad correspondiente de valores de bin, representando cada valor de bin respectivo de la segunda pluralidad correspondiente de valores de bin un bin correspondiente en la pluralidad de bins y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer. La segunda pluralidad correspondiente de valores de bin de cada sujeto respectivo de la segunda pluralidad de sujetos se obtiene de una muestra biológica correspondiente del sujeto respectivo, que comprende un segundo tipo de tejido, utilizando una segunda pluralidad correspondiente de lecturas de secuencia determinadas mediante un segundo método de secuenciación de ácido nucleico. Al menos el primer método de secuenciación de ácido nucleico difiere del segundo método de secuenciación de ácido nucleico o el primer tipo de tejido difiere del segundo tipo de tejido.

El método comprende además transformar, basándose en al menos cada función de extracción de características respectiva de la primera pluralidad de funciones de extracción de características, la segunda pluralidad respectiva de valores de bin de cada sujeto correspondiente de la segunda pluralidad de sujetos con respecto a la función de extracción de características respectiva, contribuyendo de este modo a un segundo conjunto de datos transformado con una pluralidad respectiva de valores de características para cada sujeto correspondiente.

El método comprende además utilizar el segundo conjunto de datos transformado y la indicación de la condición de cáncer de los sujetos respectivos de la segunda pluralidad de sujetos para entrenar a un primer clasificador del uno o más clasificadores en al menos un par de condiciones de cáncer del conjunto de condiciones de cáncer.

Otro aspecto de la presente descripción proporciona un medio de almacenamiento legible por ordenador no transitorio que tiene almacenadas en el mismo instrucciones de código de programa que, cuando son ejecutadas por un procesador, hacen que el procesador lleve a cabo un método para entrenar a uno o más clasificadores para discriminar entre cada condición de cáncer de un conjunto de condiciones de cáncer, comprendiendo el conjunto de condiciones de cáncer dos o más condiciones de cáncer. El método comprende proporcionar una primera pluralidad de funciones de extracción de características basadas en un primer conjunto de datos. El primer conjunto de datos comprende, para cada sujeto respectivo de una primera pluralidad de sujetos de una especie, la primera información de genotipo correspondiente que comprende (i) una primera pluralidad correspondiente de valores de bin, cada valor de bin respectivo de la primera pluralidad correspondiente de valores de bin para un bin correspondiente de una pluralidad de bins y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer. Cada bin de la pluralidad de bins representa una porción de un genoma de referencia de la especie. La primera pluralidad correspondiente de valores de bin de cada sujeto respectivo de la primera pluralidad de sujetos se obtiene de una muestra biológica correspondiente del sujeto respectivo, que comprende un primer tipo de tejido, utilizando una primera pluralidad correspondiente de lecturas de secuencia determinadas mediante un primer método de secuenciación de ácido nucleico. La primera pluralidad de funciones de extracción de características se obtiene aplicando una primera técnica de extracción de característica a los valores de bin respectivos de los sujetos respectivos del primer conjunto de datos, identificando de este modo la primera pluralidad de funciones de extracción de característica. Cada función de extracción de características de la primera pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal de los valores de bin de la totalidad o un subconjunto de la pluralidad de bins. La primera pluralidad de funciones de extracción de características discrimina colectivamente a los sujetos respectivos de la primera pluralidad de sujetos según tengan una condición de cáncer dentro del conjunto de condiciones de cáncer basándose en los valores de bin respectivos para los sujetos respectivos.

El método comprende además obtener un segundo conjunto de datos que comprende, para cada sujeto respectivo de una segunda pluralidad de sujetos de la especie, la segunda información de genotipo correspondiente que comprende (i) una segunda pluralidad correspondiente de valores de bin, representando cada valor de bin respectivo de la segunda pluralidad correspondiente de valores de bin un bin correspondiente en la pluralidad de bins y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer. La segunda pluralidad correspondiente de valores de bin de cada sujeto respectivo de la segunda pluralidad de sujetos se obtiene de una muestra biológica correspondiente del sujeto respectivo, que comprende un segundo tipo de tejido, utilizando una segunda pluralidad correspondiente de lecturas de secuencia determinadas mediante un segundo método de secuenciación de ácido nucleico. Al menos el primer método de secuenciación de ácido nucleico difiere del segundo método de secuenciación de ácido nucleico o el primer tipo de tejido difiere del segundo tipo de tejido.

El método comprende además transformar, basándose en al menos cada función de extracción de características respectiva de la primera pluralidad de funciones de extracción de características, la segunda pluralidad respectiva de valores de bin de cada sujeto correspondiente de la segunda pluralidad de sujetos con respecto a la función de extracción de características respectiva, contribuyendo de este modo a un segundo conjunto de datos transformado con una pluralidad respectiva de valores de características para cada sujeto correspondiente.

El método comprende además utilizar el segundo conjunto de datos transformado y la indicación de la condición de cáncer de los sujetos respectivos de la segunda pluralidad de sujetos para entrenar a un primer clasificador del uno o más clasificadores en al menos un par de condiciones de cáncer del conjunto de condiciones de cáncer.

Otro aspecto de la presente descripción es un sistema informático para entrenar a uno o más clasificadores para discriminar entre cada condición de cáncer de un conjunto de condiciones de cáncer, comprendiendo el conjunto de condiciones de cáncer dos o más condiciones de cáncer. El sistema informático comprende al menos un procesador y una memoria que almacena al menos un programa para la ejecución por parte del al menos un procesador. El al menos un programa comprende instrucciones para proporcionar una primera pluralidad de funciones de extracción de características basándose en un primer conjunto de datos, donde el primer conjunto de datos comprende, para cada sujeto respectivo de una primera pluralidad de sujetos de una especie, la primera información de genotipo correspondiente que comprende (i) una primera pluralidad correspondiente de valores de bin, cada valor de bin respectivo de la primera pluralidad correspondiente de valores de bin para un bin correspondiente de una pluralidad de bins y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer. Cada bin de la pluralidad de bins representa una porción de un genoma de referencia de la especie. La primera pluralidad correspondiente de valores de bin de cada sujeto respectivo de la primera pluralidad de sujetos se obtiene de una muestra biológica correspondiente del sujeto respectivo, que comprende un primer tipo de tejido, utilizando una primera pluralidad correspondiente de lecturas de secuencia determinadas mediante un primer método de secuenciación de ácido nucleico. La primera pluralidad de funciones de extracción de características se obtiene aplicando una primera técnica de extracción de característica a los valores de bin respectivos de los sujetos respectivos del primer conjunto de datos, identificando de este modo la primera pluralidad de funciones de extracción de característica. Cada función de extracción de características de la primera pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal de los valores de bin de la totalidad o un subconjunto de la pluralidad de bins. La primera pluralidad de funciones de extracción de características discrimina colectivamente a los sujetos respectivos de la primera pluralidad de sujetos según tengan una condición de cáncer dentro del conjunto de condiciones de cáncer basándose en los valores de bin respectivos para los sujetos respectivos.

El al menos un programa comprende instrucciones para obtener un segundo conjunto de datos que comprende, para cada sujeto respectivo de una segunda pluralidad de sujetos de la especie, la segunda información de genotipo correspondiente que comprende (i) una segunda pluralidad correspondiente de valores de bin, representando cada valor de bin respectivo de la segunda pluralidad correspondiente de valores de bin un bin correspondiente en la pluralidad de bins y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer. La segunda pluralidad correspondiente de valores de bin de cada sujeto respectivo de la segunda pluralidad de sujetos se obtiene de una muestra biológica correspondiente del sujeto respectivo, que comprende un segundo tipo de tejido, utilizando una segunda pluralidad correspondiente de lecturas de secuencia determinadas mediante un segundo método de secuenciación de ácido nucleico. Al menos el primer método de secuenciación de ácido nucleico o el segundo método de secuenciación de ácido nucleico es la secuenciación por metilación.

El al menos un programa comprende instrucciones para transformar, basándose en al menos cada función de extracción de características respectiva de la primera pluralidad de funciones de extracción de características, la segunda pluralidad respectiva de valores de bin de cada sujeto correspondiente de la segunda pluralidad de sujetos con respecto a la función de extracción de características respectiva, contribuyendo de este modo a un segundo conjunto de datos transformado con una pluralidad respectiva de valores de características para cada sujeto correspondiente.

El al menos un programa comprende instrucciones para utilizar el segundo conjunto de datos transformado y la indicación de la condición de cáncer de los sujetos respectivos de la segunda pluralidad de sujetos para entrenar a un primer clasificador del uno o más clasificadores en al menos un par de condiciones de cáncer del conjunto de condiciones de cáncer.

En algunas realizaciones, el primer o el segundo método de secuenciación de ácido nucleico es la secuenciación dirigida o la secuenciación del genoma completo.

En algunas realizaciones, el primer método de secuenciación de ácido nucleico es la secuenciación dirigida utilizando una pluralidad de sondas de ácido nucleico, y el segundo método de secuenciación de ácido nucleico es la secuenciación del genoma completo.

En algunas realizaciones, el primer método de secuenciación de ácido nucleico es la secuenciación del genoma completo, y el segundo método de secuenciación de ácido nucleico es la secuenciación dirigida utilizando una pluralidad de sondas de ácido nucleico.

En algunas realizaciones, el primer método de secuenciación de ácido nucleico es la secuenciación por metilación y en donde cada valor de bin de la primera pluralidad correspondiente de valores de bin es un número de fragmentos representados por la correspondiente primera pluralidad de lecturas de secuencia.

En algunas realizaciones, la primera pluralidad correspondiente de valores de bin es un número de fragmentos representados por la primera pluralidad correspondiente de lecturas de secuencia después de la aplicación de una o más condiciones de filtro.

En algunas realizaciones, la secuenciación por metilación produce un patrón de metilación correspondiente para cada fragmento respectivo de la serie de fragmentos, y una condición de filtro de la una o más condiciones de filtro es la aplicación de un umbral de valor de p (p. ej., entre 0,001 y 0,20) al patrón de metilación correspondiente. En tales realizaciones, el umbral de valor de p representa la frecuencia con la que se observa un patrón de metilación en una cohorte de sujetos sin cáncer.

En algunas realizaciones, la secuenciación por metilación produce un patrón de metilación correspondiente para cada fragmento respectivo de la serie de fragmentos, y una condición de filtro de la una o más condiciones de filtro es la aplicación del requisito de que el fragmento respectivo esté representado por un número umbral (p. ej., 2, 3, 4, 5, 6, 7, 8, 9, 10 o un entero entre 10 y 100) de lecturas de secuencia de la primera pluralidad correspondiente de lecturas de secuencia.

En algunas realizaciones, la secuenciación por metilación produce un patrón de metilación correspondiente para cada fragmento respectivo de la serie de fragmentos, y una condición de filtro de la una o más condiciones de filtro es la aplicación del requisito de que el fragmento respectivo tenga un número umbral de sitios CpG (p. ej., al menos 1, 2, 3, 4, 5, 6, 7, 8, 9 o 10 sitios CpG).

En algunas realizaciones, la secuenciación por metilación produce un patrón de metilación correspondiente para cada fragmento respectivo de la serie de fragmentos, y una condición de filtro de la una o más condiciones de filtro es un requisito de que el fragmento respectivo tenga una longitud inferior a un número umbral de pares de bases (p. ej., 1 mil, 2 mil, 3 mil o 4 mil pares de bases contiguos de longitud).

Otro aspecto de la presente descripción proporciona un método para entrenar a uno o más clasificadores para discriminar entre cada condición de cáncer de un conjunto de condiciones de cáncer, comprendiendo el conjunto de condiciones de cáncer dos o más condiciones de cáncer. El método comprende proporcionar una primera pluralidad de funciones de extracción de características basadas en un primer conjunto de datos. El primer conjunto de datos comprende, para cada sujeto respectivo de una primera pluralidad de sujetos de una especie, la primera información de genotipo correspondiente que comprende (i) una primera pluralidad correspondiente de valores de bin, cada valor de bin respectivo de la primera pluralidad correspondiente de valores de bin para un bin correspondiente de una pluralidad de bins y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer. Cada bin de la pluralidad de bins representa una porción de un genoma de referencia de la especie. La primera pluralidad correspondiente de valores de bin de cada sujeto respectivo de la primera pluralidad de sujetos se obtiene de una muestra biológica correspondiente del sujeto respectivo, que comprende un primer tipo de tejido, utilizando una primera pluralidad correspondiente de lecturas de secuencia determinadas mediante un primer método de secuenciación de ácido nucleico. La primera pluralidad de funciones de extracción de características se obtiene aplicando una primera técnica de extracción de característica a los valores de bin respectivos de los sujetos respectivos del primer conjunto de datos, identificando de este modo la primera pluralidad de funciones de extracción de característica. Cada función de extracción de características de la primera pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal de los valores de bin de la totalidad o un subconjunto de la pluralidad de bins. La primera pluralidad de funciones de extracción de características discrimina colectivamente a los sujetos respectivos de la primera pluralidad de sujetos según tengan una condición de cáncer dentro del conjunto de condiciones de cáncer basándose en los valores de bin respectivos para los sujetos respectivos.

El método comprende además obtener un segundo conjunto de datos que comprende, para cada sujeto respectivo de una segunda pluralidad de sujetos de la especie, la segunda información de genotipo correspondiente que comprende (i) una segunda pluralidad correspondiente de valores de bin, representando cada valor de bin respectivo de la segunda pluralidad correspondiente de valores de bin un bin correspondiente en la pluralidad de bins y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer. La segunda pluralidad correspondiente de valores de bin de cada sujeto respectivo de la segunda pluralidad de sujetos se obtiene de una muestra biológica correspondiente del sujeto respectivo, que comprende un segundo tipo de tejido, utilizando una segunda pluralidad correspondiente de lecturas de secuencia determinadas mediante un segundo método de secuenciación de ácido nucleico. Al menos el primer método de secuenciación de ácido nucleico o el segundo método de secuenciación de ácido nucleico es la secuenciación por metilación.

El método comprende además transformar, basándose en al menos cada función de extracción de características respectiva de la primera pluralidad de funciones de extracción de características, la segunda pluralidad respectiva de valores de bin de cada sujeto correspondiente de la segunda pluralidad de sujetos con respecto a la función de extracción de características respectiva, contribuyendo de este modo a un segundo conjunto de datos transformado con una pluralidad respectiva de valores de características para cada sujeto correspondiente.

El método comprende además utilizar el segundo conjunto de datos transformado y la indicación de la condición de cáncer de los sujetos respectivos de la segunda pluralidad de sujetos para entrenar a un primer clasificador del uno o más clasificadores en al menos un par de condiciones de cáncer del conjunto de condiciones de cáncer.

5 Otro aspecto de la presente descripción proporciona un medio de almacenamiento legible por ordenador no transitorio que tiene almacenadas en el mismo instrucciones de código de programa que, cuando son ejecutadas por un procesador, hacen que el procesador lleve a cabo un método para entrenar a uno o más clasificadores para discriminar entre cada condición de cáncer de un conjunto de condiciones de cáncer, comprendiendo el conjunto de condiciones de cáncer dos o más condiciones de cáncer. El método comprende proporcionar una primera pluralidad de funciones de extracción de características basadas en un primer conjunto de datos. El primer conjunto de datos comprende, para cada sujeto respectivo de una primera pluralidad de sujetos de una especie, la primera información de genotipo correspondiente que comprende (i) una primera pluralidad correspondiente de valores de bin, cada valor de bin respectivo de la primera pluralidad correspondiente de valores de bin para un bin correspondiente de una pluralidad de bins y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer. Cada bin de la pluralidad de bins representa una porción de un genoma de referencia de la especie. La primera pluralidad correspondiente de valores de bin de cada sujeto respectivo de la primera pluralidad de sujetos se obtiene de una muestra biológica correspondiente del sujeto respectivo, que comprende un primer tipo de tejido, utilizando una primera pluralidad correspondiente de lecturas de secuencia determinadas mediante un primer método de secuenciación de ácido nucleico. La primera pluralidad de funciones de extracción de características se obtiene aplicando una primera técnica de extracción de característica a los valores de bin respectivos de los sujetos respectivos del primer conjunto de datos, identificando de este modo la primera pluralidad de funciones de extracción de característica. Cada función de extracción de características de la primera pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal de los valores de bin de la totalidad o un subconjunto de la pluralidad de bins. La primera pluralidad de funciones de extracción de características discrimina colectivamente a los sujetos respectivos de la primera pluralidad de sujetos según tengan una condición de cáncer dentro del conjunto de condiciones de cáncer basándose en los valores de bin respectivos para los sujetos respectivos.

El método comprende además (B) obtener un segundo conjunto de datos que comprende, para cada sujeto respectivo de una segunda pluralidad de sujetos de la especie, la segunda información de genotipo correspondiente que comprende (i) una segunda pluralidad correspondiente de valores de bin, representando cada valor de bin respectivo de la segunda pluralidad correspondiente de valores de bin un bin correspondiente en la pluralidad de bins y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer. La segunda pluralidad correspondiente de valores de bin de cada sujeto respectivo de la segunda pluralidad de sujetos se obtiene de una muestra biológica correspondiente del sujeto respectivo, que comprende un segundo tipo de tejido, utilizando una segunda pluralidad correspondiente de lecturas de secuencia determinadas mediante un segundo método de secuenciación de ácido nucleico. Al menos el primer método de secuenciación de ácido nucleico o el segundo método de secuenciación de ácido nucleico es la secuenciación por metilación.

El método comprende además transformar, basándose en al menos cada función de extracción de características respectiva de la primera pluralidad de funciones de extracción de características, la segunda pluralidad respectiva de valores de bin de cada sujeto correspondiente de la segunda pluralidad de sujetos con respecto a la función de extracción de características respectiva, contribuyendo de este modo a un segundo conjunto de datos transformado con una pluralidad respectiva de valores de características para cada sujeto correspondiente.

El método comprende además utilizar el segundo conjunto de datos transformado y la indicación de la condición de cáncer de los sujetos respectivos de la segunda pluralidad de sujetos para entrenar a un primer clasificador del uno o más clasificadores en al menos un par de condiciones de cáncer del conjunto de condiciones de cáncer.

Otro aspecto de la presente descripción proporciona un sistema informático para clasificar a un sujeto de prueba según una primera condición de cáncer de un conjunto de condiciones de cáncer, comprendiendo el conjunto de condiciones de cáncer dos o más condiciones de cáncer. El sistema informático comprende al menos un procesador y una memoria que almacena al menos un programa para la ejecución por parte del al menos un procesador. El al menos un programa comprende instrucciones para obtener información de genotipo de prueba que comprende una pluralidad de valores de bin de prueba correspondiente, cada valor de bin respectivo de la pluralidad de valores de bin de prueba para un bin correspondiente de una pluralidad de bins. Cada bin de la pluralidad de bins representa una porción de un genoma de referencia de la especie. La pluralidad de valores de bin de prueba se obtiene de una muestra biológica de prueba del sujeto de prueba, utilizando una pluralidad de lecturas de secuencia de prueba correspondiente determinadas mediante un primer método de secuenciación de ácido nucleico. La pluralidad de lecturas de secuencia de prueba comprende al menos 10.000 lecturas de secuencia. La pluralidad de bins comprende al menos 100 bins. El al menos un programa comprende instrucciones para aplicar la pluralidad de valores de bin de prueba a un clasificador, entrenado en un segundo conjunto de datos transformado obtenido mediante el aprendizaje por transferencia entre un primer conjunto de datos y un segundo conjunto de datos, para hacer que el clasificador clasifique al sujeto de prueba según la primera condición de cáncer del conjunto de condiciones de cáncer. El primer conjunto de datos comprende, para cada sujeto respectivo de una primera pluralidad de sujetos de entrenamiento, la primera pluralidad de sujetos de entrenamiento que comprende al menos cincuenta sujetos, la primera información de genotipo correspondiente que comprende (i) una primera pluralidad correspondiente de valores de bin, cada valor de bin

respectivo de la primera pluralidad correspondiente de valores de bin para un bin correspondiente en la pluralidad de bins y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer. La primera pluralidad correspondiente de valores de bin de cada sujeto respectivo de la primera pluralidad de sujetos se obtiene de una muestra biológica correspondiente del sujeto respectivo, que comprende un primer tipo de tejido, utilizando una primera pluralidad correspondiente de lecturas de secuencia determinadas mediante un segundo método de secuenciación de ácido nucleico. El segundo conjunto de datos comprende, para cada sujeto respectivo de una segunda pluralidad de sujetos de la especie, la segunda información de genotipo correspondiente que comprende (i) una segunda pluralidad correspondiente de valores de bin, representando cada valor de bin respectivo de la segunda pluralidad correspondiente de valores de bin un bin correspondiente en la pluralidad de bins y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer. La segunda pluralidad correspondiente de valores de bin de cada sujeto respectivo de la segunda pluralidad de sujetos se obtiene de una muestra biológica correspondiente del sujeto respectivo, que comprende un segundo tipo de tejido, utilizando una segunda pluralidad correspondiente de lecturas de secuencia determinadas mediante un tercer método de secuenciación de ácido nucleico. En tales realizaciones, al menos el segundo método de secuenciación de ácido nucleico difiere del tercer método de secuenciación de ácido nucleico o el primer tipo de tejido difiere del segundo tipo de tejido.

En tales realizaciones, se obtiene una pluralidad de funciones de extracción de características aplicando una técnica de extracción de características a los valores de bin respectivos de los sujetos respectivos del primer conjunto de datos, identificando de este modo una pluralidad de funciones de extracción de características. Cada función de extracción de características de la pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal de los valores de bin de la totalidad o un subconjunto de la pluralidad de bins, y la pluralidad de funciones de extracción de características discrimina colectivamente a los sujetos respectivos de la primera pluralidad de sujetos según tengan una condición de cáncer dentro del conjunto de condiciones de cáncer basándose en los valores de bin respectivos para los sujetos respectivos.

Además, En tales realizaciones, cada función de extracción de características respectiva de la pluralidad de funciones de extracción de características se aplica con respecto a la segunda pluralidad respectiva de valores de bin de cada sujeto correspondiente de la segunda pluralidad de sujetos, contribuyendo de este modo al segundo conjunto de datos transformado con una pluralidad respectiva de valores de características para cada sujeto correspondiente. En algunas realizaciones, la pluralidad de valores de bin de prueba es un número de fragmentos representados por la pluralidad de lecturas de secuencia de prueba después de la aplicación de una o más condiciones de filtro. En algunas realizaciones, el primer método de secuenciación de ácido nucleico produce un patrón de metilación correspondiente para cada fragmento respectivo de la serie de fragmentos, y una condición de filtro de la una o más condiciones de filtro es la aplicación de un umbral de valor de p (p. ej., seleccionado del rango de entre 0,001 y 0,20) al patrón de metilación correspondiente, en donde el umbral de valor de p representa la frecuencia con la que se observa un patrón de metilación en una cohorte de sujetos sin cáncer.

En algunas realizaciones, el primer método de secuenciación de ácido nucleico produce un patrón de metilación correspondiente para cada fragmento respectivo de la serie de fragmentos, y una condición de filtro de la una o más condiciones de filtro es la aplicación del requisito de que el fragmento respectivo esté representado por un número umbral (p. ej., 2, 3, 4, 5, 6, 7, 8, 9, 10 o un entero entre 10 y 100) de lecturas de secuencia en la pluralidad de lecturas de secuencia de prueba.

En algunas realizaciones, el primer método de secuenciación de ácido nucleico produce un patrón de metilación correspondiente para cada fragmento respectivo de la serie de fragmentos, y una condición de filtro de la una o más condiciones de filtro es la aplicación del requisito de que el fragmento respectivo tenga un número umbral de sitios CpG (al menos 1, 2, 3, 4, 5, 6, 7, 8, 9 o 10 sitios CpG).

En algunas realizaciones, el primer método de secuenciación de ácido nucleico produce un patrón de metilación correspondiente para cada fragmento respectivo de la serie de fragmentos, y una condición de filtro de la una o más condiciones de filtro es un requisito de que el fragmento respectivo tenga una longitud inferior a un número umbral de pares de bases (p. ej., 1 mil, 2 mil, 3 mil o 4 mil pares de bases contiguas de longitud).

Breve descripción de los dibujos

Las implementaciones descritas en la presente memoria se ilustran a modo de ejemplo, y no a modo de limitación, en las figuras de los dibujos adjuntos. Números de referencia similares se refieren a partes correspondientes en todas las diversas vistas de los dibujos.

La Figura 1 es un diagrama en bloque que ilustra un ejemplo de un sistema informático según algunas realizaciones de la presente descripción.

La Figura 2 es un diagrama esquemático del procesamiento realizado según algunas realizaciones de la presente descripción.

La Figura 3 ilustra un ejemplo de estructuras de datos que representan un primer conjunto de datos y un segundo conjunto de datos, según algunas realizaciones de la presente descripción.

La Figura 4 ilustra un ejemplo del primer conjunto de datos de la Figura 3 y un ejemplo de una primera pluralidad de funciones de extracción de características generadas al aplicar una primera técnica de extracción de características al primer conjunto de datos, según algunas realizaciones de la presente descripción.

La Figura 5 ilustra la generación parcial de un segundo conjunto de datos transformado, utilizando la primera pluralidad de funciones de extracción de características y recuentos de bins para los sujetos del segundo conjunto de datos, según algunas realizaciones de la presente descripción.

La Figura 6 ilustra un ejemplo del segundo conjunto de datos de la Figura 3 y un ejemplo de una segunda pluralidad de funciones de extracción de características generadas al aplicar una segunda técnica de extracción de características al segundo conjunto de datos, según algunas realizaciones de la presente descripción.

La Figura 7 ilustra el segundo conjunto de datos transformado y las segundas características (mostradas parcialmente) obtenidas utilizando la segunda pluralidad de funciones de extracción de características, según algunas realizaciones de la presente descripción.

La Figura 8 ilustra el segundo conjunto de datos transformado que se utiliza para entrenar a un clasificador, según algunas realizaciones de la presente descripción.

La Figura 9 ilustra el segundo conjunto de datos transformado y la segunda pluralidad de funciones de extracción de características que se usan juntos para entrenar a un clasificador, según algunas realizaciones de la presente descripción.

Las Figuras 10A, 10B, 10C, 10D, 10E, 10F, 10G y 10H son diagramas de flujo que ilustran ejemplos de métodos de entrenamiento de un clasificador para discriminar entre una condición de cáncer de un conjunto de condiciones de cáncer, según algunas realizaciones de la presente descripción.

La Figura 11 es un gráfico de barras que ilustra una comparación de la capacidad de discernir un tejido de origen (TOO) utilizando un clasificador CCGA + TCGA (izquierda) o un clasificador CCGA WGS (derecha) a través de un conjunto de datos de prueba extraído del CCGA del ejemplo 2, según algunas realizaciones de la presente descripción.

La Figura 12 ilustra los resultados de la detección de varios tipos de cáncer utilizando un clasificador CCGA + TCGA frente a un clasificador CCGA WGS en un conjunto de datos de prueba extraído del estudio CCGA del ejemplo 2, según algunas realizaciones de la presente descripción.

La Figura 13 ilustra valores predictivos positivos en un conjunto de datos de prueba extraídos del estudio CCGA del ejemplo 2 para diversos cánceres respectivos utilizando un clasificador CCGA + TCGA frente a un clasificador CCGA WGS, según algunas realizaciones de la presente descripción.

La Figura 14 son gráficos que ilustran la sensibilidad de la predicción para varios tipos de cáncer utilizando un clasificador CCGA+ TCGA WGS frente a un clasificador CCGA WGS en un conjunto de datos de prueba extraído del estudio CCGA del ejemplo 2, según algunas realizaciones de la presente descripción.

La Figura 15 son gráficos que ilustran los valores predictivos positivos (PPV, por sus siglas en inglés) para varios tipos de cáncer utilizando un clasificador WGS CCGA+ TCGA frente a un clasificador WGS CCGA en un conjunto de datos de prueba extraído del estudio CCGA del ejemplo 2, según algunas realizaciones de la presente descripción.

La Figura 16 ilustra la aplicación de los métodos según la presente descripción para discriminar entre muestras de tejido canceroso y no canceroso. El gráfico de la izquierda ilustra la sensibilidad vs. (1-especificidad) de la predicción de la presencia de cánceres de alta señal utilizando los datos del clasificador CCGA WGS (la línea 1602 comienza con una sensibilidad de aproximadamente 0,2) y el clasificador de datos de aprendizaje por transferencia WGS CCGA + TCGA (la línea 1604 comienza con una sensibilidad de aproximadamente 0,4). El gráfico de la derecha ilustra los positivos verdaderos (gris claro) y falsos (negros) resultantes de la detección del cáncer utilizando el clasificador de aprendizaje por transferencia CCGA + TCGA y el clasificador CCGA. Como se muestra en el gráfico de la probabilidad del clasificador CCGA + TCGA vs. la probabilidad del clasificador CCGA, el enfoque de aprendizaje por transferencia descrito proporciona una alta sensibilidad a la detección del cáncer.

La Figura 17 ilustra la normalización por GC de los recuentos de bins, según la presente descripción.

La Figura 18 ilustra el uso de un límite de valor de p para seleccionar ciertos fragmentos para la contribución al recuento de bins, donde el valor de p se calcula utilizando un modelo de Markov entrenado utilizando los fragmentos de metilación de una cohorte de sujetos sin cáncer, según un aspecto de la presente descripción.

La Figura 19 es un diagrama de flujo que describe un proceso de secuenciación de ácido nucleico, según un aspecto de la presente descripción.

La Figura 20 es una ilustración de una parte del proceso de secuenciación de ácidos nucleicos para obtener información de metilación y vectores de estado de metilación, según un aspecto de la presente descripción.

La Figura 21 ilustra la generación de una estructura de datos para un grupo de control, según un aspecto de la presente descripción.

La Figura 22 ilustra un diagrama de flujo que describe un proceso para determinar fragmentos anómalamente metilados a partir de una muestra, según un aspecto de la presente descripción.

La Figura 23 es una ilustración de los bins (bloques) de un genoma de referencia, según un aspecto de la presente descripción.

Descripción detallada

Ahora se hará referencia en detalle a las realizaciones, cuyos ejemplos se ilustran en los dibujos adjuntos. En la siguiente descripción detallada se exponen numerosos detalles específicos para proporcionar una comprensión completa de la presente descripción. Sin embargo, resultará evidente para un experto en la técnica que la presente descripción puede ponerse en práctica sin estos detalles específicos. En otros casos, los métodos, procedimientos, componentes, circuitos y redes bien conocidos no se han descrito en detalle para no oscurecer innecesariamente aspectos de las realizaciones.

Las implementaciones descritas en la presente memoria proporcionan diversas soluciones técnicas para entrenar a un clasificador para que discrimine entre condiciones de cáncer de un conjunto de condiciones de cáncer. Las técnicas descritas utilizan un enfoque de aprendizaje por transferencia donde las funciones de extracción de características se aprenden de un conjunto de datos y se aplican a un nuevo conjunto de datos para obtener valores. Los valores caracterizan los datos del nuevo conjunto de datos y pueden usarse como base para entrenar un clasificador que asigne etiquetas a los sujetos de los que se obtiene el nuevo conjunto de datos. Los ejemplos de los valores incluyen una indicación de que los datos de una muestra biológica son indicativos de cáncer o una indicación de que los datos de una muestra biológica no son indicativos de cáncer (p. ej., “no cancerosos”).

Definiciones

Como se utiliza en la presente memoria, la expresión “patrón de metilación anormal” o “patrón de metilación anómalo” se refiere a un vector de estado de metilación o un estado de metilación de una molécula de ADN que tiene el vector de estado de metilación que se espera que se encuentre en una muestra menos frecuentemente que un valor umbral. En una realización particular 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 . De este modo, 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 en el grupo de control sano. 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 (p. ej., 0,1, 0,01, 0,001, 0,0001, etc.) puede definirse como un patrón de metilación anormal. 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. Los métodos proporcionados en la presente memoria usan regiones genómicas que tienen un patrón de metilación anómalo. Se puede determinar que una región genómica tiene un patrón de metilación anómalo cuando los fragmentos de ADNlc que corresponden a la región genómica o se originan en ella tienen vectores de estado de metilación que aparecen menos frecuentemente que un valor umbral en las muestras de referencia. Las muestras de referencia pueden ser muestras de sujetos de control o sujetos sanos. La frecuencia con la que aparece un vector de estado de metilación en las muestras de referencia se puede representar como una puntuación de valor de p . Cuando los fragmentos de ADNlc que corresponden a la región genómica o se originan en ella no tienen un único vector de estado de metilación uniforme, la región genómica puede tener múltiples puntuaciones de valor de p para múltiples vectores de estado de metilación. En este caso, los múltiples núcleos de valores de p se pueden sumar o promediar antes de compararlos con el valor umbral. Se pueden adoptar varios métodos conocidos en la técnica para comparar las puntuaciones de valor de p correspondientes a la región genómica y al valor umbral, que incluyen, aunque no de forma limitativa, la media aritmética, la media geométrica, la media armónica, la mediana, la moda, etc.

Como se utiliza en la presente memoria, el término “alrededor de” o “aproximadamente” puede significar dentro de un rango de error aceptable para el valor particular determinado por un experto en la técnica, que puede depender en parte de cómo se mida o determine el valor, p. ej., las limitaciones del sistema de medición. Por ejemplo, “aproximadamente” puede significar dentro de 1 o más de 1 desviación estándar, según la práctica en la técnica. “Aproximadamente” puede significar un rango de $\pm 20\%$, $\pm 10\%$, $\pm 5\%$ o $\pm 1\%$ de un valor dado. El término “aproximadamente” puede significar dentro de un orden de magnitud, dentro de 5 veces o dentro de 2 veces un valor. Cuando se describan valores particulares en la solicitud y en las reivindicaciones, a menos que se indique lo contrario, se debe suponer que el término “aproximadamente” significa dentro de un rango de error aceptable para el valor en particular. El término “aproximadamente” puede tener el significado comúnmente entendido por un experto en la técnica. El término “aproximadamente” puede referirse a $\pm 10\%$. El término “aproximadamente” puede referirse a $\pm 5\%$.

Como se utiliza en la presente memoria, el término “muestra biológica”, “muestra de paciente” o “muestra” se refiere a cualquier muestra tomada de un sujeto, que puede reflejar un estado biológico asociado con el sujeto, y que incluye ADN libre de células. Los ejemplos de muestras biológicas incluyen, aunque no de forma limitativa, sangre, sangre completa, plasma, suero, orina, líquido cefalorraquídeo, heces, saliva, sudor, lágrimas, líquido pleural, líquido pericárdico o líquido peritoneal del sujeto. Una muestra biológica puede incluir cualquier tejido o material obtenido de un sujeto vivo o muerto. Una muestra biológica puede ser una muestra libre de células. Una muestra biológica puede comprender un ácido nucleico (p. ej., ADN o ARN) o un fragmento del mismo. El término “ácido nucleico” puede referirse al ácido desoxirribonucleico (ADN), al ácido ribonucleico (ARN) o a cualquier híbrido o fragmento de los mismos. El ácido nucleico de la muestra puede ser un ácido nucleico libre de células. Una muestra puede ser una muestra líquida o una muestra sólida (p. ej., una muestra de células o tejidos). Una muestra biológica puede ser un fluido corporal, tal como sangre, plasma, suero, orina, secreciones vaginales, fluido de hidrocele (p. ej., de los testículos), líquido de lavado vaginal, líquido pleural, líquido ascítico, líquido cefalorraquídeo, saliva, sudor, lágrimas, esputo, líquido de lavado broncoalveolar, líquido de secreción del pezón, líquido de aspiración de diferentes partes del cuerpo (p. ej., tiroides, mama), etc. Una muestra biológica puede ser una muestra de heces. En varias realizaciones, la mayoría del ADN de una muestra biológica que se ha enriquecido para obtener ADN libre de células (p. ej., una muestra de plasma obtenida mediante un protocolo de centrifugación) puede estar libre de células (p. ej., más del 50 %, 60 %, 70 %, 80 %, 90 %, 95 % o 99 % del ADN puede estar libre de células). Una muestra biológica puede tratarse para alterar físicamente la estructura tisular o celular (p. ej., por centrifugación y/o lisis celular), liberando así los componentes intracelulares en una solución que puede contener además enzimas, reguladores, sales, detergentes y similares que pueden usarse para preparar la muestra para el análisis.

Como se utiliza en la presente memoria, el término “cáncer” o “tumor” se refiere a una masa anormal de tejido en la que el crecimiento de la masa excede y no se coordina con el crecimiento del tejido normal. Un cáncer o tumor se puede definir como “benigno” o “maligno” dependiendo de las siguientes características: grado de diferenciación celular, incluida la morfología y la funcionalidad, velocidad de crecimiento, invasión local y metástasis. Un tumor “benigno” puede estar bien diferenciado, tener un crecimiento característicamente más lento que un tumor maligno y permanecer localizado en el sitio de origen. Además, en algunos casos, un tumor benigno no tiene la capacidad de infiltrarse, invadir o hacer metástasis en sitios distantes. Un tumor “maligno” puede estar poco diferenciado (anaplasia), tener un crecimiento característicamente rápido acompañado de una infiltración, invasión y destrucción progresivas del tejido circundante. Además, un tumor maligno puede tener la capacidad de hacer metástasis en sitios distantes.

Como se utiliza en la presente memoria, el término “condición de cáncer” se refiere al cáncer de mama, cáncer de pulmón, cáncer de próstata, cáncer colorrectal, cáncer renal, cáncer de útero, cáncer de páncreas, cáncer de esófago, linfoma, cáncer de cabeza/cuello, cáncer de ovario, cáncer hepatobiliar, melanoma, cáncer de cuello uterino, mieloma múltiple, leucemia, cáncer de tiroides, cáncer de vejiga y cáncer gástrico. El término “condición de cáncer” también se refiere a una condición “no cancerosa” o a no tener cáncer o tener una condición no cancerosa. Una condición de cáncer puede ser una etapa predeterminada de un cáncer de mama, una etapa predeterminada de un cáncer de pulmón, una etapa predeterminada de un cáncer de próstata, una etapa predeterminada de un cáncer colorrectal, una etapa predeterminada de un cáncer renal, una etapa predeterminada de un cáncer de útero, una etapa predeterminada de un cáncer de páncreas, una etapa predeterminada de un cáncer de esófago, una etapa predeterminada de un linfoma, una etapa predeterminada de un cáncer de cabeza/cuello, una etapa predeterminada de un cáncer de ovario, una etapa predeterminada de un cáncer hepatobiliar, una etapa predeterminada de un melanoma, una etapa predeterminada de un cáncer de cuello uterino, una etapa predeterminada de un mieloma múltiple, una etapa predeterminada de una leucemia, una etapa predeterminada de un cáncer de tiroides, una etapa predeterminada de un cáncer de vejiga o una etapa predeterminada de un cáncer gástrico. Una condición de cáncer también puede ser una métrica de supervivencia, que puede ser una probabilidad predeterminada de supervivencia para un período de tiempo predeterminado. Por ejemplo, la medida de supervivencia se puede definir como la diferencia de tiempo (p. ej., años o meses) entre la fecha del diagnóstico inicial de una enfermedad o condición (p. ej., cáncer) hasta la fecha de fallecimiento del paciente debido a esa enfermedad o condición.

El término “clasificación” puede referirse a cualquier o cualesquiera número(s) u otro(s) carácter(es) que esté(n) asociado(s) con una propiedad particular de una muestra. Por ejemplo, un símbolo “+” (o la palabra “positivo”) puede significar que una muestra se clasifica como portadora de delecciones o amplificaciones. En otro ejemplo, el término

“clasificación” puede referirse a una cantidad de tejido tumoral en el sujeto y/o muestra, al tamaño del tumor en el sujeto y/o muestra, a una etapa del tumor en el sujeto, a una carga tumoral en el sujeto y/o muestra y a la presencia de metástasis tumoral en el sujeto. La clasificación puede ser binaria (p. ej., positiva o negativa) o tener más niveles de clasificación (p. ej., una escala del 1 al 10 o del 0 al 1). Los términos “límite” y “umbral” pueden referirse a números predeterminados utilizados en una operación. Por ejemplo, un tamaño límite puede referirse a un tamaño por encima del cual se excluyen los fragmentos. Un valor umbral puede ser un valor por encima o por debajo del cual se aplica una clasificación particular. Cualquiera de estos términos se puede utilizar en cualquiera de estos contextos.

Como se utiliza en la presente descripción, los términos “ácido nucleico libre de células”, “ADN libre de células” o “ADNlc” se refieren indistintamente a fragmentos de ácido nucleico que circulan en el cuerpo de un sujeto (p. ej., torrente sanguíneo) y se originan en una o más células sanas y/o en una o más células cancerosas.

Como se utiliza en la presente memoria, la expresión “ácidos nucleicos libres de células” se refiere a moléculas de ácido nucleico que se pueden encontrar fuera de las células, en fluidos corporales tales como sangre, sangre completa, plasma, suero, orina, líquido cefalorraquídeo, heces, saliva, sudor, lágrimas, líquido pleural, líquido pericárdico o líquido peritoneal de un sujeto. Los ácidos nucleicos libres de células se usan indistintamente como ácidos nucleicos circulantes. Los ejemplos de ácidos nucleicos libres de células incluyen, aunque no de forma limitativa, ARN, ADN mitocondrial o ADN genómico.

Como se usan en la presente memoria, los términos “control”, “muestra de control”, “referencia”, “muestra de referencia”, “normal” y “muestra normal” describen una muestra de un sujeto que no tiene una condición particular o que está sano por lo demás. En un ejemplo, un método tal como se describe en la presente memoria puede realizarse en un sujeto que tiene un tumor, donde la muestra de referencia es una muestra tomada de un tejido sano del sujeto. Se puede obtener una muestra de referencia del sujeto o de una base de datos. La referencia puede ser, p. ej., un genoma de referencia que se utiliza para mapear las lecturas de secuencia obtenidas de la secuenciación de una muestra del sujeto. Un genoma de referencia puede referirse a un genoma haploide o diploide con el que se pueden alinear y comparar las lecturas de secuencia de la muestra biológica y una muestra constitucional. Un ejemplo de muestra constitucional puede ser el ADN de los glóbulos blancos obtenidos del sujeto. Para un genoma haploide, solo puede haber un nucleótido en cada locus. Para un genoma diploide, se pueden identificar loci heterocigotos; cada locus heterocigoto puede tener dos alelos, donde cualquiera de los alelos puede permitir una coincidencia para la alineación con el locus.

Como se utiliza en la presente memoria, el término “sitio CpG” se refiere a una región de una molécula de ADN donde a un nucleótido de citosina le sigue 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; el fosfato une dos nucleótidos cualesquiera en el ADN. Las citosinas en los dinucleótidos CpG se pueden metilar para formar 5-metilcitosina.

Como se utiliza en la presente memoria, el término “falso positivo” (FP) se refiere a un sujeto que no tiene una condición. El falso positivo puede referirse a un sujeto que no tiene un tumor, un cáncer, una condición precancerosa (p. ej., una lesión precancerosa), un cáncer localizado o con metástasis, una enfermedad maligna o está sano por lo demás. El término falso positivo puede referirse a un sujeto que no tiene una condición, pero que se identifica como portador de la condición mediante un ensayo o método de la presente descripción.

Como se utiliza en la presente memoria, el término “falso negativo” (FN) se refiere a un sujeto que tiene una condición. Falso negativo puede referirse a un sujeto que tiene un tumor, un cáncer, una condición precancerosa (p. ej., una lesión precancerosa), un cáncer localizado o con metástasis, o una enfermedad maligna. El término falso negativo puede referirse a un sujeto que tiene una condición, pero se identifica como no portador de la condición mediante un ensayo o método de la presente descripción.

Como se utiliza en la presente memoria, el término “sano” se refiere a un sujeto que posee buena salud. Un sujeto sano puede demostrar la ausencia de cualquier enfermedad maligna o no maligna. Un “individuo sano” puede tener otras enfermedades o condiciones no relacionadas con la condición que se analiza y, normalmente, no puede considerarse “sano”.

Como se utiliza en la presente memoria, los términos “hipometilado” o “hipermetilado” 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 %, 85 %, 90 % o 95 %, o cualquier otro porcentaje dentro del rango de 50 %-100 %) no está metilado o está metilado, respectivamente.

Como se utiliza en la presente memoria, el término “nivel de cáncer” se refiere a si existe cáncer (p. ej., presencia o ausencia), una etapa de un cáncer, el tamaño del tumor, la presencia o ausencia de metástasis, una concentración estimada de la fracción tumoral, un valor de carga mutacional tumoral total, la carga tumoral total del cuerpo y/u otra medida de la gravedad de un cáncer (p. ej., la recurrencia del cáncer). El nivel de cáncer puede ser un número u otras marcas distintivas, tales como símbolos, letras del alfabeto y colores. El nivel puede ser cero. El nivel de cáncer también puede incluir condiciones (estados) premalignas o precancerosas asociadas con mutaciones o un número de

mutaciones. El nivel de cáncer se puede utilizar de varias maneras. Por ejemplo, las pruebas de detección pueden comprobar si hay cáncer en alguien que no se sabía previamente que tenía cáncer. La evaluación puede involucrar a una persona a la que se le ha diagnosticado cáncer en la investigación destinada a monitorizar la progresión del cáncer a lo largo del tiempo, estudiar la efectividad de las terapias o determinar el pronóstico. En una realización, el pronóstico se puede expresar como la probabilidad de que un sujeto muera de cáncer, o la probabilidad de que el cáncer progrese después de una duración o tiempo específico, o la probabilidad de que el cáncer haga metástasis. La detección puede comprender la “exploración” o puede comprender la verificación de si una persona con características que sugieren cáncer (p. ej., síntomas u otras pruebas positivas) tiene cáncer. Un “nivel de patología” puede referirse al nivel de patología asociado con un patógeno, donde el nivel puede ser el descrito anteriormente para el cáncer. Cuando el cáncer se asocia con un patógeno, el nivel de cáncer puede ser un tipo de nivel de patología.

Como se utiliza en la presente memoria, la expresión “vector de estado de metilación” se refiere 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, $\langle Mx, Mx+J, Mx+2 \rangle$, $\langle Mx, Mx+1, Ux+2 \rangle$, ..., $\langle Ux, Ux+1, Ux+2 \rangle$ 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.

Como se utiliza en la presente memoria, un “metiloma” puede ser una medida de la cantidad o el grado de metilación de ADN en una pluralidad de sitios o loci en un genoma. El metiloma puede corresponder a todo un genoma, a una parte sustancial de un genoma o a una o varias porciones relativamente pequeñas de un genoma. Un “metiloma tumoral” puede ser un metiloma de un tumor de un sujeto (p. ej., un ser humano). Un metiloma tumoral se puede determinar utilizando tejido tumoral o ADN tumoral libre de células en plasma. Un metiloma tumoral puede ser un ejemplo de un metiloma de interés. Un metiloma de interés puede ser un metiloma de un órgano que puede contribuir con ácido nucleico, p. ej., ADN, a un fluido corporal (p. ej., un metiloma de células cerebrales, huesos, pulmones, corazón, músculos, riñones, etc.). El órgano puede ser un órgano trasplantado.

Como se utiliza en la presente memoria, el término “índice de metilación” para cada sitio genómico (p. ej., un sitio CpG) puede referirse a la proporción de lecturas de secuencia que muestran metilación en el sitio con respecto al número total de lecturas que cubren ese sitio. La “densidad de metilación” de una región puede ser el número de lecturas en los sitios dentro de una región que muestran metilación dividido por el número total de lecturas que cubren los sitios de la región. Los sitios pueden tener características específicas (p. ej., los sitios pueden ser sitios CpG). La “densidad de metilación de CpG” de una región puede ser el número de lecturas que muestran la metilación de CpG dividido por el número total de lecturas que cubren los sitios CpG en la región (p. ej., un sitio CpG particular, sitios CpG dentro de una isla CpG o una región más grande). Por ejemplo, la densidad de metilación para cada bin de 100 kb en el genoma humano se puede determinar a partir del número total de citosinas no convertidas (que pueden corresponder a citosina metilada) en los sitios CpG como una proporción de todos los sitios CpG cubiertos por las lecturas de secuencia mapeadas en la región de 100 kb. Este análisis también se puede realizar para otros tamaños de bin, p. ej., 50 kb o 1 Mb, etc. Una región puede ser un genoma completo o un cromosoma o parte de un cromosoma (p. ej., un brazo cromosómico). El índice de metilación de un sitio CpG puede ser el mismo que la densidad de metilación de una región cuando la región solo incluye ese sitio CpG. La “proporción de citosinas metiladas” puede referirse al número de sitios de citosina, “C”, que se ha demostrado que están metilados (por ejemplo, no se convierten después de la conversión con bisulfito) con respecto al número total de residuos de citosina analizados, p. ej., incluidas las citosinas fuera del contexto de CpG, en la región. El índice de metilación, la densidad de metilación y la proporción de citosinas metiladas son ejemplos de “niveles de metilación”.

Como se utiliza en la presente memoria, el término “perfil de metilación” (también denominado estado de metilación) puede incluir información relacionada con la metilación de ADN para una región. La información relacionada con la metilación de ADN puede incluir un índice de metilación de un sitio CpG, una densidad de metilación de los sitios CpG en una región, una distribución de los sitios CpG en una región contigua, un patrón o nivel de metilación para cada sitio CpG individual dentro de una región que contiene más de un sitio CpG, y la metilación no CpG. Un perfil de metilación de una parte sustancial del genoma puede considerarse equivalente al metiloma. La “metilación de ADN” en los genomas de los mamíferos puede referirse a la adición de un grupo metilo en la posición 5 del anillo heterocíclico de la citosina (p. ej., para producir 5-metilcitosina) entre los dinucleótidos de CpG. La metilación de citosina puede ocurrir en las citosinas en

otros contextos de secuencia, por ejemplo 5'-CHG-3' y 5'-CHH-3', donde H es adenina, citosina o timina. La metilación de citosina también puede estar en forma de 5-hidroximetilcitosina. La metilación de ADN puede incluir la metilación de nucleótidos distintos de citosina, tales como la N6-metiladenina.

Como se utiliza en la presente memoria, el “valor predictivo negativo” o “NPV” se puede calcular mediante $TN/(TN+FN)$ o la fracción negativa verdadera de todos los resultados negativos de las pruebas. El valor predictivo negativo puede verse afectado intrínsecamente por la prevalencia de una condición en una población y la probabilidad previa a la prueba de la población que se pretende someter a las pruebas. El término “valor predictivo positivo” o “PPV” se puede calcular mediante $TP/(TP+FP)$ o la fracción verdadera positiva de todos los resultados positivos de las pruebas. El PPV puede verse afectado intrínsecamente por la prevalencia de una condición en una población y

por la probabilidad previa a la prueba de la población que pretende someter a las pruebas. Ver, p. ej., O'Marcaigh y Jacobson, "Estimating The Predictive Value of a Diagnostic Test, How to Prevent Misleading or Confusing Results", Clin. Ped. 1993, 32(8): 485-491.

Como se utiliza en la presente memoria, los términos "ácido nucleico" y "molécula de ácido nucleico" se usan indistintamente. Los términos se refieren a ácidos nucleicos de cualquier forma de composición, tales como ácido desoxirribonucleico (ADN, p. ej., ADN complementario (ADNc), ADN genómico (ADNg) y similares) y/o análogos de ADN (p. ej., que contienen análogos de bases, análogos de azúcar y/o una cadena principal no nativa y similares), todos los cuales pueden estar en forma monocatenaria o bicatenaria. A menos que se limite de otro modo, un ácido nucleico puede comprender análogos conocidos de nucleótidos naturales, algunos de los cuales pueden funcionar de manera similar a los nucleótidos de origen natural. Un ácido nucleico puede estar en cualquier forma útil para llevar a cabo procesos en la presente memoria (p. ej., lineal, circular, superenrollado, monocatenario, bicatenario y similares). En algunas realizaciones, un ácido nucleico puede proceder de un único cromosoma o fragmento del mismo (p. ej., una muestra de ácido nucleico puede proceder de un cromosoma de una muestra obtenida de un organismo diploide). En ciertas realizaciones, los ácidos nucleicos comprenden nucleosomas, fragmentos o partes de nucleosomas o estructuras similares a nucleosomas. Los ácidos nucleicos a veces comprenden proteínas (p. ej., histonas, proteínas de unión al ADN y similares). Los ácidos nucleicos analizados mediante los procesos descritos en la presente memoria a veces están sustancialmente aislados y no están sustancialmente asociados con proteínas u otras moléculas. Los ácidos nucleicos también incluyen derivados, variantes y análogos del ADN sintetizado, replicado o amplificado a partir de polinucleótidos monocatenarios ("sentido" o "antisentido", cadena "positiva" o cadena "negativa", marco de lectura "directo" o marco de lectura "inverso") y bicatenarios. Los desoxirribonucleótidos incluyen desoxiadenosina, desoxicitidina, desoxiguanosina y desoxitimidina. Se puede preparar un ácido nucleico utilizando un ácido nucleico obtenido de un sujeto como plantilla.

Como se utiliza en la presente memoria, el término "genoma de referencia" se refiere a cualquier genoma particular conocido, secuenciado o caracterizado, ya sea parcial o completo, de cualquier organismo o virus que pueda usarse para hacer referencia a secuencias identificadas de un sujeto. Los genomas de referencia ilustrativos utilizados para sujetos humanos, así como para muchos otros organismos, se proporcionan en el navegador de genomas en línea alojado en el Centro Nacional de Información Biotecnológica ("NCBI") o la Universidad de California en Santa Cruz (UCSC). Un "genoma" se refiere a la información genética completa de un organismo o virus, expresada en secuencias de ácido nucleico. Como se utiliza en la presente memoria, una secuencia de referencia o genoma de referencia frecuentemente es una secuencia genómica ensamblada o parcialmente ensamblada de un individuo o múltiples individuos. En algunas realizaciones, un genoma de referencia es una secuencia genómica ensamblada o parcialmente ensamblada de uno o más individuos humanos. El genoma de referencia puede verse como un ejemplo representativo del conjunto de genes de una especie. En algunas realizaciones, un genoma de referencia comprende secuencias asignadas a los cromosomas. Los genomas de referencia humanos ilustrativos incluyen, aunque no de forma limitativa, NCBI construcción 34 (equivalente de UCSC: hgl6), NCBI construcción 35 (equivalente de UCSC: hgl7), NCBI construcción 36.1 (equivalente de UCSC: hg18), GRCh37 (equivalente de UCSC: hgl9) y GRCh38 (equivalente de UCSC: hg38).

Como se utiliza en la presente memoria, el término "lecturas de secuencia" o "lecturas" se refiere a secuencias de nucleótidos producidas mediante cualquier proceso de secuenciación descrito en la presente memoria o conocido en la técnica. Las lecturas se pueden generar a partir de un extremo de los fragmentos de ácido nucleico ("lecturas de un solo extremo") y, a veces, se generan a partir de ambos extremos de los ácidos nucleicos (p. ej., lecturas de extremos emparejados, lecturas de doble extremo). En algunas realizaciones, las lecturas de secuencia (p. ej., lecturas de un solo extremo o de extremos emparejados) pueden generarse a partir de una o ambas cadenas de un fragmento de ácido nucleico objetivo. La longitud de la lectura de secuencia se asocia frecuentemente con la tecnología de secuenciación particular. Los métodos de alto rendimiento, por ejemplo, proporcionan lecturas de secuencia que pueden variar en tamaño desde decenas hasta cientos de pares de bases (pb). En algunas realizaciones, las lecturas de secuencia tienen una longitud media, mediana de longitud o longitud promedio de aproximadamente 15 pb a 900 pb de longitud (p. ej., aproximadamente 20 pb, aproximadamente 25 pb, aproximadamente 30 pb, aproximadamente 35 pb, aproximadamente 40 pb, aproximadamente 45 pb, aproximadamente 50 pb, aproximadamente 55 pb, aproximadamente 60 pb, aproximadamente 65 pb, aproximadamente 70 pb, aproximadamente 75 pb, aproximadamente 80 pb, aproximadamente 85 pb, aproximadamente 90 pb, aproximadamente 95 pb, aproximadamente 100 pb, aproximadamente 110 pb, aproximadamente 120 pb, aproximadamente 130, aproximadamente 140 pb, aproximadamente 150 pb, aproximadamente 200 pb, aproximadamente 250 pb, aproximadamente 300 pb, aproximadamente 350 pb, aproximadamente 400 pb, aproximadamente 450 pb, aproximadamente 500 pb). En algunas realizaciones, las lecturas de secuencia tienen una longitud media, mediana de longitud o longitud promedio de aproximadamente 1000 pb, 2000 pb, 5000 pb, 10.000 pb o 50.000 pb o más. La secuenciación por nanoporos, por ejemplo, puede proporcionar lecturas de secuencia que pueden variar en tamaño desde decenas hasta cientos o miles de pares de bases. La secuenciación paralela de Illumina puede proporcionar lecturas de secuencia que no varían tanto, por ejemplo, la mayoría de las lecturas de secuencia pueden ser inferiores a 200 pb. Una lectura de secuencia (o lectura de secuenciación) puede referirse a la información de secuencia correspondiente a una molécula de ácido nucleico (p. ej., una cadena de nucleótidos). Por ejemplo, una lectura de secuencia puede corresponder a una cadena de nucleótidos (p. ej., de aproximadamente 20 a aproximadamente 150) de parte de un fragmento de ácido nucleico, puede corresponder a una cadena de nucleótidos en uno o ambos

extremos de un fragmento de ácido nucleico, o puede corresponder a nucleótidos de todo el fragmento de ácido nucleico. Una lectura de secuencia se puede obtener de diversas formas, p. ej., utilizando técnicas de secuenciación o utilizando sondas, p. ej., en matrices de hibridación o sondas de captura, o técnicas de amplificación, tales como la reacción en cadena de la polimerasa (PCR) o la amplificación lineal utilizando un único cebador o amplificación isotérmica.

Como se utiliza en la presente memoria, los términos “secuenciación”, “determinación de secuencia” y similares, tal como se usan en la presente memoria, se refieren generalmente a todos y cada uno de los procesos bioquímicos que pueden usarse para determinar el orden de las macromoléculas biológicas, tales como los ácidos nucleicos o las proteínas. Por ejemplo, los datos de secuenciación pueden incluir todas o una porción de las bases nucleotídicas en una molécula de ácido nucleico, tal como un fragmento de ADN.

Como se utiliza en la presente memoria, el término “amplitud de secuenciación” se refiere a qué fracción de un genoma de referencia particular (p. ej., el genoma de referencia humano) o parte del genoma se ha analizado. El denominador de la fracción puede ser un genoma con repeticiones enmascaradas y, por lo tanto, el 100 % puede corresponder a todo el genoma de referencia menos las partes enmascaradas. Un genoma con repeticiones enmascaradas puede referirse a un genoma en el que las repeticiones de secuencia están enmascaradas (p. ej., las lecturas de secuencia se alinean con las porciones no enmascaradas del genoma). Se puede enmascarar cualquier parte de un genoma y, por lo tanto, uno puede centrarse en cualquier parte particular de un genoma de referencia. La secuenciación amplia puede referirse a la secuenciación y al análisis de al menos el 0,1 % del genoma.

Como se utiliza en la presente memoria, el término “profundidad de secuenciación” se refiere al número de veces que un locus se cubre con una lectura de secuencia consensuada correspondiente a una molécula diana de ácido nucleico única alineada con el locus; p. ej., la profundidad de secuenciación es igual al número de moléculas diana de ácido nucleico únicas que cubren el locus. El locus puede ser tan pequeño como un nucleótido, o tan grande como un brazo cromosómico, o tan grande como un genoma completo. La profundidad de secuenciación se puede expresar como “Yx”, p. ej., 50x, 100x, etc., donde “Y” se refiere al número de veces que un locus se cubre con una secuencia correspondiente a un objetivo de ácido nucleico; p. ej., el número de veces que se obtiene información de secuencia independiente que cubre el locus particular. En algunas realizaciones, la profundidad de secuenciación corresponde al número de genomas que se han secuenciado. La profundidad de secuenciación también se puede aplicar a múltiples loci, o a todo el genoma, en cuyo caso Y puede referirse al número medio o promedio de veces que se secuencia un loci o un genoma haploide, o un genoma completo, respectivamente. Cuando se cita una profundidad media, la profundidad real de los diferentes loci incluidos en el conjunto de datos puede abarcar un rango de valores. La secuenciación ultraprofunda puede referirse a una profundidad de secuenciación al menos 100 veces mayor en un locus.

Como se utiliza en la presente memoria, el término “sensibilidad” o “tasa de verdaderos positivos” (TPR) se refiere al número de verdaderos positivos dividido por la suma del número de verdaderos positivos y falsos negativos.

La sensibilidad puede caracterizar la capacidad de un ensayo o método para identificar correctamente una proporción de la población que realmente tiene una condición. Por ejemplo, la sensibilidad puede caracterizar la capacidad de un método para identificar correctamente el número de sujetos dentro de una población que tiene cáncer. En otro ejemplo, la sensibilidad puede caracterizar la capacidad de un método para identificar correctamente el uno o más marcadores indicativos de cáncer.

Como se utiliza en la presente memoria, el término “variante de un solo nucleótido” o “SNV” se refiere a una sustitución de un nucleótido en una posición (p. ej., sitio) de una secuencia de nucleótidos, p. ej., una secuencia correspondiente a una molécula de ácido nucleico diana de un individuo, por un nucleótido que es diferente del nucleótido en la posición correspondiente en un genoma de referencia. Una sustitución de una primera nucleobase X por una segunda nucleobase Y puede denominarse “X>Y”. Por ejemplo, una SNV de citosina por timina puede denominarse “C>T”. En algunas realizaciones, una SNV no produce un cambio en la expresión de los aminoácidos (una variante sinónima). En algunas realizaciones, una SNV produce un cambio en la expresión de aminoácidos (una variante no sinónima).

Como se utiliza en la presente memoria, los términos “perfil de tamaño” y “distribución de tamaño” pueden referirse a los tamaños de los fragmentos de ADN en una muestra biológica. Un perfil de tamaño puede ser un histograma que proporciona una distribución de una cantidad de fragmentos de ADN en una variedad de tamaños. Varios parámetros estadísticos (también denominados parámetros de tamaño o simplemente parámetros) pueden distinguir un perfil de tamaño de otro.

Un parámetro puede ser el porcentaje de fragmento de ADN de un tamaño o rango de tamaños particular con respecto a todos los fragmentos de ADN o con respecto a fragmentos de ADN de otro tamaño o rango.

Como se utiliza en la presente memoria, el término “especificidad” o “tasa de verdaderos negativos” (TNR) se refiere al número de verdaderos negativos dividido por la suma del número de verdaderos negativos y falsos positivos. La especificidad puede caracterizar la capacidad de un ensayo o método para identificar correctamente una proporción de la población que realmente no tiene una condición. Por ejemplo, la especificidad puede caracterizar la capacidad

de un método para identificar correctamente el número de sujetos dentro de una población que no tiene cáncer. En otro ejemplo, la especificidad caracteriza la capacidad de un método para identificar correctamente uno o más marcadores indicativos de cáncer.

5 Como se utiliza en la presente memoria, el término “sujeto” se refiere a cualquier organismo vivo o no vivo, que incluye, aunque no de forma limitativa, un animal humano (p. ej., un ser humano masculino, un ser humano femenino, un feto, una mujer embarazada, un niño o similar), un animal no humano, una planta, una bacteria, un hongo o un protista. Cualquier animal humano o no humano puede servir como sujeto, incluidos, aunque no de forma limitativa, mamíferos, reptiles, aves, anfibios, peces, ungulados, rumiantes, bovinos (p. ej., vacas), equinos (p. ej., caballos), caprinos y
10 ovinos (p. ej., ovejas, cabras), porcinos (p. ej., cerdos), camélidos (p. ej., camello, llama, alpaca), mono, simio (p. ej., gorila, chimpancé), ursido (p. ej., oso), aves de corral, perro, gato, ratón, rata, pez, delfín, ballena y tiburón. En algunas realizaciones, un sujeto es un sujeto masculino o un sujeto femenino de cualquier edad (p. ej., un hombre, una mujer o un niño).

15 Como se utiliza en la presente memoria, el término “tejido” puede corresponder a un grupo de células que se agrupan como una unidad funcional. Se puede encontrar más de un tipo de célula en un solo tejido. Los diferentes tipos de tejido pueden estar formados por diferentes tipos de células (p. ej., hepatocitos, células alveolares o células sanguíneas), pero también pueden corresponder a tejido de diferentes organismos (madre vs. feto) o a células sanas vs. células tumorales. El término “tejido” generalmente puede referirse a cualquier grupo de células que se encuentran
20 en el cuerpo humano (p. ej., tejido cardíaco, tejido pulmonar, tejido renal, tejido nasofaríngeo, tejido orofaríngeo). En algunos aspectos, el término “tejido” o “tipo de tejido” puede usarse para referirse a un tejido en el cual se origina un ácido nucleico libre de células. En un ejemplo, los fragmentos de ácido nucleico viral pueden derivarse del tejido sanguíneo. En otro ejemplo, los fragmentos de ácido nucleico viral pueden derivarse del tejido tumoral.

25 Como se utiliza en la presente memoria, el término “verdadero positivo” (TP) se refiere a un sujeto que tiene una condición.

“Verdadero positivo” puede referirse a un sujeto que tiene un tumor, un cáncer, una condición precancerosa (p. ej., una lesión precancerosa), un cáncer localizado o con metástasis, o una enfermedad no maligna. “Verdadero positivo”
30 puede referirse a un sujeto que tiene una condición, y se identifica como portador de la condición mediante un ensayo o método de la presente descripción.

Como se utiliza en la presente memoria, el término “verdadero negativo” (TN) se refiere a un sujeto que no tiene una condición o no tiene una condición detectable. El verdadero negativo puede referirse a un sujeto que no tiene una
35 enfermedad o una enfermedad detectable, tal como un tumor, un cáncer, una condición precancerosa (p. ej., una lesión precancerosa), un cáncer localizado o con metástasis, una enfermedad no maligna o un sujeto que por lo demás está sano. El verdadero negativo puede referirse a un sujeto que no tiene una condición o no tiene una condición detectable, o se identifica como no portador de la condición mediante un ensayo o método de la presente descripción.

40 La terminología utilizada en la presente memoria tiene el propósito de describir únicamente casos particulares y no pretende ser limitante. Como se utiliza en la presente memoria, la forma singular “un”, “una” y “el/la” también pretenden incluir las formas del plural, a menos que el contexto indique claramente lo contrario. Además, en la medida en que los términos “que incluye”, “incluye”, “tener”, “tiene”, “con” o variantes de los mismos se usen en la descripción
45 detallada y/o en las reivindicaciones, dichos términos pretenden ser inclusivos de una manera similar al término “que comprende”.

A continuación se describen varios aspectos con referencia a ejemplos de aplicaciones con fines ilustrativos. Debe entenderse que se exponen numerosos detalles, relaciones y métodos específicos para proporcionar una comprensión completa de las características descritas en la presente memoria. Sin embargo, un experto en la técnica pertinente
50 reconocerá fácilmente que las características descritas en la presente memoria se pueden poner en practicar sin uno o más de los detalles específicos o con otros métodos. Las características descritas en la presente memoria no se limitan al orden ilustrado de las acciones o eventos, ya que algunas acciones pueden ocurrir en un orden diferente y/o simultáneamente con otras acciones o eventos. Además, no todas las acciones o eventos ilustrados son necesarios para implementar una metodología según las características descritas en la presente memoria.

55 Realizaciones ilustrativas del sistema

Los detalles de un sistema ilustrativo se describen ahora junto con la Figura 1. La Figura 1 es un diagrama en bloque que ilustra un sistema 100 según algunas implementaciones. El sistema 100 en algunas implementaciones incluye al
60 menos una o más unidades de procesamiento CPU 102 (también denominadas procesadores), una o más interfaces 104 de red, una pantalla 106 que tiene una interfaz 108 de usuario, un dispositivo 110 de entrada, una memoria 111 y uno o más buses 114 de comunicación para interconectar estos componentes. El uno o más buses 114 de comunicación incluyen opcionalmente un sistema de circuitos (a veces denominado conjunto de chips) que interconectan y controlan las comunicaciones entre los componentes del sistema. La memoria 111 puede ser una
65 memoria no persistente, una memoria persistente o cualquier combinación de las mismas. La memoria no persistente típicamente incluye memoria de acceso aleatorio de alta velocidad, tal como DRAM, SRAM, RAM DDR, ROM,

EEPROM, memoria flash, mientras que la memoria persistente típicamente incluye CD-ROM, discos versátiles digitales (DVD) u otro almacenamiento óptico, casetes magnéticos, cinta magnética, almacenamiento en disco magnético u otros dispositivos de almacenamiento magnético, dispositivos de almacenamiento en disco magnético, dispositivos de almacenamiento en disco óptico, dispositivos de memoria flash u otros dispositivos de almacenamiento de estado sólido no volátil. Independientemente de su implementación específica, la memoria 111 comprende al menos un medio de almacenamiento legible por ordenador no transitorio, y almacena en el mismo instrucciones ejecutables por ordenador que pueden estar en forma de programas, módulos y estructuras de datos.

En algunas realizaciones, como se muestra en la Figura 1, la memoria 111 almacena lo siguiente:

- un sistema operativo 116, que incluye procedimientos para gestionar diversos servicios básicos del sistema y para realizar tareas dependientes del hardware;
- un módulo (o instrucciones) 118 de comunicación de red opcional para conectar el sistema 100 con otros dispositivos y/o a una red de comunicación;
- un módulo 120 de extracción de características para extraer funciones de extracción de características de los conjuntos de datos;
- un primer conjunto 122 de datos que comprende, para cada sujeto de una primera pluralidad de sujetos (sujeto 1-1, sujeto 1-2,... sujeto 1-Z), la primera información de genotipo (126-1-1,... 126-1-Z) que comprende una primera pluralidad de valores de bin (p. ej., recuentos de bin) (128-1-1,...) correspondiendo cada valor de bin respectivo a un bin de una pluralidad de bins (1, 2,..., N), y una indicación de una condición de cáncer (130-1-1,...) del sujeto respectivo (sujeto 1-1, sujeto 1-2,... sujeto 1-Z) de un conjunto de condiciones de cáncer;
- una primera pluralidad de funciones 132 de extracción de características que se basan en el primer conjunto 122 de datos, codificando independientemente cada función de extracción de características (132-1,..., 132-P) de la primera pluralidad de funciones de extracción de características una función lineal o no lineal de valores de bin (128-1-1,...) de la totalidad o un subconjunto de la pluralidad de bins;
- un segundo conjunto 124 de datos que comprende, para cada sujeto de una segunda pluralidad de sujetos (sujeto 2-1,... sujeto 2-T), una segunda información de genotipo (134-2-1,... 134-2-T) que comprende una segunda pluralidad de valores de bin (p. ej., recuentos de bin) (138-2-1,...), correspondiendo cada valor de bin respectivo a un bin de la pluralidad de bins (1, 2,..., N), y una indicación de una condición de cáncer (140-2-1,...) del sujeto respectivo (sujeto 2-1, sujeto 2-2,... sujeto 2-T) del conjunto de condiciones de cáncer;
- una segunda pluralidad de funciones 142 de extracción de características que se basan en el segundo conjunto 124 de datos, codificando independientemente cada función de extracción de características (142-1,..., 142-S) de la segunda pluralidad de funciones de extracción de características una función lineal o no lineal de valores de bin (138-2-1,...) de la totalidad o un subconjunto de la pluralidad de bins;
- un segundo conjunto 150 de datos transformado que comprende una pluralidad respectiva de valores de característica, también denominados en la presente memoria características (valor 152-1,... de característica, valor 152-T de característica), para cada sujeto correspondiente de la segunda pluralidad de sujetos (sujeto 2-1,... sujeto 2-T), el segundo conjunto 150 de datos transformado generado basándose en la segunda pluralidad respectiva de valores de bin (p. ej., recuentos de bin) (138-2-1,...) de cada sujeto correspondiente de la segunda pluralidad de sujetos (sujeto 2-1,... sujeto 2-T) con respecto a la función de extracción de características respectiva de la primera pluralidad de funciones 132 de extracción de características. Como se muestra en la Figura 1, los valores de características (valor 152-1,... de característica, valor 152-T de característica) comprenden una pluralidad de valores de características para cada sujeto, p. ej., los valores 152-1 de características para el sujeto 2-1 incluyen las características 154-2-1-2-1,..., 154-2-1-2-P, y los valores 152-T de características para el sujeto 2-T incluyen las características 154-2-T-2-1,..., 154-2-T-2-P), de modo que el segundo conjunto de datos transformado tiene el conjunto completo de valores de características para cada sujeto del segundo conjunto de datos; y
- un primer clasificador 160 entrenado en al menos un par de condiciones de cáncer del conjunto de condiciones de cáncer.

En diversas implementaciones, uno o más de los elementos identificados anteriormente se almacenan en uno o más de los dispositivos de memoria mencionados anteriormente y corresponden a un conjunto de instrucciones para realizar una función descrita anteriormente. Los módulos, datos o programas identificados anteriormente (p. ej., conjuntos de instrucciones) no necesitan implementarse como programas, procedimientos, conjuntos de datos o módulos de software separados y, por lo tanto, varios subconjuntos de estos módulos y datos pueden combinarse o

disponerse de otro modo en varias implementaciones. En algunas implementaciones, la memoria 111 almacena opcionalmente un subconjunto de los módulos y estructuras de datos identificados anteriormente. Además, en algunas realizaciones, la memoria almacena módulos y estructuras de datos adicionales no descritos anteriormente. En

algunas realizaciones, uno o más de los elementos identificados anteriormente se almacenan en un sistema informático distinto del sistema 100, que es direccionable para el sistema 100, de modo que el sistema 100 puede recuperar la totalidad o una porción de dichos datos cuando sea necesario.

Aunque la Figura 1 representa un “sistema 100”, la figura pretende ser más una descripción funcional de las diversas características que pueden estar presentes en los sistemas informáticos que un esquema estructural de las implementaciones descritas en la presente memoria. En la práctica, y como reconocen los expertos en la técnica, los elementos que se muestran por separado se pueden combinar y algunos elementos se pueden separar. Además, aunque la Figura 1 representa ciertos datos y módulos en la memoria 111 (que puede ser una memoria persistente o no persistente), debe apreciarse que estos datos y módulos, o porción(es) de los mismos, pueden almacenarse en más de una memoria. Por ejemplo, en algunas realizaciones, al menos el primer conjunto 122 de datos, el segundo conjunto 124 de datos, la primera pluralidad de funciones 132 de extracción de características y la segunda pluralidad de funciones 142 de extracción de características se almacenan en un dispositivo de almacenamiento remoto que puede formar parte de una infraestructura basada en la nube. En algunas realizaciones, al menos el primer conjunto 122 de datos y el segundo conjunto 124 de datos se almacenan en una infraestructura basada en la nube. En algunas realizaciones, el segundo conjunto 150 de datos transformado y el primer clasificador 160 también se pueden almacenar en el(los) dispositivo(s) de almacenamiento remoto.

Si bien se ha descrito un ejemplo de un sistema según la presente descripción con referencia a la Figura 1, ahora se detallan los métodos según la presente descripción. Cualquiera de los métodos según las realizaciones de la presente descripción puede hacer uso de cualquiera de los ensayos, algoritmos o técnicas, o combinaciones de los mismos, descritos en la solicitud de patente estadounidense n.º 15/793.830, presentada el 25 de octubre de 2017 y/o en la solicitud de patente internacional n.º PCT/US17/58099, presentada el 24 de octubre de 2017, para determinar una condición de cáncer en un sujeto de prueba o la probabilidad de que el sujeto tenga la condición de cáncer.

La Figura 2 ilustra una visión general de las técnicas según algunas realizaciones de la presente descripción. En las realizaciones descritas, se emplea un enfoque de aprendizaje por transferencia en el que las funciones de extracción de características obtenidas de un primer conjunto de datos se utilizan para entrenar a un clasificador para clasificar una condición de cáncer basándose en los datos de un segundo conjunto de datos. En algunas realizaciones, el primer conjunto de datos se obtiene del TCGA (Ejemplo 1) y el segundo conjunto de datos se obtiene del CCGA (Ejemplo 2). Sin embargo, debe apreciarse que las realizaciones según la presente descripción no se limitan a conjuntos de datos específicos. Por ejemplo, en algunas realizaciones, el aprendizaje por transferencia se utiliza entre un primer conjunto de datos que es el recuento del número de copias de fragmentos agrupados en bins y un segundo conjunto de datos que es el recuento de fragmentos de metilación aberrantes agrupados en bins (p. ej., tras la aplicación de un filtro de valor de p como se describe en la presente memoria).

Independientemente del tipo o tipos específicos de conjuntos de datos empleados según las técnicas descritas, en algunas realizaciones, las funciones de extracción de características se aprenden del primer conjunto de datos y las funciones de extracción de características se utilizan para extraer características basándose en el segundo conjunto de datos. Las técnicas pueden implementarse en cualquier sistema informático adecuado que comprenda al menos un procesador y una memoria que almacene al menos un programa para la ejecución por parte del al menos un procesador. Como se muestra en la Figura 2, en algunas realizaciones, se utiliza un primer conjunto 122 de datos que se genera utilizando muestras 204 biológicas obtenidas de cada sujeto respectivo de una primera pluralidad de sujetos. El primer conjunto 122 de datos puede incluir, para cada sujeto respectivo de una primera pluralidad de sujetos de una especie (p. ej., un ser humano), la primera información de genotipo correspondiente, tal como, p. ej., la información de genotipado 126-1-1,..., 126-1-Z mostrada en la Figura 1. La información de genotipado del primer conjunto de datos comprende (i) una primera pluralidad correspondiente de valores de bin, cada valor de bin respectivo de la primera pluralidad correspondiente de valores de bin para un bin correspondiente de una pluralidad de bins y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer, donde cada bin de la pluralidad de bins representa una porción de un genoma de referencia de la especie. Por lo tanto, la Figura 3 ilustra un ejemplo del primer conjunto de datos que comprende, para cada sujeto 1-1, 1-2,..., 1-Z de la primera pluralidad de sujetos, una primera pluralidad correspondiente de valores de bin para cada sujeto (cnt_{1-1-1} , cnt_{1-1-2} , cnt_{1-1-3} , ..., cnt_{1-1-N} para el sujeto 1-1; cnt_{1-2-1} , cnt_{1-2-2} , cnt_{1-2-3} , ..., cnt_{1-2-N} para el sujeto 1-2; ..., cnt_{1-Z-1} , cnt_{1-Z-2} , cnt_{1-Z-3} , ..., cnt_{1-Z-N} para el sujeto 1-Z) y una indicación respectiva de una condición de cáncer (un “valor” o una “etiqueta”) (L_{1-1} , L_{1-2} , ..., L_{1-Z}) para cada sujeto.

La Figura 23 es una ilustración de bins de un genoma de referencia, según algunas realizaciones de la presente descripción. Un genoma de referencia (o un subconjunto del genoma de referencia) se divide en una o más etapas, p. ej., para casos de uso que implican un ensayo de metilación dirigido (p. ej., cuando el primer y/o el segundo conjunto de datos incluyen datos de metilación agrupados en bins). Por ejemplo, en algunas realizaciones, el genoma de referencia se divide en bins (bloques) de sitios CpG. En algunas de tales realizaciones, cada bin queda definido cuando hay una separación entre dos sitios CpG adyacentes que supera un umbral, p. ej., más de 200 pares de bases (pb), 300 pb, 400 pb, 500 pb, 600 pb, 700 pb, 800 pb, 900 pb o 1000 pb, entre otros valores. Los bins pueden variar en función del tamaño de los pares de bases. En el caso donde el primer o el segundo conjunto de datos sean datos de metilación de la secuenciación dirigida, un tamaño común para los bins es de alrededor de 200 pb, con un rango de aproximadamente 30 pb a aproximadamente 1000 pb o más. En algunas realizaciones, cada bin está entre 30 pb y

5000 pb. En algunas realizaciones, cuando un bin respectivo de una pluralidad de bins es mayor que un tamaño umbral (p. ej., 900 pb, 1000 pb, 1100 pb, etc.), el bin respectivo se subdivide en ventanas de una longitud determinada, p. ej., 500 pb, 600 pb, 700 pb, 800 pb, 900 pb, 1000 pb, 1100 pb, 1200 pb, 1300 pb, 1400 pb o 1500 pb, entre otros valores, y cada una de dichas ventanas recibe su propio valor de bin independiente. En otras realizaciones, las ventanas pueden tener una longitud de 200 pb a 10 kilopares de bases (kpb), de 500 pb a 2 kpb, o de aproximadamente 1 kpb. Las ventanas (p. ej., adyacentes) pueden superponerse en varios pares de bases o en un porcentaje de la longitud, p. ej., el 10 %, el 20 %, el 30 %, el 40 %, el 50 % o el 60 %, entre otros valores. En algunas realizaciones, donde un bin se divide en una pluralidad de ventanas, cada función de extracción de características de la presente descripción codifica independientemente una función lineal o no lineal de los valores de ventana para cada una de las ventanas del bin respectivo. En algunas realizaciones, en lugar de dividir los bins más grandes en ventanas, dichos bins más grandes se dividen en bins más pequeños. En algunas realizaciones, dichos bins más pequeños se superponen entre sí, mientras que en otras realizaciones no se superponen entre sí.

Continuando con el bloque 1002 de la Figura 10 A, cada muestra de las muestras 204 biológicas de los primeros sujetos puede comprender un primer tipo de tejido, y la primera pluralidad correspondiente de valores de bin de cada sujeto respectivo de la primera pluralidad de sujetos puede generarse utilizando una primera pluralidad correspondiente de lecturas de secuencia determinadas mediante un primer método de secuenciación de ácido nucleico.

Como se muestra esquemáticamente en la Figura 2, se puede obtener y proporcionar una primera pluralidad de funciones 132 de extracción de características (“primeras funciones de extracción de características”) basándose en el primer conjunto 122 de datos. La primera pluralidad de funciones 132 de extracción de características se puede obtener aplicando una primera técnica de extracción de características a los valores de bin respectivos de los sujetos respectivos del primer conjunto 122 de datos. En algunas realizaciones, cada función de extracción de características de la primera pluralidad de funciones 132 de extracción de características codifica independientemente una función lineal o no lineal de los valores de bin de la totalidad o un subconjunto de la pluralidad de bins. La Figura 4 ilustra un ejemplo de la primera pluralidad de funciones 132 de extracción de características, donde cada función 1-1 de extracción de características, 1-2,..., 1-P se muestra esquemáticamente basándose en valores de bin respectivos, de modo que cada función de extracción de características codifica una función lineal o no lineal de los valores de bin respectivos de todos o un subconjunto de los bins 1,..., N. Por lo tanto, en este ejemplo, la función 1-1 de extracción de características se basa en los valores w_{1-1-1}^* (bin 1), w_{1-1-2}^* (bin 2), w_{1-1-3}^* (bin 3), w_{1-1-4}^* (bin 4), ..., w_{1-1-N}^* (bin N), donde, en algunas realizaciones, w_{1-1-1} , w_{1-1-2} , w_{1-1-3} , w_{1-1-4} , ..., w_{1-1-N} son los pesos respectivos de la función de extracción de características y bin 1... bin N son los recuentos de bin medidos de los bins correspondientes 1... N. En este ejemplo, la función 1-2 de extracción de características se basa en los valores w_{1-2-1}^* (bin 1), w_{1-2-2}^* (bin 2), w_{1-2-3}^* (bin 3), w_{1-2-4}^* (bin 4), ..., w_{1-2-N}^* (bin N); y la función de extracción de características se basa en los valores w_{1-P-1}^* (bin 1), w_{1-P-2}^* (bin 2), w_{1-P-3}^* (bin 3), w_{1-P-4}^* (bin 4), ..., w_{1-P-N}^* (bin N). Se apreciará que estos son ejemplos lineales de funciones de extracción de características y que la presente descripción abarca además ejemplos no lineales de funciones de extracción de características.

Independientemente de la forma específica en la que se generen las funciones de extracción de características, en algunas realizaciones, la primera pluralidad de funciones de extracción de características discrimina colectivamente a los sujetos respectivos de la primera pluralidad de sujetos como portadores de una condición de cáncer dentro de un conjunto de condiciones de cáncer basándose en los valores de bin respectivos para los sujetos respectivos. En algunas realizaciones, los métodos descritos permiten discriminar a un sujeto según tenga una condición de cáncer del conjunto de condiciones de cáncer frente a otra condición de cáncer del conjunto de condiciones de cáncer. En algunas realizaciones, los métodos descritos permiten discriminar a un sujeto según tenga una condición de cáncer del conjunto de condiciones de cáncer frente a todas las demás condiciones (cancerosas o no cancerosas) del conjunto de condiciones de cáncer.

En algunas realizaciones, la aplicación de la primera técnica de extracción de características comprende aplicar un algoritmo de reducción de dimensiones a los sujetos del primer conjunto 122 de datos.

La Figura 2 también ilustra que se puede utilizar un segundo conjunto 124 de datos que se genera utilizando muestras 224 biológicas obtenidas de cada sujeto respectivo de una segunda pluralidad de sujetos. El segundo conjunto 124 de datos puede incluir, para cada sujeto respectivo de una segunda pluralidad de sujetos de una especie (p. ej., un ser humano), la segunda información de genotipo correspondiente, tal como, p. ej., la información de genotipado 134-1,..., 134-1-T mostrada en la Figura 1. La información de genotipado en el segundo conjunto 124 de datos comprende (i) una segunda pluralidad correspondiente de valores de bin, cada valor de bin respectivo de la segunda pluralidad correspondiente de valores de bin para un bin correspondiente de una pluralidad de bins y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer, donde cada bin de la pluralidad de bins representa una porción de un genoma de referencia de la especie. Por lo tanto, la Figura 3, panel inferior, ilustra un ejemplo del segundo conjunto de datos que comprende, para cada sujeto 2-1, 2-2,..., 2-T de la segunda pluralidad de sujetos, una segunda pluralidad correspondiente de valores de bin para cada sujeto (cnt 2-1-1, cnt 2-1-2, cnt 2-1-3,..., cnt 2-1-N para el sujeto 2-1); cnt2-2-1, cnt2-2-2, cnt2-2-3, ..., cnt2-2-N para el sujeto 2-2; ...; cnt 2-T-1, cnt 2-T-2, cnt 2-T-3,..., cnt 2-T-N para el sujeto 2-T) y una indicación de una condición de cáncer (un valor o una “etiqueta”) (L 2-1, L 2-2,..., L 2-T) para cada sujeto.

En algunas realizaciones, cada muestra en las muestras 224 biológicas de los segundos sujetos puede comprender un segundo tipo de tejido, y se puede utilizar un segundo método de secuenciación para generar la correspondiente segunda pluralidad de valores de bin de cada sujeto respectivo de la segunda pluralidad de sujetos. Además, en algunas realizaciones, el primer método de secuenciación difiere del segundo método de secuenciación o el primer tipo de tejido difiere del segundo tipo de tejido. Por ejemplo, en algunas realizaciones, el primer método de secuenciación es la secuenciación del genoma completo o dirigida y la información de bin del primer conjunto de datos comprende los recuentos del número de copias de fragmento de ácido nucleico libre de células y el segundo conjunto de datos es la secuenciación con bisulfitos del genoma completo o del genoma objetivo y el segundo conjunto de datos comprende los recuentos de fragmentos de metilación aberrantes, donde dichos recuentos de fragmentos de metilación aberrantes se determinan utilizando los métodos descritos en la presente memoria, tales como los de los ejemplos 5 y 10 siguientes.

En algunas realizaciones, las funciones de extracción de características aprendidas del primer conjunto de datos se aplican al segundo conjunto de datos para extraer características del segundo conjunto de datos. En algunas realizaciones, como se mencionó anteriormente, el primer conjunto de datos se obtiene del TCGA y el segundo conjunto de datos se obtiene del CCGA, aunque se puede(n) utilizar otro(s) tipo(s) de conjuntos de datos adicional o alternativamente. Como se muestra en la Figura 2, se puede generar un segundo conjunto 150 de datos transformado basándose de la primera pluralidad de funciones 132 de extracción de características aplicando la primera pluralidad de funciones 132 de extracción de características a la segunda pluralidad respectiva de valores de bin de cada sujeto correspondiente de la segunda pluralidad de sujetos (el segundo conjunto 124 de datos). La segunda pluralidad respectiva de valores de bin de cada sujeto correspondiente de la segunda pluralidad de sujetos se transforma con respecto a la función de extracción de características respectiva de la primera pluralidad de funciones 132 de extracción de características. El segundo conjunto de datos transformado comprende una pluralidad respectiva de características (o valores de característica) para cada sujeto correspondiente. De esta manera, la Figura 5 muestra, a manera de ejemplo, representaciones esquemáticas de valores (2-1-2-1, 2-1-2-2,..., 2-1-2-P) de características para el sujeto 2-1 y valores (2-2-2-1, 2-2-2-2,..., 2-2-2-P) de características para el sujeto 2-2 del segundo conjunto 150 de datos transformado. Los valores de características se generan de forma similar para cada uno de los sujetos.

En algunas realizaciones, una característica (también denominada en la presente memoria “valor de característica”) es el resultado computacional de la entrada de recuentos de bin (u otros valores de bin) en una función de extracción de características. Como se ha descrito anteriormente, una función de extracción de características puede ser una combinación lineal o no lineal de valores de bin. Los valores de características pueden determinar colectivamente un vector para el sujeto. Por ejemplo, en las realizaciones en las que cada función de extracción de características es un componente principal, cada valor de característica incluye los valores de bin proyectados sobre el componente principal particular.

El segundo conjunto 150 de datos transformado se usa, junto con la indicación de la condición de cáncer de los sujetos respectivos de la segunda pluralidad de sujetos (segundo conjunto 124 de datos, Figuras 2 y 3) para entrenar a un clasificador 160 en al menos un par de condiciones de cáncer del conjunto de condiciones de cáncer. Como se muestra en la Figura 2, el clasificador 160 (p. ej., el primer clasificador 160 de la Figura 1), que puede comprender uno o más clasificadores (incluidos los clasificadores de diferentes tipos), se entrena utilizando etiquetas de clase de cáncer para la información de genotipo obtenida de una muestra biológica de cada sujeto del segundo conjunto de datos, p. ej., las etiquetas de clase L_{2-1} , L_{2-2} , ..., L_{2-T} (Figura 3) para cada sujeto. La Figura 8 ilustra un ejemplo de un segundo conjunto de datos transformado que incluye valores de características para el sujeto 2-1 (valores 2-1-2-1, 2-1-2-2,..., 2-1-2-P de características), el sujeto 2-2 (valores 2-2-2-1, 2-2-2-2,..., 2-2-2-P de características),..., el sujeto 2-T (valores 2-T-2-1, 2-T-2-2,..., 2-T-2-P de características). Como también se muestra esquemáticamente en la Figura 8, el segundo conjunto de datos transformado se aplica, junto con las etiquetas de clase reales (L_{1-1-a} , L_{1-2-a} , ..., L_{1-Z-a}), para entrenar a un clasificador de modo que pueda generar etiquetas de clase predichas (L_{1-1-p} , L_{1-2-p} , ..., L_{1-T-p}).

El clasificador 160 puede ser un clasificador de cualquier tipo adecuado. Por ejemplo, el clasificador 160 puede ser una regresión (p. ej., una regresión logística multivariante), una red neuronal, una red neuronal convolucional, una máquina de vectores de soporte, un árbol de decisiones o una técnica de agrupamiento.

En algunas de tales realizaciones, dichos clasificadores generan una única condición de cáncer para un conjunto de condiciones de cáncer determinado al introducir los valores de características de un sujeto respectivo del segundo conjunto de datos transformado. En algunas de tales realizaciones, dichos clasificadores generan una probabilidad separada de cada condición de cáncer en la condición de cáncer determinada al introducir los valores de características para un sujeto respectivo del segundo conjunto de datos transformado. La condición de cáncer (o la probabilidad de condición de cáncer) de cada condición de cáncer del conjunto de condiciones de cáncer generadas por el clasificador durante el entrenamiento se comparan con las condiciones de cáncer reales de los sujetos del segundo conjunto de datos transformado. En algunas realizaciones, particularmente en los clasificadores en forma de redes neuronales convolucionales, los errores en la asignación de condición de cáncer cometidos por el clasificador, verificados con respecto al segundo conjunto de datos transformado, se retropropagan a través de los pesos del clasificador para entrenar al clasificador. Por ejemplo, en el caso ilustrativo donde el clasificador es una red neuronal convolucional, los pesos de filtro de los respectivos filtros en las capas convolucionales de la red se ajustan en dicha

retropropagación. En una realización ilustrativa, el clasificador es una red neuronal entrenada contra los errores en las asignaciones de condiciones de cáncer realizadas por el clasificador, en vista de las condiciones de cáncer reales de los sujetos del segundo conjunto de datos transformado, mediante un descenso de gradiente estocástico con el método de aprendizaje adaptativo AdaDelta (Zeiler, 2012 “ADADELTA: an adaptive learning rate method”, CorR, vol. abs/1212.5701), y el algoritmo de retropropagación proporcionado en Rumelhart y col., 1988, “Neurocomputing: Foundations of research”, cap. Learning Representations by Back- propagating Errors, págs. 696-699, Cambridge, MA, USA: editorial MIT Press.

El clasificador 160 entrenado se utiliza para clasificar a un sujeto de prueba con respecto a una condición de cáncer utilizando como entrada al clasificador información de genotipo sobre el sujeto de prueba, basándose la información de genotipado en datos de lectura de secuencia en una muestra biológica de prueba obtenida del sujeto de prueba. De esta manera, el clasificador 160 se utiliza para predecir un valor 214 relacionado con el cáncer (p. ej., una etiqueta que indica una condición de cáncer o no cáncer, una etapa del cáncer, el tiempo de supervivencia, etc.) para el sujeto de prueba, como se muestra en la Figura 2.

Además, como también se muestra en la Figura 2, en algunas realizaciones, además de aplicar la primera técnica de extracción de característica a los valores de bin respectivos de los sujetos respectivos del primer conjunto 122 de datos, se aplica una segunda técnica de extracción de característica a los respectivos valores de bin de los sujetos respectivos del segundo conjunto 124 de datos, identificando de este modo una segunda pluralidad de funciones 142 de extracción de características (“funciones de extracción de segundas características”). La Figura 6 ilustra un ejemplo de la segunda pluralidad de funciones 142 de extracción de características, donde cada función 2-1, 2-2,..., 2-S de extracción de características se muestra esquemáticamente basándose en valores de bin respectivos, de modo que cada función de extracción de características codifica una función lineal o no lineal de los valores de bin respectivos de 53 todos o un subconjunto de los bins 1,..., N. Cada función de extracción de características de la segunda pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal de los valores de bin de la totalidad o un subconjunto de la pluralidad de bins. En este ejemplo, la función 2-1 de extracción de características se basa en los valores de bin $W_{2-1-1} * (\text{bin } 1)$, $W_{2-1-2} * (\text{bin } 2)$, $W_{2-1-3} * (\text{bin } 3)$, $W_{2-1-4} * (\text{bin } 4)$,..., $W_{2-1-N} * (\text{bin } N)$; la función 2-2 de extracción de características se basa en los valores de bin $W_{2-2-1} * (\text{bin } 1)$, $W_{2-2-2} * (\text{bin } 2)$, $W_{2-2-3} * (\text{bin } 3)$, $W_{2-2-4} * (\text{bin } 4)$,..., $W_{2-2-N} * (\text{bin } N)$; y la función 2-S de extracción de características se basa en los valores de bin $W_{2-S-1} * (\text{bin } 1)$, $W_{2-S-2} * (\text{bin } 2)$, $W_{2-S-3} * (\text{bin } 3)$, $W_{2-S-4} * (\text{bin } 4)$,..., $W_{2-S-N} * (\text{bin } N)$. Se apreciará que estos son ejemplos lineales de funciones de extracción de características y que la presente descripción abarca además ejemplos no lineales de funciones de extracción de características. La segunda pluralidad de funciones de extracción de características discrimina colectivamente a los sujetos respectivos de la segunda pluralidad de sujetos según tengan una condición de cáncer dentro del conjunto de condiciones de cáncer basándose en los valores de bin respectivos para los sujetos respectivos.

En algunas realizaciones, como se muestra en la Figura 2, las características o valores de características obtenidos utilizando la segunda pluralidad de funciones 142 de extracción de características se incluyen en el segundo conjunto 150 de datos transformado para entrenar el clasificador 160. La Figura 7 ilustra las segundas características obtenidas utilizando la segunda pluralidad de funciones 142 de extracción de características para el sujeto 2-1 y utilizadas junto con el segundo conjunto 150 de datos transformado. Como se muestra, los segundos valores de características (2-1-2'-1,..., 2-1-2'-S) (o la segunda pluralidad de valores de característica) obtenidos utilizando las segundas funciones 142 de extracción de características se añaden a los valores (2-1-2-1,..., 2-1-2-P) de características en el segundo conjunto 150 de datos transformado que se obtuvieron basándose en las primeras funciones 132 de extracción de características. La Figura 9 ilustra un ejemplo de un segundo conjunto de datos transformado que incluye valores de características para el sujeto 2-1 (valores 2-1-2-1,..., 2-1-2-P de características), el sujeto 2-2 (valores 2-2-2-1,..., 2-2-P de características),..., el sujeto 2-T (valores 2-T-2-1,..., 2-T-2-P de características), combinados con segundos valores de características para el sujeto 2-1 (valores 2-1-2'-1,..., 2-1-2'-S de características), sujeto 2-2 (valores de características 2-2-2'-1,..., 2-2-2'-S),..., sujeto 2-T (valores 2-T-2'-1,..., 2-T-2'-S de características). Como también se muestra esquemáticamente en la Figura 9, el segundo conjunto de datos transformado y la segunda pluralidad de valores de características se aplican, junto con valores o etiquetas reales (L_{1-1-a} , L_{1-2-a} , ..., L_{1-Z-a}), para entrenar a un clasificador de modo que pueda generar valores o etiquetas predichos (L_{1-1-p} , L_{1-2-p} , ..., L_{1-T-p}).

El segundo conjunto 124 de datos se puede aplicar a las segundas funciones 142 de extracción de características para producir las segundas características (o los segundos valores de característica) que se pueden utilizar además del segundo conjunto 150 de datos transformado para entrenar al clasificador 160 de diferentes maneras. De esta manera, el segundo conjunto 150 de datos transformado puede comprender los segundos valores de característica. Por ejemplo, en las realizaciones en las que el primer conjunto de datos es un conjunto de datos del TCGA y el segundo conjunto de datos es un conjunto de datos del CCGA, las funciones de extracción de características basadas en el TCGA y las funciones de extracción de características basadas en el CCGA se aplican por separado a los datos del CCGA antes de combinarse como entrada en el clasificador. En realizaciones alternativas en las que el primer conjunto de datos es un conjunto de datos del TCGA y el segundo conjunto de datos es un conjunto de datos del CCGA, utilizar la segunda pluralidad de funciones de extracción de características además del segundo conjunto de datos transformado para entrenar al primer clasificador implica aplicar la segunda pluralidad de funciones de extracción de características con respecto al segundo conjunto de datos transformado antes de entrenar al clasificador. En otras palabras, las funciones de extracción de características basadas en el TCGA se aplican a los datos del CCGA para

obtener datos del CCGA procesados, y las funciones de extracción de características basadas en el CCGA se aplican luego a los datos del CCGA ya procesados.

En algunas realizaciones, las primeras funciones 132 de extracción de características se pueden podar para generar un número reducido de funciones 208 de extracción de características, como se muestra en la Figura 2. Esto se puede hacer utilizando las primeras funciones 132 de extracción de características y las indicaciones respectivas de la condición de cáncer de los sujetos respectivos de la primera pluralidad de sujetos, eliminando de este modo una serie de funciones de extracción de características de las primeras funciones 132 de extracción de características. La poda puede comprender aplicar a las primeras funciones 132 de extracción de características una técnica de regresión, tal como, p. ej., una regresión lógica, tal como una regresión logística regularizada. En las realizaciones en las que se realiza la poda de las primeras funciones 132 de extracción de características, el número reducido de las primeras funciones 208 de extracción de características se utiliza para generar el segundo conjunto de datos transformado, como se muestra en la Figura 2. Las segundas funciones 142 de extracción de características se pueden podar de manera similar para generar un número reducido de segundas funciones 228 de extracción de características, como también se muestra en la Figura 2. De esta manera, el segundo conjunto de datos se puede aplicar al número reducido de segundas funciones 228 de extracción de características, generando de este modo los segundos valores de características que se pueden utilizar para entrenar al clasificador 160. Por lo tanto, la Figura 2 ilustra esquemáticamente que el segundo conjunto 150 de datos transformado puede generarse basándose en los valores de características obtenidos al aplicar el segundo conjunto 124 de datos al número reducido de primeras funciones 208 de extracción de características, y basándose en los segundos valores de características obtenidos al aplicar el segundo conjunto 124 de datos al número reducido de segundas funciones de extracción de características 228.

Las Figuras 10A-10H ilustran un ejemplo de un método según algunas realizaciones de la presente descripción.

Bloques de datos 1000-1020

Como se muestra en el bloque 1000, el método puede implementarse mediante el sistema informático 100 (Figura 1) para entrenar a un clasificador para que discrimine entre cada condición de cáncer de un conjunto de condiciones de cáncer. El sistema informático 100 comprende al menos un procesador 102 y la memoria 111 que almacena al menos un programa para la ejecución por parte del al menos un procesador. El al menos un programa comprende instrucciones para realizar el procesamiento mostrado en las Figuras 10A-10H y que se describe en detalle a continuación.

En el bloque 1002 de la Figura 10A, se proporciona una primera pluralidad de funciones 132 de extracción de características basadas en un primer conjunto 122 de datos. El primer conjunto 122 de datos comprende, para cada sujeto respectivo de una primera pluralidad de sujetos (p. ej., 10 o más sujetos, 20 o más sujetos, 50 o más sujetos, 100 o más sujetos) de una especie (p. ej., un ser humano), la primera información de genotipo correspondiente (126-1-1,..., 126-1-Z) que comprende (i) una primera pluralidad correspondiente de valores de bin (p. ej., valores de bin 128-1-1 para N bins para el sujeto 1-1, como se muestra en Figura 1), cada valor de bin respectivo de la primera pluralidad correspondiente de valores de bin para un bin correspondiente de una pluralidad de bins y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer (p. ej., una indicación 130-1-1 de una condición de cáncer para el sujeto 1-1, como se muestra en la Figura 1).

Cada bin de la pluralidad de bins representa una porción de un genoma de referencia de la especie, y la primera pluralidad correspondiente de valores de bin de cada sujeto respectivo de la primera pluralidad de sujetos se obtiene de una muestra biológica del sujeto respectivo, que comprende un primer tipo de tejido, utilizando una primera pluralidad correspondiente de lecturas de secuencia determinadas por un primer método de secuenciación de ácido nucleico utilizando un primer método de secuenciación. El ejemplo 10 proporciona un ejemplo de un primer método de secuenciación en el que la información de metilación se obtiene de las lecturas de secuencia y se utiliza para formar valores de bin.

La primera pluralidad de funciones de extracción de características se obtiene aplicando una primera técnica de extracción de características a los valores de bin respectivos de los sujetos respectivos del primer conjunto de datos, identificando de este modo la primera pluralidad de funciones de extracción de características, donde cada función de extracción de características de la primera pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal de los valores de bin de la totalidad o un subconjunto de la pluralidad de bins, y la primera pluralidad de funciones de extracción de características discrimina colectivamente a los sujetos respectivos de la primera pluralidad de sujetos según tengan una condición de cáncer dentro del conjunto de condiciones de cáncer basándose en los valores de bin respectivos para los sujetos respectivos. La especie puede ser humana, aunque debe apreciarse que los métodos descritos se pueden aplicar a otros tipos de especies.

Los valores de bin de la primera pluralidad correspondiente de valores de bin de un sujeto de la primera pluralidad de sujetos se pueden obtener de varias maneras, incluyendo el uso de lecturas de secuencia y/o tecnologías de micromatrices que usan una cuantificación relativa en la que la intensidad de una señal (en un punto [p. ej., un punto de ADN]) se compara con la intensidad de la señal del mismo punto en una condición diferente, y la identidad de la característica se conoce por su posición. Se puede utilizar cualquier número adecuado de lecturas de secuencia. Por

ejemplo, en algunas realizaciones, la primera pluralidad correspondiente de valores de bin de un sujeto respectivo de la primera pluralidad de sujetos se determina utilizando más de 1000, más de 3000, más de 5000, más de 10.000 o más de 20.000, más de 50.000 o más de 100.000 lecturas de secuencia que se toman colectivamente de la muestra biológica del sujeto respectivo según el primer método de secuenciación. Las lecturas de secuencia utilizadas para formar la primera pluralidad de valores de bin de un sujeto respectivo de la primera pluralidad de sujetos en todo el genoma de la especie pueden ser al menos 10 veces, al menos 20 veces o al menos 40 veces. En algunas realizaciones, cada lectura de secuencia respectiva utilizada para formar la primera pluralidad de valores de bin de un sujeto respectivo de la primera pluralidad de sujetos incluye (i) una primera porción que se puede mapear en el genoma de la especie y (ii) una segunda porción. En algunas realizaciones, las lecturas de secuencia utilizadas para formar la primera pluralidad de valores de bin de un sujeto respectivo de la primera pluralidad de sujetos se filtran de modo que solo las lecturas de secuencia cuya primera porción es inferior a 160 nucleótidos se usan para formar los valores de bin.

En algunas realizaciones, la información de genotipo se obtiene mediante la secuenciación del genoma completo o la secuenciación por panel dirigido de una muestra biológica de los sujetos. Por ejemplo, la secuenciación se puede realizar mediante secuenciación del genoma completo y la tasa de cobertura promedio de la pluralidad de lecturas de secuencia tomadas de una muestra biológica de un sujeto de entrenamiento es de al menos 1 vez, 2 veces, 3 veces, 4 veces, 5 veces, 6 veces, 7 veces, 8 veces, 9 veces, 10 veces, al menos 20 veces, al menos 30 veces, al menos 40 veces, al menos 50 veces, al menos 100 veces o al menos 200 veces en todo el genoma del sujeto de prueba. Cuando se realiza la secuenciación (basada en la metilación o no metilación) utilizando un panel objetivo, la tasa de cobertura promedio de la pluralidad de lecturas de secuencia tomadas de una muestra biológica de un sujeto de entrenamiento es al menos 200 veces, 200 veces, 500 veces, 1000 veces, al menos 2000 veces, al menos 3000 veces, o al menos 4000 veces, al menos 5000 veces, al menos 10.000 veces, al menos 20.000 veces, al menos 30.000 veces o al menos 50.000 veces en regiones seleccionadas del genoma del sujeto de prueba.

En algunas realizaciones, la muestra biológica es plasma. En algunas realizaciones, la muestra biológica comprende sangre, sangre completa, plasma, suero, orina, líquido cefalorraquídeo, materia fecal, saliva, sudor, lágrimas, líquido pleural, líquido pericárdico o líquido peritoneal del sujeto. En algunas realizaciones, la muestra biológica consiste en sangre, sangre completa, plasma, suero, orina, líquido cefalorraquídeo, materia fecal, saliva, sudor, lágrimas, líquido pleural, líquido pericárdico o líquido peritoneal del sujeto.

En algunas realizaciones, la muestra biológica se procesa para extraer ácidos nucleicos libres de células en preparación para el análisis de secuenciación. A modo de ejemplo no limitativo, en algunas realizaciones, el ácido nucleico libre de células se extrae de una muestra de sangre recolectada de un sujeto en tubos con EDTA K2. Las muestras se procesan dentro de las dos horas de la recolección mediante doble centrifugación, primero, de la sangre durante diez minutos a 1000 g y, a continuación, del plasma durante diez minutos a 2000 g. A continuación, el plasma se almacena en alícuotas de 1 ml a -80 °C. De esta manera, se prepara una cantidad adecuada de plasma (p. ej., 1-5 ml) a partir de la muestra biológica para la extracción de ácidos nucleicos sin células. En algunas de tales realizaciones, el ácido nucleico libre de células se extrae utilizando el kit QIAamp Circulating Nucleic Acid (Qiagen) y se eluye en un regulador de suspensión de ADN (Sigma). En algunas realizaciones, el ácido nucleico libre de células purificado se almacena a -20 °C hasta el uso. Ver, por ejemplo, Swanton y col., 2017, "Phylogenetic ctDNA analysis depicts early stage lung cancer evolution", *Nature*, 545 (7655): 446-451.

Se pueden utilizar otros métodos equivalentes para preparar ácido nucleico libre de células utilizando métodos biológicos con fines de secuenciación, y todos estos métodos están dentro del alcance de la presente descripción.

En algunas realizaciones, el ácido nucleico libre de células que se obtiene de la primera muestra biológica está en cualquier forma de ácido nucleico, o una combinación de las mismas. Por ejemplo, en algunas realizaciones, el ácido nucleico libre de células que se obtiene de una muestra biológica es una mezcla de ARN y ADN.

El tiempo entre la obtención de una muestra biológica y la realización de un ensayo, tal como un ensayo de secuenciación, puede optimizarse para mejorar la sensibilidad y/o la especificidad del ensayo o método. En algunas realizaciones, se puede obtener una muestra biológica inmediatamente antes de realizar un ensayo. En algunas realizaciones, se puede obtener una muestra biológica y almacenarla durante un período de tiempo (p. ej., horas, días o semanas) antes de realizar un ensayo. En algunas realizaciones, se puede realizar un ensayo en una muestra dentro de 1 día, 2 días, 3 días, 4 días, 5 días, 6 días, 1 semana, 2 semanas, 3 semanas, 4 semanas, 5 semanas, 6 semanas, 7 semanas, 8 semanas, 3 meses, 4 meses, 5 meses, 6 meses, 1 año o más de 1 año después de haber obtenido la muestra del sujeto de entrenamiento.

En algunas realizaciones, la información de genotipo para cada sujeto respectivo de la primera pluralidad de sujetos se obtiene mediante secuenciación por panel dirigida, en la que las lecturas de secuencia tomadas de una muestra biológica de un sujeto de entrenamiento para formar el constructo de genotipo 126 tienen una cobertura de al menos 50.000 veces para este panel de genes objetivo, una cobertura de al menos 55.000 veces para este panel de genes objetivo, o una cobertura de al menos 70.000 veces para este panel de genes objetivo. En algunas de tales realizaciones, el panel de genes objetivo está entre 450 y 500 genes. En algunas realizaciones, el panel de genes

objetivo está dentro del rango de 500 ± 5 genes, dentro del rango de 500 ± 10 genes o dentro del rango de 500 ± 25 genes.

En algunas realizaciones, el primer método de secuenciación es un ensayo de secuenciación del genoma completo. Un ensayo de secuenciación del genoma completo se refiere a un ensayo físico que genera lecturas de secuencia para un genoma completo o una porción sustancial del genoma completo que se puede utilizar para determinar grandes variaciones, tales como variaciones en el número de copias o aberraciones en el número de copias. Dicho ensayo físico puede emplear técnicas de secuenciación del genoma completo o técnicas de secuenciación del exoma completo.

En algunas realizaciones, el primer método de secuenciación comprende la secuenciación con bisulfito del genoma completo. En algunas de tales realizaciones, la secuenciación con bisulfito del genoma completo identifica uno o más vectores de estado de metilación tal como se describe, por ejemplo, en la solicitud de patente estadounidense n.º 16/352.602, titulada "Anomalous Fragment Detection and Classification", presentada el 13 de marzo de 2019, o según cualquiera de las técnicas descritas en la solicitud de patente estadounidense n.º 15/931.022, titulada "Model Based Featurization and Classification", presentada el 13 de mayo de 2020.

En algunas realizaciones, la información de genotipo se genera a partir de un conjunto de datos del TCGA, como se describe en el ejemplo 2 más adelante.

En algunas realizaciones, la información de genotipo se genera como se describe en los ejemplos 5 y 10 siguientes. Es decir, en algunas realizaciones, la información de genotipo está en forma de un vector 2052 de estado de metilación como se describe en el ejemplo 10 más adelante junto con la Figura 20. En algunas realizaciones, la información de genotipo está en forma de un vector 2052 de estado de metilación, como se describe en el ejemplo 10 más adelante junto con la Figura 20, siempre que el vector de estado de metilación cumpla con una o más condiciones de filtro descritas en la presente memoria, tal como el filtro de valor de p descrito, por ejemplo, en el ejemplo 5 junto con las Figuras 21 y 22.

En algunas realizaciones, las lecturas de secuencia se preprocesan para corregir sesgos o errores utilizando uno o más métodos tales como la normalización, la corrección de los sesgos de GC y/o la corrección de los sesgos debidos a la sobreamplificación por PCR.

El primer método de secuenciación puede comprender cualquier forma de secuenciación que pueda usarse para obtener un número de lecturas de secuencia medidas a partir de ácidos nucleicos libres de células, que incluyen, aunque no de forma limitativa, sistemas de secuenciación de alto rendimiento tales como la plataforma Roche 454, la plataforma SOLID de Applied Biosystems, la tecnología de secuenciación de ADN Helicos True Single Molecule, la plataforma de secuenciación por hibridación de Affymetrix Inc., la tecnología de molécula única en tiempo real (SMRT) de Pacific Biosciences, las plataformas de secuenciación por síntesis de 454 Life Sciences, Illumina/Solexa y Helicos Biosciences, y la plataforma de secuenciación por ligación de Applied Biosystems. La tecnología ION TORRENT de Life technologies y la secuenciación por nanoporos también se pueden utilizar para obtener lecturas 140 de secuencia del ácido nucleico libre de células obtenido de la muestra biológica.

En algunas realizaciones, la secuenciación por síntesis y la secuenciación basada en terminadores reversibles (p. ej., el analizador del genoma de Illumina; analizador del genoma II; HISEQ 2000; HISEQ 2500 [Illumina, San Diego California]) se utiliza para obtener lecturas de secuencia del ácido nucleico libre de células obtenido de una muestra biológica de un sujeto de entrenamiento con el fin de formar el constructo 126 de genotipo. En algunas de tales realizaciones, millones de fragmentos de ácido nucleico (p. ej., ADN) libres de células se someten a secuenciación en paralelo. En un ejemplo de este tipo de tecnología de secuenciación, se utiliza una celda de flujo que contiene un portaobjetos ópticamente transparente con ocho carriles individuales en cuyas superficies se unen anclajes de oligonucleótidos (p. ej., cebadores adaptadores). Una celda de flujo frecuentemente es un soporte sólido que se configura para retener y/o permitir el paso ordenado de las soluciones de reactivos sobre los analitos unidos. En algunas realizaciones, las celdas de flujo son de forma plana, ópticamente transparentes, generalmente a escala milimétrica o submilimétrica, y frecuentemente tienen canales o carriles en los que se produce la interacción analito/reactivo. En algunas realizaciones, una muestra de ácido nucleico libre de células puede incluir una señal o etiqueta que facilite la detección. En algunas de tales realizaciones, la adquisición de lecturas de secuencia del ácido nucleico libre de células obtenido de la muestra biológica incluye obtener información de cuantificación de la señal o etiqueta mediante una variedad de técnicas tales como, por ejemplo, citometría de flujo, reacción en cadena de la polimerasa cuantitativa (qPCR, por sus siglas en inglés), electroforesis en gel, análisis de chips genéticos, micromatrices, espectrometría de masas, análisis citofluorimétrico, microscopía por fluorescencia, microscopía de barrido láser confocal, citometría de barrido láser, cromatografía de afinidad, separación manual por lotes, suspensión de campo eléctrico, secuenciación y combinación de los mismos.

En algunas realizaciones, el primer conjunto de datos comprende, para cada sujeto respectivo de una primera pluralidad de sujetos de una especie, la primera información de genotipo correspondiente que comprende una primera pluralidad correspondiente de valores de bin, cada valor de bin respectivo de la primera pluralidad correspondiente de valores de bin para una pluralidad de bins. En tales realizaciones, cada bin respectivo de la pluralidad de bins

representa una porción diferente y no superpuesta del genoma de un genoma de referencia para la especie. Los bins pueden tener el mismo tamaño o tamaños diferentes. Además, En tales realizaciones, la información de genotipo para cada sujeto respectivo de la primera pluralidad de sujetos comprende un primer recuento de bins para cada bin respectivo de la primera pluralidad de bins.

En algunas realizaciones, las regiones genómicas con alta variabilidad o baja capacidad de mapeo se excluyen de la representación de bins de la pluralidad de bins, por ejemplo, utilizando los métodos descritos en Jensen y col., 2013, PLoS One 8; e57381. Ver, también, Li y Freudenber, 2014, Front. Genet. 5, pág. 318, para el análisis de la mapeabilidad.

En algunas realizaciones, los recuentos de bins se determinan utilizando cualquiera de las técnicas descritas en la solicitud de patente estadounidense n.º 16/201.912 titulada "Models for Targeted Sequencing", presentada el 27 de noviembre de 2018, o en la solicitud de patente estadounidense n.º 16/352.214 titulada "Identifying Copy Number Aberrations", presentada el 13 de marzo de 2019.

En algunas realizaciones, el primer recuento de bins representativo de la primera información de genotipo es un número fragmentos representados por lecturas de secuencia en la información de secuenciación medida a partir del ácido nucleico libre de células en la muestra biológica que se mapea en la porción diferente del genoma de la especie representada por el bin respectivo. En algunas realizaciones, dicho mapeo solo permite coincidencias perfectas. En algunas realizaciones, dicho mapeo permite algunos desajustes. En algunas realizaciones, se utiliza un programa como Bowtie 2 para realizar dicho mapeo. Ver, por ejemplo, Langmead y Salzberg, 2012, Nat Methods 9, págs. 357-359, por ejemplo, la descripción de dicho mapeo.

En algunas realizaciones, el primer recuento de bins se determina mediante un número de fragmentos de ácido nucleico únicos en el ácido nucleico libre de células de la muestra biológica que se mapean en la porción diferente del genoma de la especie representada por el bin respectivo. Dependiendo del método de secuenciación usado, cada uno de dichos fragmentos de ácido nucleico únicos puede representarse mediante un número de lecturas de secuencia. En casos típicos, esta redundancia en las lecturas de secuencia para fragmentos de ácido nucleico únicos en el ácido nucleico libre de células se resuelve utilizando técnicas de secuenciación múltiple, tales como códigos de barras, de modo que un recuento de bins para un bin respectivo represente el número de fragmentos de ácido nucleico únicos en el ácido nucleico libre de células en la muestra biológica que se mapean en la porción diferente del genoma de la especie representada por el bin respectivo, en lugar del número total de lecturas de secuencia en la pluralidad de lecturas de secuencia que se mapean en el bin respectivo. Ver Kircher y col., 2012, Nucleic Acids Research 40, n.º 1 e3, por ejemplo, la descripción sobre códigos de barras. En algunas de tales realizaciones, el primer recuento de bins se realiza en forma de puntuaciones B, que se describen, por ejemplo, en la solicitud de patente estadounidense n.º 16/352.739, titulada "Method and System for Selecting, Managing, and Analyzing Data of High Dimensionality", presentada el 13 de marzo de 2019.

En algunas realizaciones, el primer recuento de bins es un número de fragmentos cuyos patrones de metilación cumplen con una o más condiciones de filtro descritas en la presente memoria, tal como el filtro de valor de p del ejemplo 5, tras la generación de vectores de estado de metilación para dichos fragmentos tal como se describe, por ejemplo, en el ejemplo 10, y que se mapean en los bins respectivos de la pluralidad de bins.

En algunas realizaciones, los datos de secuenciación se procesan previamente para corregir sesgos o errores utilizando uno o más métodos tales como la normalización, la corrección de los sesgos de GC, la corrección de los sesgos debidos a la sobreamplificación por PCR, etc. Por ejemplo, en algunas realizaciones, para un sujeto respectivo de la primera pluralidad de sujetos, se obtiene una mediana del valor de bin de la primera pluralidad correspondiente de valores de bin para el sujeto respectivo. A continuación, cada valor de bin respectivo de la primera pluralidad de valores de bin para el sujeto respectivo se divide por esta mediana de valor, garantizando así que los valores de bin para el sujeto respectivo estén centrados en un valor conocido (p. ej., en cero):

$$bv_i^* = \frac{bv_i}{Mediana(bv_j)}$$

donde,

bv_i = el valor de bin del bin i de la primera pluralidad de valores de bin para el sujeto respectivo, bv_i^* = el valor de bin normalizado del bin i de la primera pluralidad de valores de bin para el sujeto respectivo tras esta primera normalización, y

$mediana(bv_j)$ = mediana del valor de bin de la primera pluralidad de valores de bin no normalizados para el sujeto respectivo.

En algunas realizaciones, en lugar de utilizar la mediana del valor de bin de la primera pluralidad correspondiente de valores de bin, se utiliza alguna otra medida de tendencia central, tal como una media aritmética, media ponderada, rango medio, rango cuartil medio, trimedia, media winsorizada, media o moda de la pluralidad de valores de bin del sujeto respectivo.

En algunas realizaciones, cada recuento de bins normalizado respectivo bv_i^* se normaliza aún más mediante la mediana del valor normalizado para el bin respectivo de la primera pluralidad de sujetos k :

$$bv_i^{**} = \log \left(\frac{bv_i^*}{\text{Mediana}(bv_{ik}^{**})} \right)$$

donde,

bv_i^* = el valor de bin normalizado del bin i de la primera pluralidad de valores de bin para el sujeto respectivo del primer procedimiento de normalización descrito anteriormente,

bv_i^{**} = el valor de bin normalizado del bin i para el sujeto respectivo tras esta segunda normalización descrita aquí, y

$\text{mediana}(bv_{ik}^{**})$ = la mediana del valor de bin normalizado bv_i^* para el bin i de la primera pluralidad de sujetos (k sujetos).

En algunas realizaciones, los valores de bin no normalizados bv_i están normalizados por GC. En algunas realizaciones, los valores de bin normalizados bv_i^* están normalizados por GC. En algunas realizaciones, los valores de bin normalizados bv_i^{**} están normalizados por GC. En tales realizaciones, los recuentos de GC de las lecturas de secuencia respectivas de la primera pluralidad de lecturas de secuencia de cada sujeto en la pluralidad de sujetos se agrupan en bins. Una curva que describe el recuento medio condicional de fragmentos por valor de GC se estima mediante dicha agrupación de bins (Yoon y col., 2009, Genome Research 19(9):1586) o, alternativamente, asumiendo la suavidad (Boeva y col., 2011, Bioinformatics 27 (2), pág. 268; Miller y col., 2011, PLoS ONE 6(1), pág. e16327). La curva GC resultante determina un recuento predicho para cada bin basándose en el GC del bin. Estas predicciones se pueden utilizar directamente para normalizar la señal original (p. ej., bv_i^* , bv_i o bv_i^{**}). Como ejemplo no limitativo, en el caso de agrupación de bins y normalización directa, para cada porcentaje de G+C respectivo en el conjunto {0 %, 1 %, 2 %, 3 %, ..., 100 %}, el valor mcc, la mediana de valor de bv_i^{**} de todos los bins de la primera pluralidad de sujetos que tienen este porcentaje de G+C respectivo, se determina y se resta de los valores de bin normalizados bv_i^{**} de aquellos bins que tienen el porcentaje de G+C respectivo para formar valores de bin normalizados de GC

bv_i^{***} . En la Figura 17, la curva 1702 es un gráfico del porcentaje de G+C frente a el valor bv_i^{**} de bin en la primera pluralidad de bins en la pluralidad de sujetos. Tras la normalización del GC, los valores de bin normalizados de GC bv_i^{***} , tal como se expone en el gráfico 1704 de la Figura 17, se centran ahora en el contenido de GC, eliminando de este modo el sesgo de GC de los valores de bin. En algunas realizaciones, en lugar de utilizar la mediana de valor de bv_i^{**} de todos los bins en la primera pluralidad de sujetos que tienen este porcentaje de G+C respectivo, se utiliza

alguna otra forma de medida de la tendencia central de bv_i^{**} de todos los bins en la primera pluralidad de sujetos que tienen este porcentaje de G+C respectivo, tal como una media aritmética, media ponderada, rango medio, rango cuartil medio, trimedia, media winsorizada, media o moda. En algunas realizaciones, la curva 1702 se determina utilizando un modelo de suavizado de gráficos de dispersión ponderado localmente (p. ej., LOESS, LOWESS, etc.). Ver, por ejemplo, Benjamini y Speed, 2012, Nucleic Acids Research 40(10): e72; y Alkan y col., 2009, Nat Genet 41:1061-7. Por ejemplo, en algunas realizaciones, la curva de sesgo de GC se determina mediante la regresión LOESS del recuento mediante GC (p. ej., utilizando el paquete R “loess”) en un muestreo aleatorio (o muestreo exhaustivo) de bins de la pluralidad de sujetos. En algunas realizaciones, la curva de sesgo de GC se determina mediante la regresión LOESS del recuento mediante GC (p. ej., utilizando el paquete R “loess”), o alguna otra forma de ajuste de curvas, en un muestreo aleatorio de bins de una cohorte de sujetos jóvenes sanos que se sometieron a secuenciación utilizando las mismas técnicas de secuenciación utilizadas para la secuenciación de la primera pluralidad de sujetos.

En algunas realizaciones, los recuentos de bins se normalizan mediante el análisis de componentes principales (ACP) para eliminar los artefactos de orden superior y realizar una corrección basada en la población (personas sanas). Ver, por ejemplo, Price y col., 2006, Nat Genet 38, págs. 904-909; Leek y Storey, 2007, PLoS Genet 3, págs. 1724 a 1735; y Zhao y col., 2015, Clinical Chemistry 61(4), págs. 608-616. Dicha normalización puede ser adicional a o en lugar de

cualquiera de las técnicas de normalización identificadas anteriormente. En algunas de tales realizaciones, para entrenar la normalización por ACP, se utiliza una matriz de datos que comprende recuentos bv_i^{***} de bins normalizados con LOESS de sujetos jóvenes sanos de la primera pluralidad de sujetos (u otra cohorte que se sometió a secuenciación de la misma manera que la primera pluralidad de sujetos) y la matriz de datos se transforma en un espacio de componentes principales, obteniendo de este modo el número de N primeros componentes principales en todo el conjunto de entrenamiento. En algunas realizaciones, los 2 primeros, los 3 primeros, los 4 primeros, los 5 primeros, los 6 primeros, los 7 primeros, los 8 mejores, los 9 primeros o los 10 primeros componentes principales se utilizan para construir un modelo de regresión lineal:

$$LM(PC_1, \dots, PC_N)$$

A continuación, cada bin bv_i^{***} de cada bin respectivo de cada sujeto respectivo de la primera pluralidad de sujetos se ajusta a este modelo lineal para formar un recuento bv_i^{****} de bin correspondiente normalizado por ACP.

$$bv_i^{****} = bv_i^{***} - \text{compatible con}_{LM(PC_1, \dots, PC_N)}.$$

En otras palabras, para cada sujeto respectivo de la pluralidad de sujetos, se ajusta un modelo de regresión lineal entre sus recuentos $\{bv_1^{***}, \dots, bv_i^{***}\}$ de bin normalizados y los primeros componentes principales del conjunto de entrenamiento. Los residuos de este modelo sirven como valores $\{bv_1^{****}, \dots, bv_i^{****}\}$ de bin normalizados finales para el sujeto respectivo. Intuitivamente, los primeros componentes principales representan el ruido que se observa comúnmente en muestras sanas y, por lo tanto, eliminar dicho ruido (en la forma de los primeros componentes principales derivados de la cohorte sana) de los valores bv_i^{***} de bin puede mejorar eficazmente la normalización. Ver Zhao y col., 2015, Clinical Chemistry 61(4), págs. 608-616, para obtener más información sobre la normalización por ACP de lecturas de secuencia utilizando una población sana. Con respecto a la normalización anterior, se apreciará que todas las variables se estandarizan (p. ej., restando sus medias y dividiendo por sus desviaciones estándar) cuando sea necesario.

En algunas realizaciones de la presente descripción, el genoma humano está bajo examen. Por ejemplo, en algunas realizaciones, el genoma humano se divide en aproximadamente 30 mil bins. A continuación, algunos de los bins se retiran del examen utilizando los métodos descritos en la solicitud de patente estadounidense n.º 16/352.739, titulada "Method and System for Selecting, Managing, and Analyzing Data of High Dimensionality", presentada el 13 de marzo de 2019, hasta llegar a 23.000 bins. En tales realizaciones, cada bin tiene aproximadamente el mismo tamaño, en términos de la cantidad de un genoma de referencia humano que corresponde al bin.

Se apreciará que, en los casos donde un valor de bin es un recuento de fragmentos que se mapean en un bin, el término "valor de bin" se refiere a cualquier forma de representación de este número de fragmentos nucleicos que se mapean en el bin i dado, y que dicho valor de bin puede estar en una forma no normalizada (p. ej., bv_i) o normalizada (p. ej., bv_i^* , bv_i^{**} , bv_i^{***} , bv_i^{****} , etc.)

En algunas realizaciones, la pluralidad de bins se construye dividiendo todo o una porción de un genoma de referencia (p. ej., mamífero, humano, etc.) en bins de igual tamaño, donde cada bin representa una parte única de igual tamaño del genoma de referencia. En algunas realizaciones, la pluralidad de bins se construye dividiendo todo un genoma de referencia (p. ej., mamífero, humano, etc.) en bins de igual tamaño, donde cada bin representa una parte única de igual tamaño del genoma de referencia. En algunas realizaciones, la pluralidad de bins se construye dividiendo una porción de un genoma de referencia (p. ej., mamífero, humano, etc.) en bins de igual tamaño, donde cada bin representa una parte única de igual tamaño del genoma de referencia y la porción del genoma de referencia está entre 1 y 22 cromosomas del genoma de referencia, o al menos el 25 por ciento, al menos el 30 por ciento, al menos el 35 por ciento, al menos el 40 por ciento, al menos el 45 por ciento, al menos el 50 por ciento, al menos el 55 por ciento, al menos el 60 por ciento, al menos el 65 por ciento, al menos el 70 por ciento, al menos el 75 por ciento, al menos el 80 por ciento, al menos el 85 por ciento, al menos el 90 por ciento, al menos el 95 por ciento o al menos el 99 por ciento del genoma de referencia. En algunas de tales realizaciones, cada bin representa entre 10.000 bases y 100.000 bases, entre 20.000 bases y 300.000 bases, entre 30.000 bases y 500.000 bases, entre 40.000 bases y 1.000.000 de bases, entre 50.000 bases y 5.000.000 de bases, o entre 100.000 bases y 25.000.000 de bases del genoma de referencia.

En algunas realizaciones, la pluralidad de bins se construye dividiendo todo o una porción de un genoma de referencia (p. ej., mamífero, humano, etc.) en bins de tamaño igual o desigual, donde cada bin representa una parte única del genoma de referencia. En algunas realizaciones, la pluralidad de bins se construye de modo que al menos algunas de las regiones del genoma humano implicadas, en ausencia o presencia de cáncer (p. ej., extraídas de las regiones

identificadas en los ejemplos 4, 7, 8 y/o 9), están representadas por la pluralidad de bins, mientras que otras regiones del genoma de referencia no están representadas por los bins. En algunas realizaciones, la pluralidad de bins se construye dividiendo todo un genoma de referencia (p. ej., mamífero, humano, etc.) en bins de tamaño igual o desigual, donde cada bin representa una parte única del genoma de referencia. En algunas realizaciones, la pluralidad de bins se construye dividiendo una porción de un genoma de referencia (p. ej., mamífero, humano, etc.) en bins de tamaño igual o desigual, donde cada bin representa una parte única del genoma de referencia y la porción del genoma de referencia está entre 1 y 22 cromosomas del genoma de referencia, o al menos el 25 por ciento, al menos el 30 por ciento, al menos el 35 por ciento, al menos el 40 por ciento, al menos el 45 por ciento, al menos el 50 por ciento, al menos el 55 por ciento, al menos el 60 por ciento, al menos el 65 por ciento, al menos el 70 por ciento, al menos el 75 por ciento, al menos el 80 por ciento, al menos el 85 por ciento, al menos el 90 por ciento, al menos el 95 por ciento o al menos el 99 por ciento del genoma de referencia. Independientemente del enfoque, cada bin representa una parte única del genoma de referencia. En algunas realizaciones, particularmente cuando los valores de bin para dichos bins representan características epigenéticas de los datos de metilación obtenidos de la secuenciación dirigida para el primer o el segundo conjunto de datos implicados en el aprendizaje por transferencia, dichos bins varían en tamaño entre 30 pb y 5000 pb, entre 30 pb y 4000 pb, entre 30 pb y 3000 pb, entre 30 pb y 2000 pb, entre 30 pb y 1000 pb, o entre 40 pb y 800 pb del genoma de referencia. En realizaciones alternativas, dichos bins varían en tamaño entre 10.000 pb y 100.000 pb, entre 20.000 pb y 300.000 pb, entre 30.000 pb y 500.000 pb, entre 40.000 pb y 1.000.000 de pb, entre 50.000 pb y 5.000.000 de pb, o entre 100.000 pb y 25.000.000 de pb del genoma de referencia.

En algunas realizaciones, la pluralidad de bins se construye dividiendo todo o una porción de un genoma de referencia (p. ej., mamífero, humano, etc.) en bins de tamaño igual o desigual, donde cada bin representa una parte correspondiente del genoma de referencia. En tales realizaciones, la parte correspondiente del genoma de referencia representada por un bin de la pluralidad de bins puede superponerse a la parte correspondiente del genoma de referencia representada por otro bin de la pluralidad de bins. En algunas de tales realizaciones, la pluralidad de bins se construye dividiendo todo un genoma de referencia (p. ej., mamífero, humano, etc.) en bins de tamaño igual o desigual, donde cada bin representa una parte correspondiente superpuesta o no superpuesta del genoma de referencia. En algunas realizaciones, la pluralidad de bins se construye dividiendo una porción de un genoma de referencia (p. ej., mamífero, humano, etc.) en bins de tamaño igual o desigual, donde cada bin representa una parte superpuesta o no superpuesta del genoma de referencia y la porción del genoma de referencia está entre 1 y 22 cromosomas del genoma de referencia, o al menos el 25 por ciento, al menos el 30 por ciento, al menos el 35 por ciento, al menos el 40 por ciento, al menos el 45 por ciento, al menos el 50 por ciento, al menos el 55 por ciento, al menos el 60 por ciento, al menos el 65 por ciento, al menos el 70 por ciento, al menos el 75 por ciento, al menos el 80 por ciento, al menos el 85 por ciento, al menos el 90 por ciento, al menos el 95 por ciento o al menos el 99 por ciento del genoma de referencia. En algunas de tales realizaciones, cada bin representa entre 10.000 bases y 100.000 bases, entre 20.000 bases y 300.000 bases, entre 30.000 bases y 500.000 bases, entre 40.000 bases y 1.000.000 de bases, entre 50.000 bases y 5.000.000 de bases, o entre 100.000 bases y 25.000.000 de bases del genoma de referencia.

En algunas realizaciones, cada uno de los bins representa un sitio específico de un genoma de referencia que se ha identificado como asociado con el cáncer.

En algunas realizaciones, cada uno de los bins representa una región específica de un genoma de referencia que se ha identificado como asociada con el cáncer a través de patrones de metilación específicos de cáncer y/o de tejido en el ADNlc en relación con los controles sin cáncer. Por ejemplo, el ejemplo 4 describe 103.456 de dichas regiones distintas. Los ejemplos 7, 8 y 9 también describen varias regiones distintas. En algunas realizaciones, existe una correspondencia uno a uno entre dichos bins y estas regiones. En otras palabras, En tales realizaciones, cada bin abarca una sola región única entre aquellas identificadas en los ejemplos 4, 7, 8 y/o 9. En algunas de tales realizaciones, cada bin varía en tamaño entre 30 pb y 5000 pb, entre 30 pb y 4000 pb, entre 30 pb y 3000 pb, entre 30 pb y 2000 pb, entre 30 pb y 1000 pb, o entre 30 pb y 750 pb. En algunas realizaciones, en el caso donde las regiones utilizadas se extraigan de los ejemplos 4, 7, 8 y/o 9, cada bin incluye entre 1 y 590 dinucleótidos de citosina-guanina (CpG). En algunas realizaciones, algunos de los bins representan regiones que están hipometiladas en el estado de cáncer en relación con el estado normal libre de cáncer. En algunas realizaciones, algunos de los bins representan regiones que están hipermetiladas en el estado de cáncer en relación con el estado normal libre de cáncer. En algunas realizaciones, la pluralidad de bins utilizados colectivamente abarca al menos 1000, al menos 2000, al menos 3000, al menos 4000, al menos 5000, al menos 6000, al menos 7000, al menos 8000, al menos 9000, al menos 10.000, al menos 25.000, al menos 30.000, al menos 40.000 o al menos 50.000 de las regiones identificadas en los ejemplos 4, 7, 8 y/o 9, representando cada bin de la pluralidad de bins una región única diferente de la pluralidad de regiones identificadas en los ejemplos 4, 7, 8 y/o 9. En tales realizaciones, el valor de bin para cada bin se basa en un número de fragmentos de ácido nucleico, tal como se determina a partir de la primera pluralidad correspondiente de lecturas de secuencia adquiridas de una muestra biológica de un sujeto respectivo que se mapean en el bin respectivo.

En algunas realizaciones, la pluralidad de bins se obtiene de las secuencias descritas en los ejemplos 4, 7, 8 y/o 9. En algunas de tales realizaciones, los objetivos adyacente y superpuesto (secuencia genómica dirigida por una sonda a una región de los ejemplos 4, 7, 8 y/o 9) se fusionan en regiones genómicas contiguas. En algunas realizaciones, si cada una de las regiones resultantes es menor que un número umbral de pares de bases (p. ej., 1000 pares de bases), se utiliza tal cual como un bin correspondiente en la pluralidad de bins; de lo contrario, se subdivide en subregiones (p. ej., 1000 regiones de pares de bases). Se apreciará que la presente descripción no se limita a los bins que tienen

regiones de 1000 pares de bases y que se puede utilizar cualquier valor entero positivo entre 100 pares de bases y 10 millones de pares de bases para definir los bins. Además, se apreciará que, en lugar de dividir un genoma por valores de pares de bases que forman bins, el genoma puede dividirse en bins basados en bloques de sitios CpG, tales como entre 1 y 1000 sitios CpG por bin, en lugar de considerar explícitamente las longitudes de los pares de bases para dichos bins. En algunas realizaciones, los bins se disponen de modo que los bins consecutivos se superpongan en un cierto número de pares de bases (p. ej., en el caso de bins de 1000 pares de bases, se superpongan, por ejemplo, en 500 pares de bases), que pueden o no representar un cierto número de sitios CpG. En algunas realizaciones, el tamaño de cada bin varía entre 30 pb y 5000 pb, entre 30 pb y 4000 pb, entre 30 pb y 3000 pb, entre 30 pb y 2000 pb, entre 30 pb y 1000 pb, o entre 30 pb y 750 pb.

En algunas realizaciones, la pluralidad de bins se obtiene de modo que cada bin abarca una, dos, tres, cuatro, cinco, seis, siete u ocho sondas descritas en el ejemplo 6. En algunas de tales realizaciones, los objetivos adyacente y superpuesto (secuencia genómica dirigida por una sonda del ejemplo 6) se fusionan en regiones genómicas contiguas. En algunas realizaciones, si cada una de las regiones resultantes es menor que un número umbral de pares de bases (p. ej., 1000 pares de bases), se utiliza tal cual como un bin correspondiente en la pluralidad de bins; de lo contrario, se subdivide en subregiones (p. ej., regiones de 1000 pares de bases). Se apreciará que la presente descripción no se limita a los bins que tienen regiones de 1000 pares de bases y que se puede utilizar cualquier valor entero positivo entre 100 pares de bases y 10 millones de pares de bases para definir los bins. En algunas realizaciones, los bins se disponen de modo que los bins consecutivos se superpongan en un cierto número de pares de bases (p. ej., en el caso de bins de 1000 pares de bases, se superpongan, por ejemplo, en 500 pares de bases). En algunas de tales realizaciones, cada bin varía en tamaño entre 30 pb y 5000 pb, entre 30 pb y 4000 pb, entre 30 pb y 3000 pb, entre 30 pb y 2000 pb, entre 30 pb y 1000 pb, o entre 30 pb y 750 pb.

En algunas realizaciones, la pluralidad de bins se obtiene de modo que cada bin abarque una región del genoma descrito en el ejemplo 4. En algunas de tales realizaciones, cada bin varía en tamaño entre 30 pb y 5000 pb, entre 30 pb y 4000 pb, entre 30 pb y 3000 pb, entre 30 pb y 2000 pb, entre 30 pb y 1000 pb, o entre 30 pb y 750 pb.

En algunas realizaciones, la pluralidad de bins se obtiene de modo que cada bin abarque una región del genoma descrito en el ejemplo 7. En algunas de tales realizaciones, cada bin varía en tamaño entre 30 pb y 5000 pb, entre 30 pb y 4000 pb, entre 30 pb y 3000 pb, entre 30 pb y 2000 pb, entre 30 pb y 1000 pb, o entre 30 pb y 750 pb.

En algunas realizaciones, la pluralidad de bins se obtiene de modo que cada bin abarque una región del genoma descrito en el ejemplo 8. En algunas de tales realizaciones, cada bin varía en tamaño entre 30 pb y 5000 pb, entre 30 pb y 4000 pb, entre 30 pb y 3000 pb, entre 30 pb y 2000 pb, entre 30 pb y 1000 pb, o entre 30 pb y 750 pb.

En algunas realizaciones, la pluralidad de bins se obtiene de modo que cada bin abarque una región del genoma descrito en el ejemplo 9. En algunas de tales realizaciones, cada bin varía en tamaño entre 30 pb y 5000 pb, entre 30 pb y 4000 pb, entre 30 pb y 3000 pb, entre 30 pb y 2000 pb, entre 30 pb y 1000 pb, o entre 30 pb y 750 pb.

En algunas realizaciones, la pluralidad de bins se obtiene de cualquier combinación de los bins descritos en los ejemplos 4, 7, 8 o 9. En algunas de tales realizaciones, cada bin varía en tamaño entre 30 pb y 5000 pb, entre 30 pb y 4000 pb, entre 30 pb y 3000 pb, entre 30 pb y 2000 pb, entre 30 pb y 1000 pb, o entre 30 pb y 750 pb.

En algunas realizaciones, cada bin representa la totalidad o una porción de un potenciador, promotor, UTR 5', exón, límite exón/inhibidor, intrón, límite intrón/exón, región UTR 3', plataforma CpG, orilla CpG o isla CpG en un genoma de referencia. Ver, por ejemplo, Cavalcante y Santor, 2017, "annotatr: genomic regions in context", *Bioinformatics* 33(15) 2381-2383, para obtener definiciones adecuadas de dichas regiones y dónde dichas anotaciones están documentadas para varias especies diferentes.

En algunas realizaciones, un genoma de referencia (o un subconjunto del genoma de referencia) se divide en una o más etapas, p. ej., para casos de uso que implican un ensayo de metilación dirigido. Por ejemplo, el genoma de referencia se separa en bloques (bins) de sitios CpG. Como se utiliza en la presente memoria, en este contexto, los términos "bins" y "bloques" se usan indistintamente. En algunas de tales realizaciones, cada bin (bloque) se define cuando hay una separación entre dos sitios CpG adyacentes que supera un umbral, p. ej., más de 200 pares de bases (pb), 300 pb, 400 pb, 500 pb, 600

pb, 700 pb, 800 pb, 900 pb o 1000 pb, entre otros valores. Por lo tanto, los bins (bloques) En tales realizaciones pueden variar en función del tamaño de los pares de bases. Para cada bin (bloque) respectivo, el bin respectivo se divide en ventanas de una cierta longitud, p. ej., 500 pb, 600 pb, 700 pb, 800 pb, 900 pb, 1000 pb, 1100 pb, 1200 pb, 1300 pb, 1400 pb o 1500 pb, entre otros valores. En otras realizaciones, las ventanas pueden tener una longitud de 200 bp a 10 kilopares de bases (kpb), de 500 bp a 2 kpb, o aproximadamente 1 kpb. Las ventanas (p. ej., que son adyacentes) pueden superponerse en varios pares de bases o un porcentaje de la longitud, p. ej., el 10 %, el 20 %, el 30 %, el 40 %, el 50 % o el 60 %, entre otros valores.

En algunas realizaciones, a continuación se analizan las lecturas de secuencia derivadas de fragmentos de ADN utilizando un proceso de creación de ventanas. En particular, un procesador de secuencias escanea los bins ventana

por ventana y lee fragmentos dentro de cada ventana. Dichas ventanas de bins se ilustran en la Figura 23. En algunas realizaciones, los fragmentos se originan en tejido y/o ADNlc de alta señal. Las muestras de ADNlc de alta señal se pueden determinar mediante un modelo de clasificación binaria, etapa del cáncer u otro indicador. Al dividir el genoma de referencia (p. ej., utilizando bins y ventanas), se facilita la paralelización computacional. Además, se reducen los recursos computacionales necesarios para procesar un genoma de referencia dirigiéndose a las secciones de pares de bases que incluyen sitios CpG, mientras se omiten otras secciones que no incluyen sitios CpG. Ver, por ejemplo, la solicitud de patente estadounidense n.º 15/931.022, titulada "Model Based Featurization and Classification", presentada el 13 de mayo de 2020.

En algunas realizaciones, cada valor de bin respectivo de la primera pluralidad correspondiente de valores de bin para un bin correspondiente de una pluralidad de bins para un sujeto respectivo se determina identificando el número de fragmentos de ácido nucleico representados en una primera pluralidad correspondiente de lecturas de secuencia obtenidas de una muestra biológica del sujeto respectivo, que se mapean en la región genómica representada por el bin correspondiente.

En algunas realizaciones, cada valor de bin respectivo de la primera pluralidad correspondiente de valores de bin para un bin correspondiente de una pluralidad de bins para un sujeto respectivo es el número de fragmentos de ácido nucleico representados en una primera pluralidad correspondiente de lecturas de secuencia obtenidas de una muestra biológica del sujeto respectivo, que se mapean en la región genómica representada por el bin correspondiente.

En algunas realizaciones, cada valor de bin respectivo de la primera pluralidad correspondiente de valores de bin para un bin correspondiente de una pluralidad de bins para un sujeto respectivo es el número de fragmentos de ácido nucleico representados en una primera pluralidad correspondiente de lecturas de secuencia, obtenidas de una muestra biológica del sujeto respectivo, que (i) se mapean en la región genómica representada por el bin correspondiente y (ii) cumplen con una o más condiciones de filtro. A continuación se proporcionan ejemplos no limitativos de condiciones de filtro.

Filtrado del valor de p basándose en vectores de metilación. En algunas realizaciones, una condición de filtro de la una o más condiciones de filtro es el requisito de que cada fragmento tenga un valor de p que esté por debajo de un valor umbral, donde el valor de p se determina mediante el filtrado del valor de p como se describe en el ejemplo 5. El objetivo de dicha condición de filtro es aceptar y utilizar fragmentos anómalamente metilados basándose en sus vectores de estado de metilación. Por ejemplo, para cada fragmento de una muestra, se realiza una determinación de si el fragmento está anómalamente metilado (mediante el análisis de las lecturas de secuencia derivadas del mismo) en relación con un vector de estado de metilación esperado de una muestra sana, utilizando el vector de estado de metilación correspondiente al fragmento. En algunas realizaciones, se calcula el valor de p para cada vector de estado de metilación que describe una probabilidad de observar ese vector de estado de metilación u otros vectores de estado de metilación incluso menos probables en el grupo de control sano (como se describe, por ejemplo, en la publicación de solicitud de patente estadounidense n.º 201 9/0287652). El proceso para calcular un valor de p se

describe en el ejemplo 5. En algunas realizaciones, el valor umbral es 0,01 (p. ej., p debe ser $<0,01$ En tales realizaciones, como se determina mediante los métodos descritos en el ejemplo 5, En tales realizaciones). En algunas realizaciones, el umbral es 0,001, 0,005, 0,01, 0,015, 0,02, 0,05 o 0,10. En algunas realizaciones, el umbral está entre 0,0001 y 0,20. En tales realizaciones, solo aquellos fragmentos que tienen un valor de p por debajo del valor umbral contribuyen al recuento de bins. En algunas realizaciones, se utiliza un modelo de Markov (p. ej., un modelo oculto de Markov [HMM, por sus siglas en inglés]) para determinar la probabilidad de que se observe una secuencia de estados de metilación (que comprende, p. ej., "M" o "U") para cada fragmento respectivo representado por la pluralidad de lecturas de secuencia, dado un conjunto de probabilidades que determinan, para cada estado de la secuencia de ácido nucleico del fragmento respectivo, la probabilidad de observar el siguiente estado de la secuencia. En algunas realizaciones, el conjunto de probabilidades se obtiene entrenando el HMM. Dicho entrenamiento implica calcular parámetros estadísticos (p. ej., la probabilidad de que un primer estado pase a un segundo estado [la probabilidad de transición] y/o la probabilidad de que se observe un estado de metilación dado en un sitio de CpG respectivo [la probabilidad de emisión]), dado un conjunto de datos de entrenamiento inicial de secuencias de estado de metilación observadas (p. ej., patrones de metilación) obtenidas de una cohorte de sujetos sin cáncer. En algunas realizaciones, el HMM se entrena mediante un entrenamiento supervisado (p. ej., utilizando muestras donde se conocen la secuencia subyacente y los estados observados). En algunas realizaciones alternativas, el HMM se entrena mediante un entrenamiento no supervisado (p. ej., aprendizaje de Viterbi, estimación de máxima verosimilitud, entrenamiento de maximización de expectativas y/o entrenamiento de Baum-Welch). Por ejemplo, un algoritmo de maximización de expectativas, tal como el algoritmo Baum-Welch, estima las probabilidades de transición y emisión de las secuencias de muestra observadas y genera un modelo probabilístico parametrizado que explica mejor las secuencias observadas. Dichos algoritmos repiten el cálculo de una función de verosimilitud hasta que se maximice el número esperado de estados predichos correctamente. Ver, p. ej., Yoon, 2009, "Hidden Markov Models and their Applications in Biological Sequence Analysis", Curr. Genomics. Sep; 10(6): 402-415, doi: 10.2174/138920209789177575. Ver también el Ejemplo 5 para un análisis adicional de los modelos de Markov.

La Figura 18 ilustra el filtrado de fragmentos representados por una pluralidad de lecturas de secuencia obtenidas de un sujeto eliminando cada fragmento representado que no cumple con un umbral de valor p, según algunas

realizaciones de la presente descripción. El filtro se aplica al patrón de metilación de cada fragmento respectivo representado por la primera pluralidad respectiva de lecturas de secuencia de un sujeto correspondiente, utilizando los patrones de metilación observados en una población de entrenamiento de sujetos sin cáncer, tal como se describe en el ejemplo 5. Cada patrón de metilación respectivo de cada fragmento respectivo (p. ej., fragmento uno,..., fragmento N) comprende uno o más sitios (p. ej., sitios CpG) de metilación correspondientes identificados con un identificador de sitio de metilación y un patrón de metilación correspondiente, representados como una secuencia de 1 y 0, donde cada "1" representa un sitio CpG metilado del uno o más sitios CpG y cada "0" representa un sitio CpG no metilado del uno o más sitios CpG. Los patrones de metilación observados en los fragmentos de la población de entrenamiento sin cáncer, tal como se describe en el ejemplo 5, se utilizan para construir una distribución de estado de metilación para los estados de sitio CpG representados colectivamente por los fragmentos de la población de entrenamiento sin cáncer (p. ej., el sitio CpG A, el sitio CpG B,..., sitio CpG ZZZ).

Como se representa en la Figura 18, las probabilidades de transición entre estados para un modelo oculto de Markov (HMM, por sus siglas en inglés) se representan mediante las flechas de la ilustración y se pueden determinar utilizando la distribución de estado de metilación para los estados de sitio CpG (p. ej., utilizando un algoritmo de maximización de expectativas tal como el algoritmo de Baum-Welch), entrenando de este modo el HMM. Para cada fragmento respectivo representado por la pluralidad respectiva de lecturas de secuencia de un sujeto, el HMM entrenado se utiliza para determinar la probabilidad de incidencia del patrón de metilación del fragmento respectivo (p. ej., utilizando un algoritmo directo). Los fragmentos cuyas probabilidades de incidencia no cumplen con un primer valor umbral del primer filtro de umbral (p. ej., sus patrones de metilación son demasiado comunes entre los fragmentos de la población de referencia) se descartan (p. ej., no contribuyen al recuento de bins).

Tamaño mínimo de bolsa. En algunas realizaciones, una condición de filtro de la una o más condiciones de filtro es el requisito de que cada fragmento tenga un tamaño de bolsa mayor que un entero umbral. En otras palabras, que cada fragmento esté representado por más del entero umbral de lecturas de secuencia de la pluralidad de lecturas de secuencia. Por ejemplo, en el caso donde el entero umbral sea uno, cada fragmento debe estar representado por más de una lectura de secuencia de la pluralidad de lecturas de secuencia. En algunas realizaciones, el entero umbral es 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 o un entero entre 10 y 100.

Cantidad mínima de sitios CpG. En algunas realizaciones, una condición de filtro de la una o más condiciones de filtro es el requisito de que cada fragmento cubra un primer número umbral de sitios CpG y sea inferior a una segunda longitud umbral en términos de pares de bases. Por ejemplo, en el caso donde el primer umbral sea 1 sitio CpG y el segundo umbral 1000 pares de bases, cada fragmento debe cubrir más de un sitio CpG y tener una longitud inferior a 1000 pares de bases. En algunas realizaciones, cada fragmento debe cubrir al menos 1, 2, 3, 4, 5, 6, 7, 8, 9 o 10 sitios CpG. En algunas realizaciones, cada fragmento debe tener una longitud inferior a mil, 2 mil, 3 mil o 4 mil pares de bases contiguas.

Hipermetilación o hipometilación. En algunas realizaciones, una condición de filtro de la una o más condiciones de filtro es un requisito de que cada fragmento esté hipermetilado. En algunas realizaciones, una condición de filtro de la una o más condiciones de filtro es un requisito de que cada fragmento esté hipometilado. En algunas realizaciones, la condición de filtro depende del bin. Por ejemplo, la publicación de patente internacional n.º WO2019/195268, titulada "Methylation Markers and Targeted Methylation Probe Panels", presentada el 2 de abril de 2019, describe varias regiones del genoma humano que tienen un estado hipermetilado que está asociado con una o más condiciones de cáncer, así como varias regiones del genoma humano que tienen un estado hipometilado asociado con una o más condiciones de cáncer. En consecuencia, en algunas realizaciones de la presente descripción, uno o más bins de la pluralidad de bins representan cada uno una región genómica correspondiente de las regiones descritas en el documento WO2019/19528, y la condición de filtro de la una o más condiciones de filtro (a) requiere la selección de fragmentos que están hipermetilados cuando se seleccionan fragmentos que se mapean en un bin que representa una región del genoma humano que tiene un estado hipermetilado que está asociado con una o más condiciones de cáncer de los sitios CpG, como se indica en el documento WO2019/195268, y (b) requiere la selección de fragmentos que están hipometilados cuando se seleccionan fragmentos que se mapean en un bin que representa una región del genoma humano que tiene un estado hipometilado que está asociado con una o más condiciones de cáncer de los sitios CpG, como se indica en el documento WO2019/195268.

Como otro ejemplo, la solicitud internacional n.º PCT/US2020/015082, titulada "Detecting Cancer, Cancer Tissue or Origin, or Cancer Type", presentada el 24 de enero de 2020, también describe varias regiones del genoma humano que tienen un estado hipermetilado que está asociado con una o más condiciones de cáncer, así como varias regiones del genoma humano que tienen un estado hipometilado que está asociado con una o más condiciones de cáncer. En consecuencia, en algunas realizaciones de la presente descripción, uno o más bins de la pluralidad de bins representan cada uno una región genómica correspondiente de las regiones descritas en la solicitud internacional n.º PCT/US2020/015082 y la condición de filtro de la una o más condiciones de filtro (a) requieren la selección de fragmentos que están hipermetilados cuando se seleccionan fragmentos que se mapean en un bin que representa una región del genoma humano que tiene un estado hipermetilado que está asociado con una o más condiciones de cáncer de los sitios CpG, como se indica en la solicitud internacional n.º PCT/US2020/015082, y (b) requieren la selección de fragmentos que están hipometilados cuando se seleccionan fragmentos que se mapean en un bin que representa una región del genoma humano que tiene un estado hipometilado que está asociado con una o más

condiciones de cáncer de los sitios CpG, como se indica en la solicitud internacional n.º PCT/US2020/015082. En algunas realizaciones, la una o más condiciones de filtro requieren que se cumpla con el umbral de valor de p y que ese fragmento esté hipermetilado. En algunas realizaciones, la una o más condiciones de filtro requieren que se cumpla con el umbral de valor de p y que ese fragmento esté hipometilado. En algunas realizaciones, la una o más condiciones de filtro son diferentes para cada bin. Por ejemplo, para un bin de la pluralidad de bins, la una o más condiciones de filtro requieren que se cumpla con el umbral de valor de p y que ese fragmento esté hipometilado, mientras que para un segundo bin de la pluralidad de bins, la una o más condiciones de filtro requieren que se cumpla con el umbral de valor de p y ese fragmento esté hipermetilado.

En algunas realizaciones, se impone cualquier combinación de las condiciones de filtro descritas. Por ejemplo, en algunas realizaciones, cada valor de bin respectivo de la primera pluralidad correspondiente de valores de bin para un bin correspondiente de una pluralidad de bins para un sujeto respectivo es el número de fragmentos de ácido nucleico representados en una primera pluralidad correspondiente de lecturas de secuencia, obtenidas de una muestra biológica del sujeto respectivo, que (i) se mapean en la región genómica representada por el bin correspondiente, (ii) tienen un valor de p inferior a 0,01 tras la aplicación del patrón de metilación a un modelo de Markov entrenado, (iii) tienen un tamaño de bolsa superior a uno, (iv) cubren al menos 1 sitio CpG y (v) tienen una longitud inferior a 1000 pares de bases. Si bien se ha presentado un análisis de las condiciones de filtro en relación con los sujetos de un primer conjunto de datos que se utiliza en el aprendizaje por transferencia (es decir, la aplicación a los sujetos de un segundo conjunto de datos), se apreciará que la aplicación de dichas condiciones de filtro se aplica típicamente tanto a los sujetos del conjunto de datos de entrenamiento (el primer conjunto de datos) como al conjunto de datos que se somete a un aprendizaje por transferencia (el segundo conjunto de datos). En algunas realizaciones, las condiciones de filtro que se aplican a los fragmentos del primer conjunto de datos son diferentes de las condiciones de filtro que se aplican a los fragmentos del segundo conjunto de datos.

Superposición de bins. En algunas realizaciones, cuando un fragmento se superpone a varios bins, se le asigna (contribuye al recuento) en cada bin al que se superpone. En algunas realizaciones, cuando un fragmento se superpone a múltiples bins, se le asigna (contribuye al recuento) el bin con el que se superpone más.

Volviendo a la Figura 10A, con referencia al bloque 1003, cada función de extracción de características de la primera pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal del valor de bin de la totalidad o un subconjunto de la pluralidad de bins en forma de un peso independiente para cada bin respectivo de la pluralidad de bins o el subconjunto de la pluralidad de bins.

En algunas realizaciones, como se ha descrito anteriormente, cada función de extracción de características de la primera pluralidad de funciones de extracción de características (obtenidas utilizando el primer conjunto de datos) puede ser una función de extracción de características que es una función lineal o no lineal. Un ejemplo de una función de extracción de características no lineal (que se puede aplicar, p. ej., al segundo conjunto de datos) es:

función 1 = $A * \text{de extracción de características} ([\text{recuento de bins del bin B}] * [\text{recuento de bins del bin C}]) + E * (\text{recuento de bins del bin F})$,

donde

A y E son pesos, cada uno de los cuales toma la forma de un valor real positivo o negativo, y

B, C y F son bins de la pluralidad de bins.

Un ejemplo de una función de extracción de características lineal (que se puede aplicar, p. ej., al segundo conjunto de datos) es:

función 2 = $G * \text{de extracción de características} [\text{recuento de bins del bin H}] + I * [\text{recuento de bins del bin J}] + K * (\text{recuento de bins del bin L})$,

donde

G, I y K son pesos, cada uno de los cuales toma la forma de un valor real positivo o negativo, y

H, J y L son bins de la pluralidad de bins.

Por supuesto, el ejemplo anterior son funciones de extracción de características en las que solo se han especificado unos pocos bins. En la práctica, las funciones de extracción de características especifican muchos más bins que los pocos que se ilustran aquí. Por ejemplo, en algunas realizaciones, los valores y pesos de bin para 50, 100, 200, 1000 o 2000 o más bins o combinaciones de los mismos se especifican mediante una función de extracción de características. En algunas realizaciones, los valores y pesos de bin para al menos el diez por ciento, el veinte por ciento, el cuarenta por ciento, el sesenta por ciento, el ochenta y cinco por ciento o todos los bins aplicables se

especifican mediante una función de extracción de características. En tales realizaciones, algunos de los bins tienen un peso cero.

Una característica (también denominada en la presente memoria valor de característica) que se obtiene utilizando una función de extracción de características puede ser el resultado computacional de introducir los recuentos de bin (del segundo conjunto de datos) en esa función de extracción de características. Por ejemplo, continuando con el ejemplo simplificado anterior, un ejemplo de una característica, tal como la característica 2-1 para el sujeto 2-1 del segundo conjunto de datos, extraída mediante la función 1 de extracción de características puede ser:

Función 2-1 = A_x [recuento de bin del bin B para el sujeto 2-1] * [recuento de bin del bin C para el sujeto 2-1] + E * (recuento de bin del bin F para el sujeto 2-1).

En algunas realizaciones, los valores de características determinan colectivamente un vector para el sujeto. Por ejemplo, en las realizaciones en las que cada función de extracción de características es un componente principal, cada valor de característica incluye los valores de bin proyectados sobre el componente principal particular.

En algunas realizaciones, como se muestra en el bloque 1004, el conjunto de condiciones de cáncer comprende tres o más condiciones de cáncer. Para cada condición de cáncer respectiva del conjunto de condiciones de cáncer, hay dos o más sujetos de la primera pluralidad de sujetos que tienen la condición de cáncer respectiva, y para cada par respectivo de condiciones de cáncer del conjunto de condiciones de cáncer, la aplicación de la primera técnica de extracción de características se realiza aplicando una instancia de la primera técnica de extracción de características a los valores de bin respectivos de los sujetos respectivos del primer conjunto de datos que tienen una primera condición de cáncer o una segunda condición de cáncer del par respectivo de condiciones de cáncer. Cada instancia de la primera técnica de extracción de características contribuye con un subconjunto correspondiente de la primera pluralidad de funciones de extracción de características a la primera pluralidad de funciones de extracción de características. El conjunto de condiciones de cáncer puede consistir en al menos cinco, o al menos diez o al menos 20 condiciones de cáncer únicas. En algunas realizaciones, el conjunto de condiciones de cáncer consiste en 22 condiciones de cáncer únicas. El conjunto de condiciones de cáncer puede incluir cualquier número adecuado de condiciones de cáncer únicas. En algunas realizaciones, cada subconjunto correspondiente de la primera pluralidad de funciones de extracción de características consta de entre cuatro y cien funciones de extracción de características.

La primera técnica de extracción de características puede implicar cualquier técnica adecuada. Por ejemplo, con referencia al bloque 1006, en algunas realizaciones, la aplicación de la primera técnica de extracción de características comprende aplicar un algoritmo de reducción de dimensiones a los sujetos del primer conjunto de datos que tienen una primera condición de cáncer o una segunda condición de cáncer del par respectivo de condiciones de cáncer, identificando de este modo el subconjunto correspondiente de la primera pluralidad de funciones de extracción de características. En algunas realizaciones, una vez que se combinan los subconjuntos que contribuyen a la primera pluralidad de funciones de extracción de características, la primera pluralidad de funciones de extracción de características puede tener entre 500 y 20.000 funciones, entre 500 y 15.000 funciones, entre 500 y 10.000 funciones, entre 500 y 5000 funciones, entre 500 y 4000 funciones, entre 500 y 3000 funciones, entre 500 y 2000 funciones, o entre 500 y 1000 funciones. En algunas realizaciones, la primera pluralidad de funciones de extracción de características puede incluir menos de 500 funciones, menos de 200 funciones o incluso menos de 100 funciones. Además, en algunas realizaciones, cada subconjunto correspondiente de la primera pluralidad de funciones de extracción de características tiene un número menor de funciones; p. ej., consta de entre 2 y 100 funciones de extracción de características. En tales realizaciones, los métodos que se centran en la clasificación de un número limitado de cánceres (p. ej., un tipo de cáncer específico) pueden incluir un pequeño número de funciones de extracción de características.

El algoritmo de reducción de dimensiones puede ser un algoritmo de reducción de dimensiones lineales o un algoritmo de reducción de dimensiones no lineales, como se muestra en el bloque 1008 de la Figura 10B. En algunas realizaciones, el algoritmo de reducción de dimensiones es un algoritmo de análisis de componentes principales, un algoritmo de análisis factorial, un mapeo de Sammon, un análisis de componentes curvilíneos, un algoritmo de incrustación de vecinos estocásticos (SNE), un algoritmo de Isomap, un algoritmo de despliegue de varianza máxima, un algoritmo de incrustación localmente lineal, un algoritmo t-SNE, un algoritmo de factorización matricial no negativa, un algoritmo de análisis de componentes principales de kernel, un algoritmo de análisis de componentes principales de kernel basado en gráficos, un algoritmo de análisis discriminante lineal, un algoritmo de análisis discriminante generalizado, un algoritmo de aproximación y proyección de variedad uniforme (UMAP), un algoritmo de Large Vis, un algoritmo de mapas propios laplacianos o un algoritmo de análisis discriminante lineal de Fisher. Ver, por ejemplo, Fodor, 2002, "A survey of dimension reduction techniques", Center for Applied Scientific Computing, Lawrence Livermore National, Technical Report UCRL-ID-148494; Cunningham, 2007, "Dimension Reduction", University College Dublin, Informe técnico UCD-CSI-2007-7, Zahorian y col., 2011, "Nonlinear Dimensionality Reduction Methods for Use with Automatic Speech Recognition", Speech Technologies. doi:10.5772/16863. ISBN 978-953-307-996-7; y Lakshmi y col. (18 de agosto de 2016). 6.^a Conferencia Internacional sobre Computación Avanzada (IACC) del IEEE de 2016, págs. 31-34, doi:10.1109/IACC.2016.16, ISBN 978-1-4673-8286-1- En consecuencia, con referencia al bloque 1010, en algunas realizaciones, el algoritmo de reducción de dimensiones es un algoritmo de análisis de componentes principales (ACP), y cada función de extracción de características respectiva del subconjunto

correspondiente de la primera pluralidad de funciones de extracción de características comprende un componente principal respectivo derivado del ACP. En tales realizaciones, el subconjunto correspondiente de la primera pluralidad de funciones de extracción de características puede limitarse a un número umbral de componentes principales calculado por el algoritmo de ACP. El número umbral de componentes principales puede ser, por ejemplo, 50, 100, 1000, 1500 o cualquier otro número. En algunas realizaciones, a cada componente principal calculado por el algoritmo de ACP se le asigna un valor propio mediante el algoritmo ACP, y el subconjunto correspondiente de la primera pluralidad de funciones de extracción de características se limita al número umbral de componentes principales a los que se asignan los valores propios más altos.

En algunas realizaciones, con referencia al bloque 1012, las técnicas descritas implican podar la primera pluralidad de funciones de extracción de características utilizando la primera pluralidad de funciones de extracción de características y las indicaciones respectivas de la condición de cáncer de los sujetos respectivos de la primera pluralidad de sujetos, eliminando de este modo una serie de funciones de extracción de características de la primera pluralidad de funciones de extracción de características. Se utilizan técnicas computacionales (p. ej., regresión regularizada/regresión logística regularizada, etc.) para determinar qué características son más informativas con el fin de discriminar entre diferentes tipos de cáncer utilizando el enfoque descrito. De esta manera, se reduce un número de la primera pluralidad de funciones de extracción de características, como también se muestra en el bloque 208 de la Figura 2.

La poda provoca un cierto grado de reducción en el número de funciones de extracción de características de la primera pluralidad de funciones de extracción de características. Por ejemplo, con referencia al bloque 1014, la poda puede provocar una reducción de al menos cuatro veces en el número de funciones de extracción de características de la primera pluralidad de funciones de extracción de características. En algunas realizaciones, la poda puede dar como resultado una reducción de al menos 6, al menos 7 o al menos 8 veces. Por ejemplo, en una realización, la primera pluralidad de funciones de extracción de características puede tener 50 características por cada par de condiciones de cáncer, y la poda puede dar como resultado 6 o 7 características. Dicha poda enriquece ventajosamente el conjunto de datos (reduce la dispersión) y reduce aún más el conjunto de datos, lo que sirve para mejorar la eficiencia y, al mismo tiempo, reduce la cantidad de recursos informáticos necesarios para gestionar el conjunto de datos.

En consecuencia, en algunas realizaciones, la poda de la primera pluralidad de funciones de extracción de características comprende hacer una regresión del primer conjunto de datos basándose en la totalidad o un subconjunto de la primera pluralidad de funciones de extracción de características utilizando un algoritmo de regresión, para asignar de este modo un coeficiente de regresión a cada función de extracción de características respectiva de la primera pluralidad de funciones de extracción de características. La regresión puede conducir a la clasificación de una condición de cáncer del conjunto de condiciones de cáncer para cada sujeto de la primera pluralidad de sujetos. Los coeficientes de regresión, que pueden tener valores positivos o negativos, asignados a cada función de extracción de características se pueden utilizar para determinar qué función de extracción de características de la primera pluralidad de funciones de extracción de características se debe conservar para el posterior procesamiento. Por ejemplo, con referencia al bloque 1016, en algunas implementaciones, se pueden eliminar las funciones de extracción de características a las que el algoritmo de regresión asigna un coeficiente negativo.

En algunas realizaciones, con referencia al bloque 1018, la poda utilizando un algoritmo de regresión implica asignar un coeficiente de regresión a cada función de extracción de características respectiva de la primera pluralidad de funciones de extracción de características, de modo que las funciones de extracción de características a las que se les asigna un coeficiente que no cumple con un determinado umbral de coeficientes se eliminan.

En algunas realizaciones, aplicar un algoritmo de regresión comprende dividir el primer conjunto de datos en una pluralidad de conjuntos, donde cada conjunto de la pluralidad de conjuntos incluye dos o más sujetos que padecen una primera condición de cáncer y dos o más sujetos que padecen una segunda condición de cáncer, hacer una regresión independiente de cada conjunto respectivo de la pluralidad de conjuntos basándose en la totalidad o un subconjunto de la primera pluralidad de funciones de extracción de características en todos los sujetos del conjunto respectivo en función de la indicación respectiva de condición de cáncer en el sujeto del conjunto respectivo utilizando un algoritmo de regresión para asignar de este modo un coeficiente de regresión correspondiente, de una pluralidad de coeficientes de regresión, a cada característica respectiva de la primera pluralidad de funciones de extracción de características. Además, las características de la primera pluralidad de funciones de extracción de características tienen asignado un coeficiente que cumple con un umbral de coeficiente para al menos un porcentaje umbral de la pluralidad de conjuntos. La pluralidad de conjuntos puede constar de entre cinco y cincuenta conjuntos, diez conjuntos o cualquier otro número de conjuntos. El umbral de coeficiente puede ser cero o puede ser mayor que cero.

El algoritmo de regresión puede ser cualquier tipo de regresión. Por ejemplo, en algunas realizaciones, con referencia al bloque 1020 de la Figura 10B, el algoritmo de regresión es una regresión logística. Los algoritmos de regresión logística se describen en Agresti, *An Introduction to Categorical Data Analysis*. 1996, capítulo 5, págs. 103-144, John Wiley & Son, Nueva York. En algunas realizaciones, la regresión logística supone

$$P(x_i) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})},$$

donde,

$x_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ son la pluralidad correspondiente de valores de características para el i -ésimo sujeto correspondiente 126/134, donde el i -ésimo sujeto correspondiente tiene una primera condición de cáncer ($Y = 1$) o una segunda condición de cáncer ($Y = 0$);

$Y \in \{0, 1\}$ es una etiqueta de clase que tiene el valor "1" cuando el sujeto correspondiente i tiene la primera condición de cáncer y tiene el valor "0" cuando el sujeto correspondiente i tiene la segunda condición de cáncer,

β_0 es una intersección, y

$\beta_j = (j = 1, \dots, k)$ es una pluralidad de coeficientes de regresión, donde cada coeficiente de regresión respectivo de la pluralidad de coeficientes de regresión es para una función de extracción de características correspondiente de la primera pluralidad de funciones de extracción de características.

En algunas realizaciones, la regresión logística es una regresión logística del operador de selección y contracción mínima absoluta (LASSO). En tales realizaciones, el estimador $\widehat{\beta}_0, \dots, \widehat{\beta}_k$ LASSO logístico se define como el minimizador de la probabilidad logarítmica negativa:

$$\text{Min} \left(\sum_{i=1}^n \left[-y_i(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}) + \log(1 + \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})) \right] \right),$$

sujeto a la restricción $\sum_{j=1}^k |\beta_j| \leq \lambda$, donde λ es una constante optimizada para cualquier conjunto de datos determinado.

En algunas realizaciones, el algoritmo de regresión es una regresión logística con regularización L1 o L2.

Como se indica en las ecuaciones anteriores, cada $x_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ es la pluralidad correspondiente de valores de características para el i -ésimo sujeto correspondiente y, como tal, cada x_i representa una función de extracción de características correspondiente. Además, cada $\beta_j = (j = 1, \dots, k)$ es el coeficiente de regresión para la función de extracción de características correspondiente. En algunas realizaciones, las funciones de extracción de características que tienen un coeficiente de regresión correspondiente que no cumple con un valor umbral se podan (eliminan) de la pluralidad de funciones de extracción de características. En algunas realizaciones, este valor umbral es cero. Por lo tanto, En tales realizaciones, las funciones de extracción de características que tienen un coeficiente de regresión correspondiente que es cero con respecto a la regresión descrita anteriormente se eliminan de la pluralidad de funciones de extracción de características antes de entrenar al clasificador. En algunas realizaciones, el valor umbral es 0,1. Por lo tanto, En tales realizaciones, las funciones de extracción de características que tienen un coeficiente de regresión correspondiente cuyo valor absoluto es inferior a 0,1 con respecto a la regresión descrita anteriormente se eliminan de la pluralidad de funciones de extracción de características antes de entrenar al clasificador. En algunas realizaciones, el valor umbral es un valor entre 0,1 y 0,3. Un ejemplo de tales realizaciones es el caso donde el valor umbral es 0,2. En tales realizaciones, las funciones de extracción de características que tienen un coeficiente de regresión correspondiente cuyo valor absoluto es inferior a 0,2 con respecto a la regresión descrita anteriormente se eliminan de la pluralidad de funciones de extracción de características antes de entrenar al clasificador.

Si bien las ecuaciones anteriores se refieren solo a dos condiciones de cáncer, se apreciará que se pueden aplicar por pares a más de dos condiciones de cáncer. Por ejemplo, considérese el caso en el que hay tres condiciones de cáncer, A, B y C. Por lo tanto, hay tres pares únicos que deben evaluarse: A frente a B, A frente a C y B frente a C. Se recordará que En tales realizaciones donde hay múltiples pares de condiciones de cáncer, la extracción de las características también se puede realizar por pares. Por lo tanto, en algunas realizaciones donde hay múltiples condiciones de cáncer y las funciones de extracción de características se dedujeron por pares, la poda también puede realizarse por pares, donde cada poda solo funciona para podar del subconjunto de funciones de extracción de características que se obtuvieron para un par dado de condiciones de cáncer. Por ejemplo, en el caso donde hay tres condiciones de cáncer A, B y C, la poda de las funciones de extracción de características obtenidas para A frente a B puede adoptar la forma:

$$P(x_i) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})},$$

donde,

$x_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ son la pluralidad correspondiente de valores de características para el i -ésimo sujeto correspondiente 126/134, donde el i -ésimo sujeto correspondiente tiene una condición de cáncer A ($Y = 1$) o una condición de cáncer B ($Y = 0$);

$Y \in \{0, 1\}$ es una etiqueta de clase que tiene el valor "1" cuando el sujeto i correspondiente tiene la condición de cáncer A y tiene el valor "0" cuando el sujeto i correspondiente tiene la condición de cáncer B,

β_0 es una intersección, y

$\beta_j = (j = 1, \dots, k)$ es una pluralidad de coeficientes de regresión, donde cada coeficiente de regresión respectivo de la pluralidad de coeficientes de regresión es para una función de extracción de características correspondiente del subconjunto de la primera pluralidad de funciones de extracción de características, donde cada función de extracción de características del subconjunto de funciones se obtuvo mediante la reducción de dimensiones utilizando el conjunto de sujetos que tienen una condición de cáncer A o B. En algunas realizaciones, la regresión logística es una regresión logística con operador de selección y contracción mínima absoluta (LASSO). En tales realizaciones, el estimador

LASSO logístico $\hat{\beta}_0, \dots, \hat{\beta}_k$ se define como el minimizador de la probabilidad logarítmica negativa:

$$\text{Min}(\sum_{i=1}^n [-y_i(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}) + \log(1 + \text{Exp}(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}))]),$$

sujeto a la restricción $\sum_{j=1}^k |\beta_j| \leq \lambda$, donde λ es una constante optimizada para cualquier conjunto de datos determinado.

En este caso, cada $x_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ es el subconjunto correspondiente de la pluralidad de valores de características para el i -ésimo sujeto correspondiente y, como tal, cada x_i representa una función de extracción de características correspondiente. Además, cada $\beta_j = (j = 1, \dots, k)$ es el coeficiente de regresión para una función de extracción de características correspondiente. En algunas realizaciones, las funciones de extracción de características que tienen un coeficiente de regresión correspondiente que no cumple con un valor umbral se podan (eliminan) de la pluralidad de funciones de extracción de características. Tal regresión también se repite para los otros pares de condiciones de cáncer para podar del sujeto las funciones de extracción de características asociadas con sus condiciones de cáncer. De esta manera, se seleccionó un conjunto final de funciones de extracción de características a partir del análisis de cada par de condiciones de cáncer. Por ejemplo, las funciones de extracción de características para A frente a B que inspeccionan la poda, las funciones de extracción de características para A frente a C que inspeccionan la poda, etc.

Si bien la determinación de las funciones de extracción de características se puede determinar y podar por pares, la presente descripción no es tan limitada. En algunas realizaciones alternativas, la determinación de las funciones de extracción de características y la posterior poda se realizan de forma multivariante en lugar de por pares.

Bloques de datos 1022-1044

En algunas realizaciones, con referencia al bloque 1022 de la Figura 10C, se obtiene un segundo conjunto 124 de datos que comprende, para cada sujeto respectivo de una segunda pluralidad de sujetos de la especie, la segunda información de genotipo correspondiente. La segunda información de genotipo (p. ej., 134-2-1, ..., 134-1-T) comprende, como se muestra esquemáticamente en un ejemplo del segundo conjunto de datos de la Figura 3, (i) una segunda pluralidad correspondiente de valores de bin, cada valor de bin respectivo de la segunda pluralidad correspondiente de valores de bin representa un bin correspondiente en la pluralidad de bins, y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer. La segunda pluralidad correspondiente de valores de bin de cada sujeto respectivo de la segunda pluralidad de sujetos se obtiene de una muestra biológica del sujeto respectivo, como se muestra a modo de ejemplo con referencia a la Figura 2 (el segundo conjunto 124 de datos se obtiene de las muestras 224 biológicas).

Como se describió anteriormente (bloque 1002) en relación con el primer conjunto de datos (p. ej., en algunas realizaciones, un conjunto de datos del TCGA), la primera pluralidad correspondiente de valores de bin de cada sujeto respectivo de la primera pluralidad de sujetos se obtiene de una muestra biológica obtenida de ese sujeto, donde la muestra comprende un primer tipo de tejido, utilizando un primer método de secuenciación. En las realizaciones ilustradas de la presente descripción, la muestra biológica del sujeto respectivo de la segunda pluralidad de sujetos comprende un segundo tipo de tejido, y la segunda pluralidad de valores de bin de cada sujeto respectivo se obtiene utilizando un segundo método de secuenciación. Por lo tanto, en algunas realizaciones, al menos el primer método de secuenciación difiere del segundo método de secuenciación, o el primer tipo de tejido difiere del segundo tipo de tejido. Por ejemplo, en algunas realizaciones, el segundo conjunto de datos es un conjunto de datos del CCGA. Sin embargo, debe apreciarse que las realizaciones según la presente descripción no se limitan a ningún primer y segundo conjunto de datos específico, ni a ninguna condición de cáncer específica. Además, en algunas realizaciones, el primer y el

segundo conjunto de datos se obtienen mediante el mismo método de secuenciación. En algunas realizaciones, los valores de bin en el segundo conjunto de datos se normalizan de la misma manera que se normalizaron los valores de bin del primer conjunto de datos. Además, en algunas realizaciones, se apreciará que todas las variables se estandarizan (p. ej., restando sus medias y dividiendo por sus desviaciones estándar) cuando sea necesario. En algunas realizaciones, el primer conjunto de datos son los recuentos del número de copias de fragmento obtenidos mediante la secuenciación dirigida o del genoma completo, mientras que el segundo conjunto de datos es el recuento de fragmentos de metilación anormal tal como se determina, por ejemplo, en los ejemplos 5 y 10 junto con cualquier combinación del uno o más filtros de fragmento descritos en la presente memoria, tal como el filtro de valor de p en relación con una cohorte sin cáncer.

En algunas realizaciones, con referencia al bloque 1024, cada función de extracción de características de la segunda pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal del valor de bin de la totalidad o un subconjunto de la pluralidad de bins en forma de un peso independiente para cada bin respectivo de la pluralidad de bins o el subconjunto de la pluralidad de bins.

Los valores de bin de la primera pluralidad de valores de bin de la primera información de genotipado del primer conjunto de datos y la segunda pluralidad de valores de bin de la segunda información de genotipado del segundo conjunto de datos se pueden obtener de varias maneras diferentes. Por ejemplo, en algunas realizaciones, con referencia al bloque 1026, cada valor de bin respectivo de la primera pluralidad de valores de bin o la segunda pluralidad de valores de bin representa un número de fragmentos libres de células medidos en la muestra biológica correspondiente que se mapean en un bin respectivo de la pluralidad de bins.

En algunas realizaciones, la primera o la segunda información de genotipo para cada sujeto respectivo de la primera o la segunda pluralidad de sujetos se obtiene mediante secuenciación dirigida o del genoma completo, con referencia al bloque 1028. La secuenciación dirigida o del genoma completo comprende la secuenciación regular sin metilación o la secuenciación por metilación (p. ej., basándose en la conversión con bisulfito o la conversión enzimática de los sitios CpG). En algunas realizaciones, esta secuenciación del genoma completo se utiliza para la secuenciación de una porción del genoma. En algunas realizaciones, esta porción del genoma es al menos el 10 por ciento, el 20 por ciento, el 30 por ciento, el 40 por ciento, el 50 por ciento, el 60 por ciento, el 70 por ciento, el 80 por ciento, el 90 por ciento, el 95 por ciento, el 99 por ciento, el 99,9 por ciento o la totalidad de un genoma. En algunas realizaciones, la secuenciación del genoma completo obtiene lecturas de secuencia que tienen una longitud de secuencia de 1000 pares de bases o menos con una cobertura de la porción del genoma que es de al menos 5 veces, al menos 10 veces, al menos 15 veces, al menos 20 veces, al menos 25 veces, al menos 30 veces o al menos 50 veces en toda la porción del genoma. En algunas realizaciones, la secuenciación del genoma completo obtiene lecturas de secuencia que tienen una cobertura de al menos 5 veces, al menos 10 veces, al menos 15 veces, al menos 20 veces, al menos 25 veces, al menos 30 veces, al menos 50 veces o al menos 100 veces en todo el genoma.

En algunas realizaciones, la primera información de genotipo para cada sujeto respectivo de la primera pluralidad de sujetos se obtiene mediante secuenciación dirigida, y la segunda información de genotipo para cada sujeto respectivo de la segunda pluralidad de sujetos se obtiene mediante secuenciación del genoma completo, como se muestra en el bloque 1030. Alternativamente, en algunas realizaciones, la primera información de genotipo se obtiene mediante secuenciación del genoma completo, y la segunda información de genotipo se obtiene mediante secuenciación dirigida. En algunas realizaciones, se utiliza una pluralidad de sondas de ácido nucleico en la secuenciación de ácido nucleico dirigida. En algunas realizaciones, estas sondas se mapean en las regiones identificadas en el ejemplo 4. En algunas realizaciones, se usan al menos 1000 sondas, cada una de las cuales se mapea en una ubicación única diferente en el genoma de referencia humano. En algunas realizaciones, al menos 3000 sondas, al menos 4000 sondas, al menos 5000 sondas, al menos 6000 sondas, al menos 7000 sondas, al menos 8000 sondas, al menos 9000 sondas, al menos 10000 sondas, al menos 15000 sondas, al menos 20000 sondas, al menos 25000 sondas, al menos 30000 sondas o al menos 4000 sondas, cada una de las cuales se mapean en una ubicación única diferente en el genoma de referencia humano, se usan en la secuenciación de ácido nucleico dirigida. En algunas realizaciones, cada sonda se diseña para mapearse en un bin particular de la pluralidad de bins y la pluralidad de bins es para una región específica respectiva de un genoma de referencia que se ha identificado como asociado con el cáncer a través de patrones de metilación específicos del cáncer y/o tejido en el ADNlc en relación con los controles sin cáncer, como se describió anteriormente y en el ejemplo 4. En algunas realizaciones, dichas sondas se usan para la secuenciación dirigida (p. ej., la secuenciación por metilación) y las regiones del genoma de referencia en las que estas sondas se mapean de forma única se secuencian con una cobertura de al menos 5 veces, al menos 10 veces, al menos 15 veces, al menos 20 veces, al menos 25 veces, al menos 30 veces, al menos 50 veces, al menos 100 veces, al menos 250 veces, al menos 500 veces, o al menos 1000 veces (para las porciones del genoma en las que se mapean las sondas).

En algunas realizaciones, con referencia al bloque 1032, cada valor de bin respectivo de la primera pluralidad de valores de bin o la segunda pluralidad de valores de bin representa un número de fragmentos que se mapean en un bin respectivo de la pluralidad de bins o un patrón de metilación respectivo medido en la muestra biológica correspondiente para el bin respectivo de la pluralidad de bins. La técnica descrita para utilizar el aprendizaje por transferencia para determinar si un sujeto tiene una condición de cáncer se puede implementar utilizando patrones de metilación medidos en las muestras biológicas de los sujetos. Adicional o alternativamente, dado que los datos de

secuenciación por metilación pueden proporcionar el número de copias, se puede realizar un análisis del número de copias con los datos de secuenciación por metilación.

En realizaciones según la presente descripción, una muestra biológica de un sujeto respectivo de la primera pluralidad o la segunda pluralidad de sujetos puede ser cualquier tipo de muestra obtenida de un sujeto. En algunas realizaciones, la muestra biológica comprende sangre, sangre completa, plasma, suero, orina, líquido cefalorraquídeo, materia fecal, saliva, sudor, lágrimas, líquido pleural, líquido pericárdico o líquido peritoneal de un sujeto respectivo. En algunas realizaciones, la muestra biológica consiste en sangre, sangre completa, plasma, suero, orina, líquido cefalorraquídeo, materia fecal, saliva, sudor, lágrimas, líquido pleural, líquido pericárdico o líquido peritoneal de un sujeto respectivo.

En algunas realizaciones, el aprendizaje por transferencia se basa en el uso de diferentes muestras biológicas de los mismos sujetos. En consecuencia, en algunas realizaciones, el primer y el segundo conjunto de datos pueden generarse basados en el análisis de las muestras adquiridas de la misma cohorte. En tales realizaciones, cada muestra biológica del primer conjunto de datos es un primer tipo de tejido, y cada muestra biológica del segundo conjunto de datos es un segundo tipo de tejido, y al menos algunos de los sujetos del primer conjunto de datos también están en el segundo conjunto de datos. Por ejemplo, las muestras de tejido de un conjunto de datos del CCGA se pueden utilizar para extraer características, y estas características se pueden aplicar a muestras de ADNlc, para las mismas cohortes de sujetos.

En algunas realizaciones, con referencia al bloque 1034, el primer tipo de tejido es sangre y la muestra biológica para cada sujeto respectivo de la primera pluralidad o la segunda pluralidad de sujetos es sangre, y el segundo tipo de tejido es tejido mamario, tejido pulmonar, tejido prostático, tejido colorrectal, tejido renal, tejido uterino, tejido pancreático, tejido esofágico, tejido de cabeza/cuello, tejido ovárico, tejido hepatobiliar, tejido de cuello uterino, tejido tiroideo, tejido vesical. La muestra biológica para cada sujeto respectivo de la segunda pluralidad de sujetos consiste en el segundo tipo de tejido, o la muestra biológica para cada sujeto respectivo de la segunda pluralidad de sujetos comprende un tumor sólido del segundo tipo de tejido.

En diversas realizaciones, un primer tipo de tejido y una muestra biológica para cada sujeto respectivo de la primera pluralidad de sujetos pueden seleccionarse de sangre, cualquier tipo de tejido (p. ej., tejido mamario, tejido pulmonar, tejido prostático, tejido colorrectal, tejido renal, tejido uterino, tejido pancreático, tejido esofágico, tejido de cabeza/cuello, tejido ovárico, tejido hepatobiliar, tejido de cuello uterino, tejido tiroideo, tejido vesical, etc.), o un tumor sólido de un tipo determinado de tejido. Un segundo tipo de tejido y una muestra biológica para cada sujeto respectivo de la segunda pluralidad de sujetos pueden seleccionarse de manera similar de sangre, cualquier tipo de tejido o un tumor sólido de un cierto tipo de tejido. Por ejemplo, en algunas realizaciones, la muestra biológica para cada sujeto respectivo de la primera pluralidad de sujetos comprende un tumor sólido de un determinado tipo de tejido, el segundo tipo de tejido es sangre y la muestra biológica para cada sujeto respectivo de la segunda pluralidad de sujetos es sangre. En algunas realizaciones, cada sujeto respectivo de la primera pluralidad de sujetos comprende un tumor sólido del primer tipo de tejido, y el segundo tipo de tejido es sangre y la muestra biológica para cada sujeto respectivo de la segunda pluralidad de sujetos es sangre.

Una condición de cáncer puede ser cualquier tipo de condición de cáncer. Por ejemplo, en algunas realizaciones, con referencia al bloque 1036, cada condición de cáncer respectiva de un conjunto de condiciones de cáncer se selecciona del grupo que consiste en no cáncer, cáncer de mama, cáncer de pulmón, cáncer de próstata, cáncer colorrectal, cáncer renal, cáncer de útero, cáncer de páncreas, cáncer de esófago, linfoma, cáncer de cabeza/cuello, cáncer de ovario, cáncer hepatobiliar, melanoma, cáncer de cuello uterino, mieloma múltiple, leucemia, cáncer de tiroides, cáncer de vejiga y cáncer gástrico.

El primer tipo de tejido y el segundo tipo de tejido pueden ser el mismo, con referencia al bloque 1038 de la Figura 10D. Sin embargo, en algunas realizaciones, el primer tipo de tejido y el segundo tipo de tejido son diferentes. En algunas realizaciones, como se muestra en el bloque 1040, el primer método de secuenciación o el segundo método de secuenciación generan lecturas de secuencia de ADN libre de células.

En algunas realizaciones según la presente descripción, como se ha descrito anteriormente, se puede reducir el número de la segunda pluralidad de funciones de extracción de características. Por lo tanto, con referencia al bloque 1042 de la Figura 10D, la segunda pluralidad de funciones de extracción de características se puede podar utilizando la segunda pluralidad de funciones de extracción de características y las indicaciones respectivas de la condición de cáncer de los sujetos respectivos de la segunda pluralidad de sujetos, eliminando de este modo una serie de funciones de extracción de características de la segunda pluralidad de funciones de extracción de características. De esta manera, se identifica un número reducido de funciones de extracción de características, como se muestra a modo de ejemplo en el bloque 228 de la Figura 2. En algunas realizaciones (bloque 1043), la poda comprende hacer una regresión del segundo conjunto de datos basándose en la totalidad o un subconjunto de la segunda pluralidad de funciones de extracción de características utilizando un algoritmo de regresión para asignar de este modo un coeficiente de regresión a cada función de extracción de características respectiva de la segunda pluralidad de funciones de extracción de características, y eliminar las funciones de extracción de características de la segunda pluralidad de funciones de extracción de características a las que el algoritmo de regresión asigna un coeficiente negativo. En algunas realizaciones (bloque 1044), la poda comprende hacer una regresión del segundo conjunto de

datos basándose en la totalidad o un subconjunto de la segunda pluralidad de funciones de extracción de características utilizando un algoritmo de regresión para asignar de este modo un coeficiente de regresión a cada función de extracción de características respectiva de la segunda pluralidad de funciones de extracción de características, y eliminar las funciones de extracción de características de la segunda pluralidad de funciones de extracción de características a las que el algoritmo de regresión asigna un coeficiente que no cumple con un umbral de coeficientes. Independientemente de su tipo específico, la regresión conduce a clasificar una condición de cáncer del conjunto de condiciones de cáncer para cada sujeto de la segunda pluralidad de sujetos. En algunas realizaciones, se obtienen las funciones de extracción de características para el segundo conjunto de datos y, a continuación, se podan utilizando las técnicas descritas anteriormente para el primer conjunto de datos. Además, En algunas de tales realizaciones, se obtienen y, posteriormente, se podan por pares, como se ha descrito anteriormente para el primer conjunto de datos. Además, en algunas realizaciones alternativas, se obtienen y, posteriormente, se podan de forma multivariante, como se ha descrito anteriormente para el primer conjunto de datos.

Bloque de datos 1044-1051

En algunas realizaciones, como se muestra en el bloque 1044 de la Figura 10E, los sistemas y métodos descritos para entrenar a uno o más clasificadores para discriminar entre cada condición de cáncer del conjunto de condiciones de cáncer realizan además la transformación, basándose en al menos cada función de extracción de características respectiva de la primera pluralidad de funciones de extracción de características, la segunda pluralidad respectiva de valores de bin de cada sujeto correspondiente de la segunda pluralidad de sujetos con respecto a la función de extracción de características respectiva, contribuyendo de este modo a un segundo conjunto de datos transformado que comprende una pluralidad respectiva de valores de características para cada sujeto correspondiente. Las funciones de extracción de características obtenidas utilizando el primer conjunto de datos (que pueden estar en forma de un número reducido de funciones de extracción de características generadas como resultado de la aplicación de la poda a la primera pluralidad de funciones de extracción de características) se aplican con respecto a los valores de bin de cada sujeto correspondiente de la segunda pluralidad de sujetos. Por ejemplo, en algunas realizaciones, cada función de extracción de características de la primera pluralidad de funciones de extracción de características puede codificar una función lineal o no lineal de los valores de bin de la totalidad o un subconjunto de la pluralidad de bins del primer conjunto de datos. La transformación incluye aplicar la función lineal o no lineal, generada (o “aprendida”) basándose en al menos algunos de los valores de bin del primer conjunto de datos, a los valores de bin en el segundo conjunto de datos. En tales realizaciones, se apreciará que, En tales realizaciones, el índice para la pluralidad de bins del primer y el segundo conjunto de datos es el mismo. En otras palabras, cada bin respectivo en el primer conjunto de datos tiene un bin correspondiente del segundo conjunto de datos, donde los bins correspondientes representan el mismo subconjunto del genoma de referencia de la especie.

En algunas realizaciones, la transformación implica el uso de redes neuronales convolucionales, incluidas las redes neuronales convolucionales profundas. Por lo tanto, con referencia al bloque 1046, en algunas realizaciones, la primera técnica de extracción de características es una primera red neuronal convolucional que comprende una primera pluralidad de capas convolucionales, cada capa convolucional respectiva en la primera pluralidad de capas convolucionales está asociada a un vector de peso aprendido que se obtiene mediante la retropropagación en la primera red neuronal convolucional utilizando los valores de bin respectivos y las indicaciones respectivas de la condición de cáncer de los sujetos respectivos del primer conjunto de datos, y el vector de peso aprendido respectivo de cada capa convolucional de un subconjunto de la primera pluralidad de capas convolucionales representa colectivamente la primera pluralidad de funciones de extracción de características. En tales realizaciones, la transformación (bloque 1044) comprende introducir la segunda pluralidad correspondiente de valores de bin de un sujeto respectivo de la segunda pluralidad de sujetos en una segunda red convolucional que comprende el subconjunto de la primera pluralidad de capas convolucionales, donde un vector de peso de cada capa convolucional respectiva de la segunda red neuronal convolucional se inicializa y, en algunas realizaciones, se congela en valores del vector de peso aprendido de la capa convolucional correspondiente en la primera red neuronal convolucional. Ver, por ejemplo, Yosinski y col., 2014, “How transferable are features in deep neural networks?”, *Advances in Neural Information Processing Systems* 27, páginas 3320-3328. En algunas de tales realizaciones, la primera pluralidad de capas convolucionales comprende cualquier número adecuado de capas. Por ejemplo, en algunas realizaciones, la primera pluralidad de capas convolucionales comprende tres, cuatro, cinco, seis, siete u ocho capas convolucionales, como se muestra en el bloque 1048. Como otro ejemplo, en algunas realizaciones (bloque 1050), la primera pluralidad de capas convolucionales comprende cinco capas convolucionales y el subconjunto de la primera pluralidad de capas convolucionales consiste en las tres primeras capas convolucionales de la primera red neuronal convolucional.

En algunas realizaciones, con referencia al bloque 1051 de la Figura 10E, la transformación (C) se basa en cada función de extracción de características respectiva de la primera pluralidad de funciones de extracción de características y una segunda pluralidad de funciones de extracción de características. La segunda pluralidad de funciones de extracción de características se puede obtener aplicando una segunda técnica de extracción de características a los valores de bin respectivos de los sujetos respectivos del segundo conjunto de datos, identificando de este modo una segunda pluralidad de funciones de extracción de características, como se muestra en la Figura 6. Cada función de extracción de características de la segunda pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal de los valores de bin de la totalidad o un subconjunto de la pluralidad de bins, y la segunda pluralidad de funciones de extracción de características discrimina colectivamente a

los sujetos respectivos de la segunda pluralidad de sujetos según tengan una condición de cáncer dentro del conjunto de condiciones de cáncer basándose en los valores de bin respectivos para los sujetos respectivos.

Bloques de datos 1052-1076

En algunas realizaciones, como se muestra en el bloque 1052 de la Figura 10F, los sistemas y métodos descritos para entrenar a uno o más clasificadores comprenden además, una vez que se genera el segundo conjunto de datos transformado, utilizar el segundo conjunto de datos transformado y la indicación de la condición de cáncer de los sujetos respectivos de la segunda pluralidad de sujetos para entrenar a un primer clasificador (también denominado en la presente memoria “clasificador”) del uno o más clasificadores en al menos un par de condiciones de cáncer del conjunto de condiciones de cáncer. La Figura 8, analizada anteriormente, ilustra el entrenamiento de un clasificador según las realizaciones de la presente descripción.

En algunas realizaciones, como se ha descrito anteriormente (Figura 2), además de aplicar una primera técnica de extracción de características al primer conjunto de datos para identificar una primera pluralidad de funciones de extracción de características, se aplica una segunda técnica de extracción de características al segundo conjunto de datos, p. ej., al segundo conjunto 124 de datos, para generar e identificar una segunda pluralidad de funciones de extracción de características (p. ej., las segundas funciones 224 de extracción de características de la Figura 2). Por lo tanto, con referencia al bloque 1054 de la Figura 10F, el método descrito lleva a cabo la aplicación de una segunda técnica de extracción de características a los valores de bin respectivos de los sujetos respectivos del segundo conjunto de datos, identificando de este modo una segunda pluralidad de funciones de extracción de características. Cada función de extracción de características de la segunda pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal de los valores de bin de la totalidad o un subconjunto de la pluralidad de bins, y la segunda pluralidad de funciones de extracción de características discrimina colectivamente a los sujetos respectivos de la segunda pluralidad de sujetos según tengan una condición de cáncer dentro del conjunto de condiciones de cáncer basándose en los valores de bin respectivos para los sujetos respectivos. El método según algunas realizaciones de la presente descripción incluye además aplicar la segunda pluralidad de funciones de extracción de características a los valores de bin respectivos de los sujetos respectivos del segundo conjunto de datos para obtener una pluralidad respectiva de segundos valores de características para cada sujeto correspondiente del segundo conjunto de datos. Además, con referencia al bloque 1056 de la Figura 10F, en las realizaciones en las que se obtiene la pluralidad de segundos valores de característica, el segundo conjunto de datos transformado comprende la pluralidad respectiva de los segundos valores de característica.

Además, en algunas realizaciones, con referencia al bloque 1057 de la Figura 10F, para cada par respectivo de condiciones de cáncer del conjunto de condiciones de cáncer, la aplicación de la segunda técnica de extracción de características se realiza aplicando una instancia de la segunda técnica de extracción de características a los valores de bin respectivos de los sujetos respectivos del segundo conjunto de datos que tienen una primera condición de cáncer o una segunda condición de cáncer del par respectivo de condiciones de cáncer. Cada instancia de la segunda técnica de extracción de características contribuye con un subconjunto correspondiente de la segunda pluralidad de funciones de extracción de características a la segunda pluralidad de funciones de extracción de características.

En algunas realizaciones, como se ha descrito anteriormente, la aplicación de la primera técnica de extracción de características comprende aplicar un algoritmo de reducción de dimensiones a los sujetos del primer conjunto de datos que tienen una primera condición de cáncer o una segunda condición de cáncer del par respectivo de condiciones de cáncer, identificando de este modo el subconjunto correspondiente de la primera pluralidad de funciones de extracción de características. Además, en algunas realizaciones, con referencia al bloque 1058, la aplicación de la segunda técnica de extracción de características comprende aplicar un algoritmo de reducción de dimensiones a los sujetos del segundo conjunto de datos que tienen una primera condición de cáncer o una segunda condición de cáncer en un par respectivo de condiciones de cáncer, identificando de este modo el subconjunto correspondiente de la segunda pluralidad de funciones de extracción de características. Un algoritmo de reducción de dimensiones se puede aplicar alternativamente a más de dos condiciones de cáncer, p. ej., a tres o más de tres condiciones de cáncer.

El algoritmo de reducción de dimensiones puede ser un algoritmo de reducción de dimensiones lineales o un algoritmo de reducción de dimensiones no lineales. De manera similar a un algoritmo de reducción de dimensiones aplicado a los sujetos del primer conjunto de datos, el algoritmo de reducción de dimensiones que se puede aplicar a los sujetos del segundo conjunto de datos puede ser un algoritmo de análisis de componentes principales, de modo que cada función de extracción de características respectiva del subconjunto correspondiente de la segunda pluralidad de funciones de extracción de características comprenda un componente principal respectivo derivado del algoritmo de reducción de dimensiones. En algunas realizaciones, el subconjunto correspondiente de la segunda pluralidad de funciones de extracción de características se limita a un número umbral de componentes principales calculado por el algoritmo de análisis de componentes principales. Además, en algunas realizaciones, a cada componente principal calculado por el algoritmo de análisis de componentes principales se le asigna un valor propio mediante el algoritmo de componentes principales, y el subconjunto correspondiente de la segunda pluralidad de funciones de extracción de características se limita al número umbral de componentes principales a los que se asignan los valores propios más altos.

En algunas realizaciones, el conjunto de condiciones de cáncer consiste en al menos cinco, al menos diez, al menos 20 condiciones de cáncer únicas. En algunas realizaciones, el conjunto de condiciones de cáncer consiste en 22 condiciones de cáncer únicas, como se muestra en el bloque 1060 de la Figura 10F. En algunas realizaciones, el conjunto de condiciones de cáncer consiste en 231 pares únicos de condiciones de cáncer. Debe apreciarse que el conjunto de condiciones de cáncer puede incluir cualquier otro número de pares únicos de condiciones de cáncer. Con referencia al bloque 1062, en algunas realizaciones, cada subconjunto correspondiente de la segunda pluralidad de funciones de extracción de características consta de entre 4 y 100 funciones de extracción de características. En algunas realizaciones, una vez que se combinan los subconjuntos que contribuyen a la segunda pluralidad de funciones de extracción de características, la segunda pluralidad de funciones de extracción de características puede tener entre 500 y 20.000 funciones, entre 500 y 15.000 funciones, entre 500 y 10.000 funciones, entre 500 y 5000 funciones, entre 500 y 4000 funciones, entre 500 y 3000 funciones, entre 500 y 2000 funciones, o entre 500 y 1000 funciones.

En algunas realizaciones, como se describió anteriormente, se aplica una técnica de extracción de características (ya sea la primera o la segunda técnica de extracción de características) a cada par de condiciones de cáncer de un conjunto de condiciones de cáncer. En otras realizaciones, una técnica de extracción de características se aplica simultáneamente a tres, cuatro o más de cuatro condiciones de cáncer de un conjunto de condiciones de cáncer. Además, en algunas realizaciones, la extracción de características multinomiales se realiza para todas las condiciones de cáncer de un conjunto de condiciones de cáncer simultáneamente.

En algunas realizaciones, con referencia al bloque 1064 de la Figura 10G, el conjunto de condiciones de cáncer comprende tres o más condiciones de cáncer. Como se describió anteriormente (bloque 1004), En tales realizaciones, para cada condición de cáncer respectiva del conjunto de condiciones de cáncer, puede haber dos o más sujetos de la primera pluralidad de sujetos que tienen la condición de cáncer respectiva, la aplicación de la primera técnica de extracción de características comprende además aplicar un algoritmo de reducción de dimensiones a los sujetos del primer conjunto de datos que tienen una condición de cáncer de las tres o más condiciones de cáncer, identificando de este modo la primera pluralidad de funciones de extracción de características.

Además, con referencia al bloque 1066, en las realizaciones en las que el conjunto de condiciones de cáncer comprende tres o más condiciones de cáncer, el uso (bloque 1052 de la Figura 10F) comprende utilizar la pluralidad de segundos valores de características del segundo conjunto de datos transformado para entrenar al primer clasificador, donde la aplicación de la segunda técnica de extracción de características comprende además aplicar un algoritmo de reducción de dimensiones a los sujetos del segundo conjunto de datos que tienen una condición de cáncer de las tres o más condiciones de cáncer, identificando de este modo la segunda pluralidad de funciones de extracción de características. En algunas realizaciones, el algoritmo de reducción de dimensiones es un algoritmo de reducción de dimensiones lineales o un algoritmo de reducción de dimensiones no lineales, como se muestra en el bloque 1067 de la Figura 10G. El algoritmo de reducción de dimensiones puede ser un algoritmo de análisis de componentes principales o cualquier otro tipo de algoritmo de reducción de dimensiones, que puede ser el mismo algoritmo de reducción de dimensiones aplicado a los sujetos del primer conjunto de datos, o un algoritmo de reducción de dimensiones diferente. Los ejemplos no limitativos de un algoritmo de reducción de dimensiones comprenden un algoritmo de análisis factorial, un mapeo de Sammon, un análisis de componentes curvilíneos, un algoritmo de incrustación de vecinos estocásticos (SNE), un algoritmo de Isomap, un algoritmo de despliegue de varianza máxima, un algoritmo de incrustación localmente lineal, un algoritmo t-SNE, un algoritmo de factorización de matriz no negativa, un algoritmo de análisis de componentes principales de kernel, un algoritmo de análisis de componentes principales de kernel basado en gráficos, un algoritmo de análisis discriminante lineal, un algoritmo de análisis discriminante generalizado, un algoritmo de aproximación y proyección de variedad uniforme (UMAP), un algoritmo de Large Vis, un algoritmo de mapas propios laplacianos, un algoritmo de análisis discriminante lineal de Fisher o cualquier otro algoritmo de reducción de dimensiones.

Volviendo al bloque 1052 de la Figura 10F, el clasificador entrenado puede ser uno cualquiera o más de varios clasificadores. El clasificador puede entrenarse utilizando el segundo conjunto de datos transformado y la indicación de la condición de cáncer de los sujetos respectivos de la segunda pluralidad de sujetos. En algunas realizaciones, como se ha descrito anteriormente en relación al menos con los bloques 1054 y 1056, el segundo conjunto de datos transformado comprende la pluralidad respectiva de segundos valores de característica.

En algunas realizaciones, con referencia al bloque 1068, el primer clasificador es un clasificador multivariante. En algunas realizaciones, con referencia al bloque 1070, el uno o más clasificadores consisten en el primer clasificador, y el primer clasificador es un clasificador multivariante único que discrimina cada condición de cáncer del conjunto de condiciones de cáncer. En algunas realizaciones, con referencia al bloque 1072, el primer clasificador es una regresión logística multivariante, un algoritmo de red neuronal, un algoritmo de red neuronal convolucional, un algoritmo de máquina de vector de soporte (SVM), un algoritmo de árbol de decisión, un algoritmo de agrupamiento o un algoritmo de regresión.

Los algoritmos de regresión logística se describen en Agresti, *An Introduction to Categorical Data Analysis*, 1996, capítulo 5, págs. 103-144, John Wiley & Son, Nueva York.

Los algoritmos de redes neuronales, incluidos los algoritmos de redes neuronales convolucionales, se describen en Vincent y col., 2010, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion”, *J Mach Learn Res* 11, págs. 3371-3408; Larochelle y col., 2009, “Exploring strategies for training deep neural networks”, *J Mach Learn Res* 10, págs. 1-40; y Hassoun, 1995, *Fundamentals of Artificial Neural Networks*, Massachusetts Institute of Technology.

Los algoritmos de SVM se describen en Cristianini y Shawe-Taylor, 2000, “An Introduction to Support Vector Machines”, Cambridge University Press, Cambridge; Boser y col., 1992, “A training algorithm for optimal margin classifiers”, en *Actas del quinto taller anual de la ACM sobre teoría del aprendizaje computacional*, ACM Press, Pittsburgh, Pensilvania, págs. 142-152; Vapnik, 1998, *Statistical Learning Theory*, Wiley, Nueva York; Mount, 2001, *Bioinformatics: sequence and genome analysis*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, Nueva York; Duda, *Pattern Classification*, segunda edición, 2001, John Wiley & Sons, Inc., págs. 259, 262-265; y Hastie, 2001, *The Elements of Statistical Learning*, Springer, Nueva York; y Furey y col., 2000, *Bioinformatics* 16, 906-914.

Cuando se utilizan para la clasificación, las SVM separan un conjunto dado de entrenamiento de datos binarios etiquetados con un hiperplano que está a la máxima distancia de los datos etiquetados. En los casos en los que no es posible una separación lineal, las SVM pueden funcionar en combinación con la técnica de los “kernels”, que realiza automáticamente un mapeo no lineal en un espacio de características. El hiperplano encontrado por la SVM en el espacio de características corresponde a un límite de decisión no lineal en el espacio de entrada.

Los árboles de decisión se describen generalmente en Duda, 2001, *Pattern Classification*, John Wiley & Sons, Inc., Nueva York, págs. 395-396. Los métodos basados en árboles dividen el espacio de características en un conjunto de rectángulos y, a continuación, ajustan un modelo (como una constante) en cada uno de ellos. En algunas realizaciones, el árbol de decisión es una regresión forestal aleatoria. Un algoritmo específico que se puede utilizar es un árbol de clasificación y regresión (CART). Otros algoritmos de árbol de decisiones específicos incluyen, aunque no de forma limitativa, ID3, C4.5, MART y Random Forests. CART, ID3 y C4.5 se describen en Duda, 2001, *Pattern Classification*, John Wiley & Sons, Inc., Nueva York, págs. 396-408 y págs. 411-412.

CART, MART y C4.5 se describen en Hastie y col., 2001, *The Elements of Statistical Learning*, Springer-Verlag, Nueva York, capítulo 9.

Los bosques aleatorios se describen en Breiman, 1999, “Random Forests-Random Features”, Informe técnico 567, Departamento de Estadística, U.C. Berkeley, septiembre de 1999.

El agrupamiento se describe en las páginas 211-256 de Duda y Hart, *Pattern Classification and Scene Analysis*, 1973, John Wiley & Sons, Inc., Nueva York (de aquí en adelante, “Duda 1973”). Como se describe en la sección 6.7 de Duda 1973, el problema del agrupamiento se describe como el de encontrar agrupaciones naturales en un conjunto de datos. Para identificar las agrupaciones naturales, se abordan dos cuestiones. En primer lugar, se determina una forma de medir la similitud (o disimilitud) entre dos muestras. Esta métrica (medida de similitud) se utiliza para garantizar que las muestras de un grupo se parezcan más entre sí que a las muestras de otros grupos. En segundo lugar, se determina un mecanismo para dividir los datos en grupos utilizando la medida de similitud.

Las medidas de similitud se analizan en la sección 6.7 de Duda 1973, donde se afirma que una forma de iniciar una investigación de agrupamiento es definir una función de distancia y calcular la matriz de distancias entre todos los pares de muestras del conjunto de entrenamiento. Si la distancia es una buena medida de similitud, entonces la distancia entre las entidades de referencia del mismo grupo será significativamente menor que la distancia entre las entidades de referencia de diferentes grupos. Sin embargo, como se indica en la página 215 de Duda 1973, el agrupamiento no requiere el uso de una métrica de distancia. Por ejemplo, se puede utilizar una función de similitud no métrica $s(x, x')$ para comparar dos vectores x y x' . Convencionalmente, $s(x, x')$ es una función simétrica cuyo valor es grande cuando x y x' son de alguna manera “similares”. En la página 218 de Duda 1973 se proporciona un ejemplo de una función de similitud no métrica $s(x, x')$.

Una vez que se ha seleccionado un método para medir la “similitud” o la “disimilitud” entre los puntos de un conjunto de datos, el agrupamiento requiere una función de criterio que mida la calidad del agrupamiento de cualquier partición de los datos. Las particiones del conjunto de datos que extremen la función de criterio se utilizan para agrupar los datos. Ver la página 217 de Duda 1973. Las funciones de los criterios se analizan en la sección 6.8 de Duda 1973.

Más recientemente, se ha publicado Duda y col., *Pattern Classification*, 2.ª edición, John Wiley & Sons, Inc., Nueva York. Las páginas 537-563 describen la agrupación en detalle. Se puede encontrar más información sobre las técnicas de agrupación en Kaufman y Rousseeuw, 1990, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, Nueva York, NY; Everitt, 1993, *Cluster analysis* (3.ª ed.), Wiley, Nueva York, NY; y Backer, 1995, *Computer-Assisted Reasoning in Cluster Analysis*, Prentice Hall, Upper Saddle River, Nueva Jersey. Las técnicas de agrupamiento particulares ilustrativas que se pueden utilizar en la presente descripción incluyen, aunque no de forma limitativa, el agrupamiento jerárquico (agrupamiento aglomerativo que usa el algoritmo del vecino más cercano, el algoritmo del vecino más lejano, el algoritmo de enlace promedio, el algoritmo de centroide o el algoritmo de suma de cuadrados), el agrupamiento de k-medias, el algoritmo de agrupamiento de k-medias difuso y el agrupamiento de Jarvis-Patrick.

Dicho agrupamiento puede estar en el conjunto de las primeras características $\{p_1, \dots, p_{N-k}\}$ (o los componentes principales derivados del conjunto de las primeras características).

En algunas realizaciones, el agrupamiento comprende un agrupamiento no supervisado (bloque 490) donde no se impone ninguna noción preconcebida de qué grupos deben formarse cuando se agrupa el conjunto de entrenamiento.

En algunas realizaciones, como se muestra en el bloque 1074 de la Figura 10G, el uso del segundo conjunto de datos transformado para entrenar al clasificador (bloque 1052) se repite para cada par único de condiciones de cáncer del conjunto de condiciones de cáncer, entrenando de este modo una pluralidad de clasificadores. Cada clasificador respectivo de la pluralidad de clasificadores se entrena en sujetos del segundo conjunto de datos transformado que tienen una de la primera condición de cáncer o el segundo cáncer del par respectivo de condiciones de cáncer representadas por el clasificador respectivo. En algunas realizaciones, con referencia al bloque 1076, el uno o más clasificadores consisten en el primer clasificador, y el primer clasificador se entrena en una condición de cáncer distinta de una condición de cáncer del conjunto de condiciones de cáncer.

Bloques de datos 1078-1088

El clasificador entrenado según las realizaciones de la presente descripción se utiliza para clasificar a los sujetos de prueba utilizando su información de genotipado, asignando una condición de cáncer a cada uno de los sujetos de prueba. De esta manera, con referencia al bloque 1078 de la Figura 10H, el primer clasificador se utiliza para clasificar a un sujeto de prueba con respecto a una primera condición de cáncer utilizando información de genotipo, en forma electrónica, basándose en datos de lectura de secuencia en una muestra biológica de prueba obtenida del sujeto de prueba como entrada al primer clasificador.

La muestra biológica de prueba puede ser cualquier tipo de muestra. Por ejemplo, en algunas realizaciones, con referencia al bloque 1080, la muestra biológica de prueba comprende sangre, sangre completa, plasma, suero, orina, líquido cefalorraquídeo, materia fecal, saliva, sudor, lágrimas, líquido pleural, líquido pericárdico o líquido peritoneal del sujeto de prueba. En algunas realizaciones, como se muestra en el bloque 1082, la primera condición de cáncer está en el conjunto de condiciones de cáncer.

Además, en algunas realizaciones, con referencia al bloque 1084, el uno o más clasificadores se usan para determinar la probabilidad de que un sujeto de prueba tenga cada condición de cáncer del conjunto de condiciones de cáncer utilizando información de genotipo, en formato electrónico, basándose en datos de lectura de secuencia en una muestra biológica de prueba obtenida del sujeto de prueba como entrada para cada clasificador del uno o más clasificadores. La muestra biológica de prueba puede comprender sangre, sangre completa, plasma, suero, orina, líquido cefalorraquídeo, materia fecal, saliva, sudor, lágrimas, líquido pleural, líquido pericárdico o líquido peritoneal del sujeto de prueba, como se muestra en el bloque 1086.

Como también se muestra en la Figura 10H (bloque 1088), en algunas realizaciones, el uno o más clasificadores consisten en el primer clasificador y el al menos un programa comprende además instrucciones para utilizar el primer clasificador para determinar la probabilidad de que un sujeto de prueba tenga cada condición de cáncer del conjunto de condiciones de cáncer utilizando información de genotipo, en formato electrónico, basándose en datos de lectura de secuencia en una muestra biológica de prueba obtenida del sujeto de prueba como entrada al primer clasificador.

En algunas realizaciones, el escalado de un primer valor de bin respectivo para cada bin respectivo de la pluralidad de bins para cada sujeto respectivo de la primera pluralidad de sujetos se realiza tomando una transformación logarítmica del primer valor de bin respectivo, formando de este modo un primer valor de bin transformado logarítmicamente para el bin respectivo, restando un valor medio del primer valor de bin transformado logarítmicamente de la primera pluralidad de sujetos del primer valor de bin transformado logarítmicamente del bin respectivo, formando de este modo un primer valor de bin normalizado para el bin respectivo y, posteriormente, dividiendo el primer valor de bin normalizado respectivo para el bin respectivo por una desviación estándar del primer valor de bin normalizado de la primera pluralidad de sujetos, escalando de este modo el primer valor de bin para cada bin respectivo de la pluralidad de bins para cada sujeto respectivo de la primera pluralidad de sujetos.

Adicional o alternativamente, en algunas realizaciones, el escalado de un segundo valor de bin respectivo para cada bin respectivo de la pluralidad de bins para cada sujeto respectivo de la segunda pluralidad de sujetos se realiza tomando una transformación logarítmica del segundo valor de bin respectivo, formando de este modo un segundo valor de bin transformado logarítmicamente para el bin respectivo, restando un valor medio del segundo valor de bin transformado logarítmicamente respectivo de la segunda pluralidad de sujetos del segundo valor de bin transformado logarítmicamente del bin respectivo, formando de este modo un segundo valor de bin normalizado para el bin respectivo y, posteriormente, dividiendo el segundo valor de bin normalizado respectivo para el bin respectivo por una desviación estándar del segundo valor de bin normalizado de la segunda pluralidad de sujetos, escalando de este modo el segundo valor de bin para cada bin respectivo de la pluralidad de bins para cada sujeto respectivo de la segunda pluralidad de sujetos.

En algunas realizaciones, el clasificador entrenado que usa el segundo conjunto de datos transformado se utiliza para determinar y aplicar un régimen de tratamiento a un sujeto de prueba basándose, al menos en parte, en un valor del clasificador (p. ej., una condición de cáncer predicha) tras la aplicación de los valores de bin del sujeto de prueba que se obtienen en una de las formas descritas para el primer o el segundo conjunto de datos descrito anteriormente. En algunas realizaciones, el régimen de tratamiento comprende aplicar un agente para el cáncer al sujeto de prueba basándose en la condición de cáncer determinada por el clasificador para el sujeto de prueba. Los ejemplos no limitativos de agentes para el cáncer que se pueden aplicar basándose en una salida del clasificador entrenado utilizando el segundo conjunto de datos transformado incluyen, aunque no de forma limitativa, hormonas, terapias inmunológicas, radiografías y fármacos contra el cáncer. Los ejemplos de fármacos contra el cáncer incluyen, aunque no de forma limitativa, lenalidomida, pembrolizumab, trastuzumab, bevacizumab, rituximab, ibrutinib, vacuna tetravalente contra el virus del papiloma humano (tipos 6, 11, 16 y 18), pertuzumab, pemetrexed, nilotinib, denosumab, acetato de abiraterona, promacta, imatinib, everolimus, palbociclib, erlotinib, borteomib y borteomib.

En algunas realizaciones, el sujeto de prueba ha sido tratado con un agente para el cáncer y los métodos descritos comprenden además utilizar la salida del clasificador entrenado utilizando el segundo conjunto de datos transformado para evaluar la respuesta del sujeto de prueba al agente para el cáncer.

En algunas realizaciones, el sujeto de prueba ha recibido un agente para el cáncer y los métodos descritos comprenden además utilizar una condición de cáncer predicha por el clasificador entrenado utilizando el segundo conjunto de datos transformado para determinar si se debe intensificar (p. ej., aumentar la radiación, aumentar la dosis del fármaco, etc.) o dejar de utilizar el agente para el cáncer en el sujeto.

En algunas realizaciones, el sujeto ha sido sometido a una intervención quirúrgica para tratar una condición de cáncer y el método comprende además utilizar la condición de cáncer predicha por el clasificador entrenado utilizando el segundo conjunto de datos transformado para evaluar una condición del sujeto de prueba en respuesta a la intervención quirúrgica.

En algunas realizaciones, el clasificador entrenado utilizando el segundo conjunto de datos transformado se usa, a su vez, para generar un informe que se comunica a un cuidador (p. ej., un médico, un profesional de la salud, una agencia de seguros, un pariente) asociado con un sujeto de prueba basándose, al menos en parte, en uno o más valores proporcionados por el clasificador (p. ej., una condición de cáncer predicha) tras la aplicación de los valores de bin obtenidos de un sujeto de prueba obtenidos de la misma manera que se obtuvieron los valores de bin para los sujetos en el primer o el segundo conjunto de datos descritos en la presente memoria. En algunas de tales realizaciones, el uno o más valores del clasificador indican que el sujeto de prueba tiene o no alguna de las condiciones de cáncer descritas en la presente memoria. En algunas de tales realizaciones, cada uno del uno o más valores del clasificador proporciona una posibilidad o probabilidad de que el sujeto de prueba tenga o no alguna de las condiciones de cáncer descritas en la presente memoria. En tales realizaciones, el clasificador proporciona una posibilidad o probabilidad, para cada condición de cáncer respectiva de un conjunto de condiciones de cáncer, de que el sujeto de prueba tenga la condición de cáncer respectiva, donde el conjunto de condiciones de cáncer se extrae de cualquiera de las condiciones de cáncer descritas en la presente memoria. El conjunto de condiciones de cáncer puede ser una sola condición de cáncer, dos condiciones de cáncer o cualquier número de condiciones de cáncer, algunas de las cuales se describen en la presente memoria. Como se describe en la presente memoria, ejemplos no limitativos de condiciones de cáncer son orígenes específicos del cáncer (p. ej., mama, pulmón, etc.). Como se describe en la presente memoria, ejemplos adicionales no limitativos de condiciones de cáncer son etapas específicas de un cáncer particular (p. ej., cáncer de mama en etapa I, cáncer de mama en etapa II, etc.).

EJEMPLO 1: Estudio del Atlas del Genoma del Cáncer (TCGA).

En algunas realizaciones, la información de genotipo se obtiene utilizando datos del programa de genómica del cáncer del Atlas del Genoma del Cáncer (TCGA) dirigido por el Instituto Nacional del Cáncer y el Instituto Nacional de Investigación del Genoma Humano. El conjunto de datos del TCGA comprende, entre otra información, perfiles de expresión génica de muestras de tejido disecadas de un gran número de muestras de cáncer humano. La información se obtiene utilizando plataformas de alto rendimiento que incluyen la mutación de la expresión génica, el número de copias, la metilación, etc. El conjunto de datos del TCGA es un conjunto de datos disponible públicamente que comprende más de dos petabytes de datos genómicos de más de 11.000 pacientes con cáncer, que incluye información clínica sobre los pacientes con cáncer, metadatos sobre las muestras (p. ej., el peso de una porción de la muestra, etc.) recolectadas de dichos pacientes, imágenes histopatológicas de portaobjetos de porciones de la muestra e información molecular derivada de las muestras (p. ej., expresión de ARNm/miARN, expresión de proteínas, número de copias, etc.). El conjunto de datos del TCGA incluye datos de secuenciación basados en matrices obtenidos mediante el análisis de matrices de todo el genoma utilizando la matriz Genome-Wide Human SNP Array 6.0 de Affymetrix para sujetos. El conjunto de datos del TCGA incluye dichos datos para sujetos con un cáncer particular conocido y los datos para cada sujeto respectivo provienen del tejido aislado y puro que origina el cáncer en el sujeto respectivo. En el conjunto de datos del TCGA se representan un total de 33 tipos de cáncer diferentes: mama (carcinoma ductal de mama, carcinoma lobular de mama), sistema nervioso central (glioblastoma multiforme, glioma de grado inferior), endocrino (carcinoma adrenocortical, carcinoma papilar tiroideo, paraganglioma y feocromocitoma), gastrointestinal (colangiocarcinoma, adenocarcinoma colorrectal, cáncer de esófago, carcinoma hepatocelular

hepático, adenocarcinoma ductal pancreático y cáncer de estómago), ginecológico (cáncer de cuello uterino, cistadenocarcinoma seroso de ovario, carcinosarcoma uterino y carcinoma endometrial de cuerpo uterino), cabeza y cuello (carcinoma de células escamosas de cabeza y cuello, melanoma uveal), hematológico (leucemia mieloide aguda, timoma), piel (melanoma cutáneo), tejido blando (sarcoma), torácico (adenocarcinoma de pulmón, carcinoma de células escamosas de pulmón y mesotelioma) y urológico (carcinoma de células renales cromóforas, carcinoma renal de células claras, carcinoma papilar de riñón, adenocarcinoma de próstata, cáncer de células germinales testiculares y carcinoma de vejiga urotelial). ver Blum y col., 2018, "TCGA-Analyzed Tumors", SNAPSHOT 173 (2), pág. 530.

EJEMPLO 2: *El estudio del Atlas del genoma libre de células circulantes* (CCGA, por sus siglas en inglés).

En la presente descripción se usaron sujetos del CCGA. El CCGA (NCT02889978) es un estudio prospectivo, multicéntrico, observacional, basado en ADNlc, de casos y controles de detección temprana del cáncer, que inscribió a un grupo equilibrado demográficamente de 15.254 participantes (44 % sin cáncer y 56 % con cáncer) de 142 lugares de Norteamérica con un seguimiento longitudinal, diseñado para desarrollar un único análisis de sangre para detectar más de 50 tipos de cáncer en todas las etapas del cáncer. Ver Liu y col., "Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA", Ann. Oncol 2020, <https://doi.org/10.1016/j.annonc.2020.02.011>.

El estudio del CCGA incluye un ensayo de detección de múltiples cánceres basado en ADN libre de células (ADNlc) en plasma. Se recolectaron hasta 80 ml de sangre completa de sujetos con cáncer recién diagnosticado sin tratamiento previo (C, caso) y de participantes sin diagnóstico de cáncer (no cáncer [NC], control), como se definió en el momento de la inscripción.

Todas las muestras se analizaron mediante: 1) secuenciación dirigida de ADNlc y glóbulos blancos (WBC) emparejada (60.000 veces, panel de 507 genes); un llamador conjunto eliminó las variantes somáticas derivadas de WBC y el ruido técnico residual; 2) secuenciación del genoma completo (WGS; 35 veces) de ADNlc y WBC emparejada; un novedoso algoritmo de aprendizaje automático generó puntuaciones de señales relacionadas con el cáncer; el análisis conjunto identificó eventos compartidos; y 3) secuenciación con bisulfito del genoma completo (WGBS; 34 veces) de ADNlc; se generaron puntuaciones normalizadas utilizando fragmentos anormalmente metilados. En el ensayo dirigido, las variantes somáticas de ADNlc (SNV/Indel) compatibles con WBC no tumorales representaron el 76 % de todas las variantes en NC y el 65 % en C. Coherentes con el mosaicismo somático (p. ej., la hematopoyesis clonal), las variantes compatibles con WBC aumentaron con la edad; varias eran mutaciones de pérdida de función no canónicas no informadas anteriormente. Después de la eliminación de la variante de WBC, las variantes somáticas del controlador canónico fueron altamente específicas para C (p. ej., en EGFR y PIK3CA, 0 NC tuvo variantes vs. 11 y 30, respectivamente, de C). Similarmente, de los 8 NC con alteraciones del número de copias somáticas (SCNA, por sus siglas en inglés) detectadas con WGS, cuatro se derivaron de WBC. Los datos de la WGBS del CCGA revelan CpG informativos a nivel de hiper e hipofragmentos (relación 1:2); un subconjunto de los cuales se usó para calcular las puntuaciones de metilación. Se observó una señal "similar a la del cáncer" constante en menos del 1 % de los participantes NC en todos los ensayos (lo que representa posibles cánceres no diagnosticados). Se observó una tendencia al alza en NC vs. las etapas I-III vs. la etapa IV (nonsyn, SNVS/indels por Mb [media $\pm$ SD] NC: 1,01 $\pm$ 0,86, etapas I-III: 2.43 $\pm$ 3.98; etapa IV: 6.45 $\pm$ 6.79; puntuación de WGS NC: 0,00 $\pm$ 0,08, I-III: 0.27 $\pm$ 0.98; IV: 1.95 $\pm$ 2.33; puntuación de metilación NC: 0 $\pm$ 0.50; I-III: 1.02 $\pm$ 1.77; IV: 3.94 $\pm$ 1.70). Estos datos demuestran la viabilidad de lograr una especificidad superior al 99 % para el cáncer invasivo y respaldan la promesa del ensayo de ADNlc para la detección temprana del cáncer.

EJEMPLO 3

Los inventores realizaron experimentos que demostraron una mejor detección del cáncer utilizando el enfoque de aprendizaje por transferencia según las realizaciones descritas, que se denomina en la presente memoria un enfoque que usa datos del CCGA + TCGA. Los datos de variación en el número de copias del TCGA se generan utilizando los datos de la matriz Affymetrix SNP 6.0 utilizados para identificar las regiones genómicas que se repiten y para inferir el número de copias de estas repeticiones. La matriz Genome-Wide Human SNP Array 6.0 contiene más de 946.000 sondas para la detección de la variación en el número de copias. Este ejemplo se describe con referencia a la Figura 2.

Paso 122 de la Figura 2. Para clasificar el tejido de origen, se obtuvo un primer conjunto de datos. En este ejemplo, el primer conjunto de datos son datos del TCGA descritos en el ejemplo 1 anterior. Los datos del TCGA comprenden, para cada sujeto respectivo de una primera pluralidad de sujetos (en este caso, más de 11.000 pacientes con cáncer) de una especie (humana en este ejemplo), la primera información de genotipo correspondiente que comprende los recuentos del número de copias basados en matrices del TCGA para las sondas representadas en el conjunto de datos del TCGA y una indicación de una condición de cáncer del sujeto respectivo. En este ejemplo, el conjunto de condiciones de cáncer son los 21 tipos de cáncer diferentes que están representados en los conjuntos de datos del TCGA y el CCGA, más la condición sana no cáncer. Se interpolaron los recuentos del número de copias del TCGA basados en matrices para las sondas representadas en el conjunto de datos del TCGA. No se espera que la interpolación de la representación menos densa (TCGA) a la más densa (CCGA) del genoma sea problemática porque

las aberraciones del número de copias en el cáncer frecuentemente son mucho más largas que el espacio entre los marcadores basados en matrices (p. ej., la escala de brazos cromosómicos).

Paso 132 de la Figura 2. Una vez que los datos del número de copias del conjunto de datos del TCGA se procesaron utilizando la interpolación como se describió anteriormente, los datos del número de copias del TCGA procesados (interpolados) para cada sujeto respectivo se agruparon en aproximadamente 30.000 bins que abarcaban todo el genoma y se sometieron al filtrado, p. ej., al filtro de variabilidad alta/baja, descrito en la solicitud de patente estadounidense n.º 16/352.739, titulada “Method and System for Selecting, Managing, and Analyzing Data of High Dimensionality”, presentada el 13 de marzo de 2019, hasta llegar a 23.000 bins. De esta forma, había 23.000 valores de bin correspondientes para cada sujeto en el conjunto de datos del TCGA. Cada uno de estos valores de bin representa un recuento de números de copias. El TCGA proporciona una canalización para convertir los valores de matriz en valores de recuento de números de copias.

Para cada sujeto respectivo en el conjunto de datos del TCGA, se normalizaron los 23.000 valores de bin correspondientes del sujeto respectivo. Se obtiene una mediana del valor de bin de la pluralidad correspondiente de valores de bin para el sujeto respectivo. A continuación, cada valor de bin respectivo de la pluralidad de valores de bin para el sujeto respectivo se divide por esta mediana de valor, asegurando así que los valores de bin para el sujeto respectivo estén centrados en un valor conocido (p. ej., en cero):

$$bv_i^* = \frac{bv_i}{Mediana(bv_j)}$$

donde, bv_i = el valor de bin del bin i de la pluralidad de valores de bin para el sujeto respectivo, bv_i^* = el valor de bin normalizado del bin i de la pluralidad de valores de bin para el sujeto respectivo tras esta normalización, y $mediana(bv_j)$ = la mediana del de bin en la pluralidad de valores de bin no normalizados para el sujeto respectivo. Ver la solicitud de patente estadounidense n.º 16/352.739, titulada “Method and System for Selecting, Managing, and Analyzing Data of High Dimensionality”, presentada el 13 de marzo de 2019.

A partir de los valores de bin normalizados de los sujetos respectivos en el conjunto de datos del TCGA, se aprendieron las primeras funciones (filtros de número de copias) de extracción de características que serían útiles para diferenciar los tipos de cáncer. Esto se hizo para cada par de tipos de cáncer objeto de examen.

Fueron objeto de examen 231 pares $(21 + 1) * 21/2$, donde 21 es el número de condiciones de cáncer (en este ejemplo, el origen del cáncer) y 1 representa la ausencia de cáncer. Para cada par respectivo de los 231 pares, todos los sujetos del conjunto de datos del TCGA que tenían una de las condiciones de cáncer representadas por el par se sometieron colectivamente a un análisis de componentes principales (ACP). De esta manera, se realizaron 231 análisis de componentes principales separados. Para cada ACP respectivo, se conservaron los primeros 1000 componentes principales que explican la variación en el recuento del número de copias en los 23.000 bins en los sujetos que contribuyeron al ACP (uno de los 231 pares).

Paso 208 de la Figura 2. El análisis de componentes principales (ACP) redujo la dimensionalidad del problema de clasificación al generar funciones de extracción de características (en forma de componentes principales derivados del análisis de componentes principales de cada primera pluralidad respectiva de valores de bin en el conjunto de datos del TCGA) que, al inspeccionarlos, correspondían a las ponderaciones de los bins que destacaban eventos de amplificación y delección particulares.

Para aumentar la probabilidad de que el ACP identifique aberraciones en el número de copias que difieren entre los 21 tipos de cáncer considerados en este ejemplo, se mantuvieron (mediante poda) los 50 primeros componentes principales (CP) para cada par de cánceres del TCGA que coincidían con los tipos de cáncer del CCGA y para cada uno de los 21 cánceres con respecto a los no cáncer, lo que dio como resultado un total de 231 conjuntos de 50 CP cada uno. El número 231 se obtiene considerando 231 pares $(21 + 1) * 21/2$, donde 21 es el número de condiciones de cáncer (en este ejemplo, el origen del cáncer) y 1 representa no cáncer. Como tal, la primera pluralidad de funciones de extracción de características consistía en 231×50 CP, o un total de 11.550 funciones 132 de extracción de características.

Aunque los CP se adaptan independientemente a cada par de clases de diagnóstico, se generaron por diseño sin un conocimiento explícito de los tipos de cáncer en cada par. Para abordar esta limitación, se determinó cuáles de los CP eran informativos para la clasificación del tejido de origen. Para reducir el conjunto de 11.550 CP, dentro de los datos del TCGA, se realizó una regresión logística regularizada de LI para clasificar cada uno de los 231 pares, y solo se conservaron los CP con coeficientes distintos de cero en la primera pluralidad de funciones de extracción de características. Esto dio como resultado un promedio de 6,5 CP (funciones de extracción de características) por par de condiciones de cáncer, reduciendo a 1502 el número de CP conservados. Estos CP se interpretaron como las funciones de extracción de características que mejor diferenciaban entre las clases de diagnóstico.

Pasos 224, 124, 142 y 228 de la Figura 2. El ACP generó 20 funciones adicionales de extracción de características específicas de ADNlc (segundas funciones 226 de extracción de características) en los datos de la WGS de un conjunto de entrenamiento de sujetos del CCGA.

Paso 150 de la Figura 2. Las 1502 características derivadas del conjunto de datos del TCGA (primeras funciones de extracción de características) y las 20 características derivadas de la WGS de los sujetos de entrenamiento del conjunto de datos del CCGA (segundas funciones de extracción de características) se concatenaron para formar un total de 1522 funciones de extracción de características. Para cada sujeto respectivo de los 2000 sujetos del conjunto de datos del CCGA, para cada función de extracción de características respectiva de las 1522 funciones de extracción de características, se tomó el producto escalar entre los valores de bin de la WGS normalizados del sujeto respectivo y la combinación lineal de pesos de la función de extracción de características respectiva para completar de este modo el conjunto de datos transformado con un valor de característica para el sujeto respectivo para la función de extracción de características respectiva. De esta manera, se obtuvieron 2000 vectores para el conjunto de datos transformado (de aquí en adelante denominado “conjunto de datos del TCGA + CCGA”) para enfatizar el aprendizaje por transferencia que se produce entre los conjuntos de datos del TCGA y CCGA, como se describió anteriormente. Cada vector del conjunto de datos del TCGA + CCGA representa un sujeto de los 2000 sujetos del conjunto de datos del CCGA y cada vector incluye 1522 elementos, donde cada elemento representa el valor de característica para una función de extracción de características correspondiente en las 1522 funciones de extracción de características aplicadas a los valores de bin del sujeto respectivo.

Paso 160 de la Figura 2. Los 2000 vectores para los 2000 sujetos (conjunto de entrenamiento) del conjunto de datos del TCGA + CCGA del paso 150 se usaron para entrenar una regresión logística regularizada L2 (primer clasificador 160) utilizando las etiquetas conocidas de origen del cáncer de los 2000 sujetos. Este clasificador se denomina clasificador del TCGA + CCGA en este ejemplo.

Paso 214 de la Figura 2. Una vez entrenados en el conjunto de entrenamiento, el modelo logístico entrenado (clasificador del TCGA + CCGA) se aplicó al conjunto de prueba para producir probabilidades de tejido de origen (TOO, por sus siglas en inglés). Los datos del conjunto de pruebas se procesaron previamente de la misma manera que los datos del conjunto de entrenamientos (en términos de obtención y normalización de valores de bin), excepto que se conocían las etiquetas del conjunto de pruebas (es decir, el cáncer de origen). Había más de 160 sujetos en el conjunto de prueba, y cada uno de los cánceres en estudio en este ejemplo estaba representado en el conjunto de prueba. La Figura 11 ilustra la sensibilidad general del clasificador TCGA + CCGA en comparación con un enfoque que emplea un clasificador entrenado utilizando únicamente datos del CCGA de WGS (denominado “clasificador CCGA” en la Figura 11). En la Figura 11, la sensibilidad se califica como la probabilidad de que el clasificador califique a un sujeto para un tipo de cáncer en particular, dado que el sujeto tiene ese tipo de cáncer en particular. Por ejemplo, en el caso del cáncer de pulmón, la probabilidad de que el clasificador identifique a un sujeto como portador de cáncer de pulmón cuando, de hecho, tiene cáncer. Para la Figura 11, la especificidad es para todos los tipos de cánceres considerados en el ejemplo. Como se ilustra en la Figura 11, el aprendizaje por transferencia a partir de los datos del TCGA mejora el rendimiento del clasificador de origen del cáncer (el clasificador TCGA + CCGA en forma de un modelo logit multinomial de la regresión logística) en comparación con el clasificador de origen del cáncer entrenado con los datos del CCGA de WGS sin el aprendizaje por transferencia (el “clasificador CCGA”). Se esperan resultados similares cuando se utiliza un clasificador de árbol de decisiones en lugar de una regresión logística multinomial.

La Figura 12 ilustra los resultados de la detección de varios tipos de cáncer utilizando el clasificador CCGA + TCGA en comparación con el clasificador CCGA WGS. En la Figura 12, se muestran las puntuaciones primccat (“categoría de cáncer primario”) para cada tipo de cáncer. Como muestra la Figura 12, en la mayoría de los casos, el clasificador TCGA + CCGA seleccionó el cáncer verdadero como la opción número uno (primccat) más frecuentemente que el clasificador WGS CCGA. Por ejemplo, en el caso de los cánceres de ovario de un conjunto de datos de prueba extraído de la cohorte del CCGA, el clasificador TCGA + CCGA seleccionó el cáncer de ovario el cincuenta por ciento de las veces (0,5) cuando el sujeto tenía cáncer de ovario, mientras que el clasificador CCGA WGS seleccionó el cáncer de ovario el veinticinco por ciento de las veces (0,25) cuando el sujeto tenía cáncer de ovario.

La Figura 13 ilustra los valores predictivos positivos (la probabilidad de que los sujetos con una prueba de detección positiva tengan realmente la enfermedad) para el clasificador TCGA + CCGA frente al clasificador CCGA WGS para varios cánceres respectivos con respecto a un conjunto de sujetos de prueba extraídos de la cohorte del CCGA. Por ejemplo, de los sujetos del conjunto de prueba de sujetos extraídos de la cohorte del CCGA que tienen cáncer de cabeza/cuello, el clasificador TCGA + CCGA identifica que el 47 por ciento de dichos sujetos tiene cáncer de cabeza/cuello, mientras que el clasificador CCGA WGS identifica que el 43 por ciento de esos sujetos tiene cáncer de cabeza/cuello.

La Figura 14 ilustra la sensibilidad de la predicción para el clasificador TCGA + CCGA (líneas 1402 en la Figura 14, con puntos de datos redondos) frente a un clasificador CCGA (líneas 1404 en la Figura 14, con puntos de datos cuadrados) para los varios cánceres respectivos con respecto a un conjunto de sujetos de prueba extraídos de la cohorte del CCGA. En la Figura 14, se proporciona la sensibilidad a N primero, donde N es 1 (1 primero), 2 (2 primeros) o 3 (3 primeros). Por ejemplo, si solo hay una muestra de cáncer, y es de pulmón, y si las tres primeras predicciones del clasificador (en términos de probabilidad de origen del cáncer) son 1) de mama, 2) de pulmón y 3) colorrectal, las

N primeras sensibilidades son: 1 primera: 0 %, 2 primeras: 100 %, 3 primeras: 100 %. Para la Figura 14, el clasificador CCGA se entrenó utilizando datos de secuenciación por metilación disponibles en el conjunto de datos del CCGA (datos de la WGBS descritos en el ejemplo 2 anterior) en lugar de datos de la WGS. Si bien los datos de la WGBS generalmente mejoraron la predicción del tipo de cáncer, el rendimiento utilizando los datos del CCGA+TCGA (basados en la secuenciación WGS) fue comparable.

La Figura 15 ilustra los N primeros valores predictivos positivos (VPP) de varios tipos de cáncer, donde, como en la Figura 14, N es 1, 2 o 3 excepto que ahora la métrica es PPV en lugar de sensibilidad y donde, como en la Figura 14, las líneas con círculos representan que el clasificador CCGA se entrenó utilizando los datos de secuenciación por metilación disponibles en el conjunto de datos del CCGA (los datos de la WGBS descritos en el ejemplo 2 anterior) en lugar de los datos de la WGS, y las líneas con cuadrados representan el clasificador CCGA+TCC (WGS). Como se muestra, el rendimiento utilizando el clasificador CCGA+TCGA WGS es comparable al del clasificador CCGA WGBS.

La Figura 16 ilustra la aplicación de los métodos según la presente descripción para discriminar entre muestras de tejido canceroso y no canceroso. El gráfico de la izquierda ilustra la sensibilidad vs. (1-especificidad) de la predicción de la presencia de cánceres de alta señal utilizando el clasificador WGS CCGA (línea 1602 que comienza con una sensibilidad de aproximadamente 0,2) y utilizando el clasificador CCGA + TCGA WGS (línea 1604 que comienza con una sensibilidad de aproximadamente 0,4). El gráfico de la derecha ilustra la probabilidad del clasificador CCGA + TCGA frente a la probabilidad del clasificador CCGA WGS, y muestra la alta sensibilidad de la detección del cáncer utilizando el enfoque descrito. En este gráfico, los verdaderos positivos se muestran como círculos de color gris claro, mientras que los falsos positivos se muestran como círculos negros. La mayoría de los verdaderos positivos se encuentran en la parte superior derecha del gráfico, y la mayoría de los falsos positivos se encuentran en la parte inferior izquierda del gráfico. La Figura 16 ilustra que el clasificador de aprendizaje por transferencia CCGA + TCGA WGS reduce el número de falsos positivos, aumenta el número de verdaderos positivos y casi duplica el rendimiento general de detección de cánceres de alta señal. Como se utiliza en la presente memoria, el cáncer de alta señal es cualquier cáncer distinto del de tiroides uterina, el cáncer de mama HR+ en etapa I/II y el cáncer de próstata.

Ejemplo 4: Ejemplos de bins para realizaciones de metilación.

En algunas realizaciones, los bins de la presente descripción se diseñan para abarcar solo regiones objetivo del genoma humano. Este ejemplo resume la identificación de regiones adecuadas del genoma humano para incluirse en dichos bins. Basándose en los resultados del ejemplo 2, tal como se describe con más detalle en Liu y col., “Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA”, Ann. Oncol 2020, <https://doi.org/10.1016/j.annonc.2020.02.011>, se identificaron las porciones del genoma humano (el genoma hg19, Vogelstein y col, 2013, “Cancer genome landscapes”, Science 339 1546-1558) que se predijo que contenían patrones de metilación específicos del cáncer y/o tejido en el ADNlc en comparación con los controles sin cáncer, y se seleccionaron las regiones más informativas para que estuvieran representadas por los bins de una realización de la presente descripción.

Específicamente, después del tratamiento con bisulfito, los fragmentos de ADNlc seleccionados que contenían patrones de metilación anormales en relación con los controles sin cáncer de ambas cadenas se enriquecieron utilizando sondas biotiniladas. En resumen, se diseñaron sondas de ADN biotiniladas de 120 pb para apuntar al enriquecimiento del ADN convertido con bisulfito a partir de fragmentos hipermetilados (CpG metilados al 100 %) o fragmentos hipometilados (CpG no metilados al 100 %); las sondas colocaron en mosaico las regiones objetivo con una superposición del 50 % entre las sondas adyacentes. Un algoritmo personalizado alineó las sondas candidatas con el genoma y calificó el número de eventos de mapeo dentro y fuera de la diana. Las sondas con un mapeo elevado fuera de la diana se omitieron del panel final de regiones que se representarían mediante los bins de una realización de la presente descripción.

Como se describe en la solicitud de patente estadounidense n.º 15/931.022, titulada “Model Based Featurization and Classification”, presentada el 13 de mayo de 2020, se identificó un panel de metilación dirigido, la totalidad o una porción del cual está representada por los bins de una realización de la presente descripción, que cubren 103.456 regiones distintas (17,2 Mb), que cubren 1.116.720 CpG, utilizando los datos de bisulfito del genoma completo obtenidos del subestudio CCGA-1 del CCGA. Este incluyó 363.033 CpG en 68.059 regiones (7,5 Mb) cubiertas por sondas dirigidas a fragmentos hipometilados; 585.181 CpG en 28.521 regiones (7,4 Mb) cubiertas por sondas dirigidas a fragmentos hipermetilados; y 218.506 CpG en 6.876 regiones (2,3 Mb) dirigidas a ambos tipos de fragmentos. Las regiones objetivo anormales individuales contenían entre 1 y 590 CpG, con una mediana de recuento de CpG de 3 para las regiones objetivo hipometiladas y de 6 para las regiones objetivo hipermetiladas. Los CpG estuvieron presentes en las siguientes regiones genómicas utilizando la nomenclatura de Cavalcante y Sartor, 2017, “annotatr: genomic regions in context”, Bioinformatics33(15):2381-2383: 193.818 (17 %) en la región de 1 a 5 kbp corriente arriba de los sitios de inicio de la transcripción (TSS); 278.872 (24 %) en promotores (< 1 kbp corriente arriba de TSS); 500.996 (43 %) en intrones; 292.789 (25 %) en exones; 247.752 (21 %) en los límites intrón-exón (es decir, 200 pb corriente arriba o corriente abajo de cualquier límite entre un exón y un intrón); los límites son con respecto a la cadena del gen); 134.144 (11 %) en regiones no traducidas en 5'; 28.388 (2,4 %) en regiones no traducidas en 3'; 182.174 (16 %) entre genes; y los 1.817 restantes (<1 %) no fueron anotados. Los porcentajes fueron relativos al número total

de CpG y no suman el 100 % porque cada CpG podría recibir múltiples anotaciones debido a la superposición de genes y/o transcripciones.

Ejemplo 5: *Filtrado del valor de P: asignación de un valor de p a los fragmentos basándose en su estado de metilación.*

En algunas realizaciones, un valor de p para el vector de estado de metilación correspondiente de cada fragmento respectivo representado por una pluralidad observada de lecturas de secuencia se compara con los vectores de estado de metilación de los fragmentos de un grupo de control sano. Ver, por ejemplo, la solicitud de patente estadounidense n.º 15/931.022, titulada “Model Based Featurization and Classification”, presentada el 13 de mayo de 2020. La puntuación de valor de p describe la probabilidad de observar una molécula de ácido nucleico que tenga un estado de metilación que coincida con ese vector de estado de metilación en el grupo de control sano. Para determinar si un fragmento está anómalamente metilado (debido a que su valor de p está por debajo de un determinado umbral de valor p), se utiliza un grupo de control sano (una cohorte de sujetos sin cáncer) con una mayoría de fragmentos normalmente metilados. Al realizar este análisis probabilístico para determinar fragmentos anómalos, la determinación tiene peso en comparación con el grupo de sujetos de control que forman el grupo de control sano. Para garantizar la solidez del grupo de control sano, se utiliza un número umbral de individuos sanos para obtener muestras, incluidos fragmentos de ADN (cohorte de sujetos sin cáncer).

En algunas realizaciones, se crea una estructura de datos de un grupo de control sano para un grupo de control sano (una cohorte de sujetos sin cáncer). Para crear una estructura de datos de grupo de control sano, se obtiene una pluralidad de fragmentos (p. ej., ADNlc) de una pluralidad de individuos sanos. A continuación, junto con la Figura 21, se describe un método 2100 para generar una estructura de datos para un grupo de control sano con el que se calculan puntuaciones de valor de p. A continuación, junto con la Figura 22, se describe un método para calcular una puntuación de valor de p con la estructura de datos generada.

La Figura 21 es un diagrama de flujo que describe un proceso 2100 para generar una estructura de datos para un grupo de control sano, según una realización de la presente descripción. Para crear una estructura de datos de grupo de control sano, se obtiene una pluralidad de fragmentos de ADN (p. ej., ADNlc) de una pluralidad de individuos sanos. Se identifica un vector 2052 de estado de metilación para cada fragmento, por ejemplo, mediante el proceso ilustrado en las Figuras 18 y 19 junto con el ejemplo 10 más adelante.

Con el vector 2052 de estado de metilación de cada fragmento, el vector de estado de metilación se subdivide en cadenas de sitios CpG 2105. En una realización, el vector 2052 de estado de metilación se subdivide de modo que las cadenas resultantes son todas menores que una longitud dada. Por ejemplo, un vector 2052 de estado de metilación de longitud 11 puede subdividirse en cadenas de longitud menor o igual a 3, lo que 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 inferior o igual a 4, lo que 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 2052 de estado de metilación es más corto o tiene una longitud igual a la longitud de la cadena especificada, entonces el vector de estado de metilación puede convertirse en una sola cadena que contenga todos los sitios CpG del vector.

Las cadenas se cuentan contando 2110, para cada posible sitio CpG y posibilidad de estados de metilación en el vector 2052, 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 dado y teniendo en cuenta longitudes de cadena de 3, hay 2^3 u 8 posibles configuraciones de cadena. En ese sitio CpG dado, para cada una de las 8 configuraciones de cadena posibles, se realiza un recuento del número de incidencias de cada posibilidad del vector 2052 de estado de metilación que aparecen en el grupo de control. Continuando con este ejemplo, esto puede implicar contar las siguientes cantidades: $\langle M_x, M_{x+1}, M_{x+2} \rangle$, $\langle M_x, M_{x+1}, U_{x+2} \rangle$, ..., $\langle U_x, U_{x+1}, U_{x+2} \rangle$ para cada sitio CpG inicial x en el genoma de referencia. La estructura de datos almacena los recuentos contados para cada sitio CpG inicial y posibilidad 2115 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 puede aumentar drásticamente. Por ejemplo, una longitud máxima de cadena de 4 significa que cada sitio CpG tiene al menos 2^4 números para contar cadenas de longitud 4. Aumentar la longitud máxima de la cadena a 5 significa que cada sitio CpG tiene 2^4 o 16 números adicionales para contar, lo que aumenta al doble la cantidad de números para contar (y la memoria de computadora requerida) en comparación con la longitud de la cadena anterior. La reducción del tamaño de la cadena ayuda a mantener la creación y el rendimiento de la estructura de datos (p. ej., para accesos posteriores, como se describe más adelante) razonables en términos de cálculo y almacenamiento. En segundo lugar, una consideración estadística para limitar la longitud máxima de la cadena es evitar sobreajustar los modelos corriente abajo que utilizan los recuentos de cadena. Si las cadenas largas de sitios CpG no tienen, biológicamente, un efecto fuerte en el resultado (p. ej., predicciones de anomalías que predicen la presencia de cáncer), calcular probabilidades basándose en grandes cadenas de sitios CpG puede resultar problemático, ya que requiere una cantidad significativa de datos que pueden no estar disponibles y, por lo tanto, serían demasiado escasos para que un modelo funcione adecuadamente. Por ejemplo, calcular una probabilidad de anomalía/cáncer 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 anómalo o no.

La Figura 22 es un diagrama de flujo que describe un proceso 2200 para identificar fragmentos anómalamente metilados de un individuo, según una realización. En el proceso 2200, los vectores 2052 de estado de metilación se generan a partir de fragmentos de ADNlc del sujeto utilizando los métodos descritos en el ejemplo 10 junto con las Figuras 19 y 20. Cada vector 2052 de estado de metilación se procesa de la siguiente manera.

Para un vector 2052 de estado de metilación dado, se enumeran 2230 todas las posibilidades de los vectores de estado de metilación que tienen el mismo sitio CpG inicial y la misma longitud (p. ej., conjunto de sitios CpG) en el vector de estado de metilación. Dado que cada estado de metilación generalmente está metilado o no metilado, cada sitio CpG tiene realmente dos estados posibles y, por lo tanto, el recuento de posibilidades distintas de vectores de estado de metilación depende de una potencia de 2, de modo que un vector de estado de metilación de longitud n estaría asociado con 2^n posibilidades de los vectores de estado de metilación. Con vectores de estado de metilación que incluyen estados indeterminados para uno o más sitios CpG, las posibilidades de vectores de estado de metilación se enumeran teniendo en cuenta solo los sitios CpG que tienen estados observados.

La probabilidad de observar cada posibilidad de vector de estado de metilación para el sitio CpG inicial identificado y la longitud de vector de estado de metilación se determinan accediendo a la estructura de datos del grupo de control sano. En un ejemplo, el cálculo de la probabilidad de observar una posibilidad dada utiliza una probabilidad de cadena de Markov para modelar el cálculo de probabilidad conjunta. En uno de dichos enfoques, se obtuvo un modelo de cadena de Markov basándose en los estados de metilación de una cohorte de 131 participantes sin cáncer (no incluidos en el estudio del CCGA del ejemplo 2) y se usó para asignar un valor de p a cada fragmento, que representa la probabilidad de observar los estados de metilación del fragmento en el ADNlc no cáncer. Ver Liu y col., "Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA", Ann. Oncol 2020, <https://doi.org/10.1016/j.annonc.2020.02.011>.

En otras realizaciones, se usan métodos de cálculo distintos de las probabilidades de cadena de Markov para determinar la probabilidad de observar cada posibilidad de vector de estado de metilación.

En algunas realizaciones, se calcula una puntuación de valor de p para el vector 2052 de estado de metilación utilizando las probabilidades calculadas para cada posibilidad 2250. En una realización, esto incluye identificar la probabilidad calculada correspondiente a la posibilidad de una coincidencia para el vector de estado de metilación en cuestión. Específicamente, se trata de la posibilidad de tener el mismo conjunto de sitios CpG o, de manera similar, el mismo sitio CpG inicial y la misma longitud que el vector de estado de metilación. Las sumas calculadas representan las probabilidades de todas las posibilidades que tienen probabilidades inferiores o iguales a la probabilidad identificada para generar la puntuación del valor p .

Este valor de p representa la probabilidad de observar el vector 2052 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 corresponde generalmente a un vector de estado de metilación que es raro en un individuo sano, y que hace que el fragmento se marque como anómalamente 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 individuo sano. Por ejemplo, si el grupo de control sano es un grupo sin cáncer, un valor de p bajo indica que el fragmento está anómalamente metilado en relación con el grupo sin cáncer y, por lo tanto, posiblemente sea indicativo de la presencia de cáncer en el sujeto de prueba.

Las puntuaciones de valor de p se calculan 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 cuáles de los fragmentos están anómalamente metilados, el conjunto de vectores de estado de metilación se filtra basándose en sus puntuaciones 2260 de valor de p . En una realización, el filtrado se realiza comparando las puntuaciones de valores de p con 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.

Según los resultados de ejemplo del proceso, una mediana (rango) de 2800 (entre 1500 y 12.000) fragmentos con patrones de metilación anómalos para los participantes sin cáncer durante el entrenamiento, y una mediana (rango) de 3000 (entre 1200 y 220.000) fragmentos con patrones de metilación anómalos para los participantes con cáncer en entrenamiento. Estos conjuntos filtrados de fragmentos con patrones de metilación anómalos pueden usarse para los análisis corriente abajo, como se describe en la presente memoria.

En un ejemplo, se utiliza una ventana deslizante para determinar las posibilidades de los vectores de estado de metilación y calcular los valores de p 2255. En lugar de enumerar las posibilidades y calcular los valores de p para todos los vectores de estado de metilación, se enumeran las posibilidades y se calculan los valores de p solo para una ventana de sitios CpG contiguos, 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. Se calcula una puntuación de valor de p para la ventana que incluye el primer sitio de CpG. A continuación, se “desliza” la ventana al segundo sitio CpG en el vector y se calcula otra puntuación de valor de p para la segunda ventana. Por lo tanto, para un tamaño de ventana l y longitud m de vector de metilación, cada vector de estado de metilación generará puntuaciones de valor de p $m-l+1$. 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 realización, las puntuaciones de valor de p de los vectores de estado de metilación se agregan para generar una puntuación de valor de p general.

Utilizando 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. Para dar un ejemplo realista, es posible que los fragmentos tengan hacia arriba de 54 sitios CpG. En lugar de calcular las probabilidades de 2^{54} ($\sim 1,8 \times 10^{16}$) posibilidades para generar una sola puntuación de p , se puede utilizar una ventana de sitios CpG de tamaño 5 (por ejemplo), lo que da como resultado 50 cálculos de valor de p para cada una de las 50 ventanas del vector de estado de metilación de ese fragmento. Cada uno de los 50 cálculos enumera 2^5 (32) posibilidades de vectores de estado de metilación, cuyos resultados totales dan como resultado 50×2^5 ($1,6 \times 10^3$) 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.

En las realizaciones con estados indeterminados en algunos o todos sus sitios CpG, se calcula una puntuación de valor de p sumando los sitios CpG con estados indeterminados en el vector de estado de metilación de un fragmento. Se determinan todas las posibilidades que están en consenso con todos los estados de metilación del vector de estado de metilación, excluyendo los estados indeterminados. Se asigna una probabilidad al vector de estado de metilación como una suma de las probabilidades de las posibilidades identificadas. A modo de ejemplo, una probabilidad de un vector de estado de metilación de $\langle M_1, I_2, U_3 \rangle$ se calcula como una suma de las probabilidades de las posibilidades de los vectores de estado de metilación de $\langle M_1, M_2, U_3 \rangle$ y $\langle M_1, U_2, U_3 \rangle$, ya que se observan estados de metilación para los sitios CpG 1 y 3 y están en consenso con los estados de metilación del fragmento en los sitios CpG 1 y 3. Este método de sumar sitios CpG con estados indeterminados utiliza cálculos de probabilidades de posibilidades de hasta 2^i , donde i indica el número de estados indeterminados en el vector de estados de metilación. En realizaciones adicionales, se puede implementar un algoritmo de programación dinámica para calcular la probabilidad de un vector de estados de metilación con uno o más estados indeterminados. Ventajosamente, el algoritmo de programación dinámica funciona en un tiempo computacional lineal.

En una realización, la carga computacional de calcular probabilidades y/o puntuaciones de valor de p se reduce aún más al almacenar en caché al menos algunos cálculos. Por ejemplo, los cálculos de probabilidades para las posibilidades de vectores de estado de metilación (o ventanas de los mismos) se pueden almacenar en caché en la memoria transitoria o persistente. Si otros fragmentos tienen los mismos sitios CpG, el almacenamiento en caché de las probabilidades de posibilidad permite un cálculo eficiente de los valores de puntuación de p sin necesidad de volver a calcular las probabilidades de posibilidad subyacentes. De manera equivalente, las puntuaciones de valor de p pueden calcularse para cada una de las posibilidades de vectores de estado de metilación asociados con un conjunto de sitios CpG del vector (o ventana del mismo). Las puntuaciones de valor de p pueden almacenarse en caché para 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.

En algunas realizaciones, los fragmentos anómalos se identifican como fragmentos con más de un número umbral de sitios CpG y con más de un porcentaje umbral de sitios CpG metilados (hipermetilados) o con más de un porcentaje umbral de sitios CpG no metilados (hipometilados) 2270. Los umbrales de ejemplo para la longitud de los fragmentos (o sitios CpG) incluyen más de 3, 4, 5, 6, 7, 8, 9, 10, etc. Los porcentajes umbral de ejemplo de metilación o no metilación incluyen más del 80 %, 85 %, 90 % o 95 %, o cualquier otro porcentaje dentro del rango de 50 %-100 %.

Ejemplo 6: *Sondas y paneles para ensayos de cáncer.*

En diversas realizaciones, los modelos predictivos de cáncer descritos en la presente memoria usan muestras enriquecidas utilizando un panel de detección de cáncer que comprende una pluralidad de sondas o una pluralidad de pares de sondas. En la técnica se conocen varios paneles de detección de cáncer dirigidos, por ejemplo, como se describe en el documento WO 2019/195268 titulado “Methylation Markers and Targeted Methylation Probe Panels”, presentado el 2 de abril de 2019, el documento PCT/US2019/053509, presentado el 27 de septiembre de 2019 y el documento PCT/US2020/015082 titulado “Detecting Cancer, Cancer Tissue or Origin, or Cancer Type”, presentado el 24 de enero de 2020. Por ejemplo, en algunas realizaciones, el panel de detección de cáncer puede diseñarse para incluir una pluralidad de sondas (o pares de sondas) que pueden capturar fragmentos que juntos pueden proporcionar información relevante para la detección del cáncer. En algunas realizaciones, un panel incluye al menos 50, 100, 500, 1000, 2000, 2500, 5000, 6000, 7500, 10.000, 15.000, 20.000, 25.000 o 50.000 pares de sondas. En otras realizaciones,

un panel incluye al menos 500, 1000, 2000, 5000, 10.000, 12.000, 15.000, 20.000, 30.000, 40.000, 50.000 o 100.000 sondas. La pluralidad de sondas en conjunto puede comprender al menos 0,1 millones, 0,2 millones, 0,4 millones, 0,6 millones, 0,8 millones, 1 millón, 2 millones, 3 millones, 4 millones, 5 millones, 6 millones, 7 millones, 8 millones, 9 millones o 10 millones de nucleótidos. Las sondas (o pares de sondas) se diseñan específicamente para dirigirse a una o más regiones genómicas metiladas diferencialmente en muestras cancerosas y no cancerosas. Las regiones genómicas objetivo se pueden seleccionar para maximizar la precisión de la clasificación, en función de una dotación de tamaño (determinada por la dotación de secuenciación y profundidad de secuenciación deseada).

Las muestras enriquecidas mediante un panel de detección de cáncer pueden someterse a una secuenciación dirigida. Generalmente, las muestras enriquecidas utilizando el panel de detección de cáncer puede usarse para detectar la presencia o ausencia de cáncer y/o proporcionar una clasificación de cáncer tal como el tipo de cáncer, la etapa de cáncer tal como I, II, III o IV, o proporcionar el tejido de origen donde se cree que el cáncer se origina. Dependiendo del propósito, un panel puede incluir sondas (o pares de sondas) dirigidas a regiones genómicas que están metiladas diferencialmente entre muestras cancerosas generales (pancancerosas) y muestras no cancerosas, o solo muestras cancerosas con un tipo específico de cáncer (p. ej., objetivos específicos de cáncer de pulmón). Específicamente, se diseña un panel de detección de cáncer basándose en los datos de secuenciación con bisulfito generados a partir del ADN libre de células (ADNlc) o el ADN genómico (ADNg) de individuos con cáncer y/o sin cáncer.

En algunas realizaciones, el panel de detección de cáncer diseñado mediante los métodos proporcionados en la presente memoria comprende al menos 1000 pares de sondas, cada par de los cuales comprende dos sondas configuradas para superponerse entre sí mediante una secuencia superpuesta que comprende un fragmento de 30 nucleótidos. El fragmento de 30 nucleótidos comprende al menos cinco sitios CpG, en donde al menos el 80 % de los al menos cinco sitios CpG son CpG o UpG. El fragmento de 30 nucleótidos se configura para unirse a una o más regiones genómicas en muestras cancerosas, en donde una o más regiones genómicas tienen al menos cinco sitios de metilación con un patrón de metilación anormal. Otro panel de detección de cáncer comprende al menos 2000 sondas, cada una de las cuales se diseña como una sonda de hibridación complementaria a una o más regiones genómicas. Cada una de las regiones genómicas se selecciona basándose en el criterio de que comprende (i) al menos 30 nucleótidos y (ii) al menos cinco sitios de metilación, en donde los al menos cinco sitios de metilación tienen un patrón de metilación anormal y están hipometilados o hipermetilados.

Cada una de las sondas (o pares de sondas) puede diseñarse para dirigirse a una o más regiones genómicas objetivo. Las regiones genómicas objetivo se seleccionan basándose en varios criterios diseñados para aumentar el enriquecimiento selectivo de fragmentos de ADNlc relevantes y, al mismo tiempo, disminuir el ruido y las uniones no específicas. Por ejemplo, un panel puede incluir sondas que pueden unirse selectivamente y enriquecer fragmentos de ADNlc que están diferencialmente metilados en muestras cancerosas. En este caso, la secuenciación de los fragmentos enriquecidos puede proporcionar información relevante para el diagnóstico del cáncer. Además, las sondas pueden diseñarse para dirigirse a regiones genómicas que se determina que tienen un patrón de metilación anormal y/o patrones de hipermetilación o hipometilación para proporcionar una selectividad y especificidad adicionales de la detección. Por ejemplo, las regiones genómicas se pueden seleccionar cuando las regiones genómicas tienen un patrón de metilación con un valor de p bajo según un modelo de Markov entrenado en un conjunto de muestras no cancerosas, que además cubren al menos 5 CpG, el 90 % de los cuales está metilado o no metilado. En otras realizaciones, las regiones genómicas se pueden seleccionar utilizando modelos de mezcla, como se describe en la presente memoria.

Cada una de las sondas (o pares de sondas) puede dirigirse a regiones genómicas que comprenden al menos 25 pb, 30 pb, 35 pb, 40 pb, 45 pb, 50 pb, 60 pb, 70 pb, 80 pb o 90 pb. Las regiones genómicas pueden seleccionarse en función de si contienen menos de 20, 15, 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) están metilados o no metilados en muestras no cancerosas o cancerosas.

Las regiones genómicas pueden filtrarse adicionalmente para seleccionar solo aquellas que probablemente sean informativas basándose en sus patrones de metilación, por ejemplo, los sitios CpG que están metilados diferencialmente entre muestras cancerosas y no cancerosas (p. ej., anómalamente metilados o no metilados en muestras cancerosas frente a no cancerosas). Para la selección, el cálculo se puede realizar con respecto a cada sitio CpG. En algunas realizaciones, se determina en un primer recuento el número de muestras que contienen cáncer (cancer_count) que incluyen un fragmento que se superpone a ese CpG, y se determina en un segundo recuento el número de muestras totales que contienen fragmentos que se superponen a ese sitio CpG (total). Las regiones genómicas pueden seleccionarse basándose en criterios correlacionados positivamente con el número de muestras que contienen cáncer (cancer_count) que incluyen un fragmento que se superpone a ese sitio CpG, y correlacionados inversamente con el número total de muestras que contienen fragmentos que se superponen a ese sitio CpG (total).

En una realización, se cuenta el número de muestras no cancerosas ($n_{\text{sin cáncer}}$) y el número de muestras cancerosas ($n_{\text{cáncer}}$) que tienen un fragmento que se superpone a un sitio CpG. A continuación, se calcula la probabilidad de que una muestra sea cancerosa, por ejemplo como $(n_{\text{cáncer}} + 1) / (n_{\text{cáncer}} + n_{\text{sin cáncer}} + 2)$. Los sitios CpG se clasifican según esta métrica y se añaden a un panel con avidez hasta que se agote la dotación de tamaño del panel.

Dependiendo de si se pretende que el ensayo sea un ensayo pancanceroso o un ensayo de un solo cáncer, o dependiendo del tipo de flexibilidad que se desee a la hora de elegir qué sitios de CpG contribuyen al panel, las muestras que se utilizan para el recuento de cánceres pueden variar. Un panel para diagnosticar un tipo específico de cáncer (p. ej., TOO) puede diseñarse utilizando un proceso similar. En esta realización, para cada tipo de cáncer, y para cada sitio CpG, la ganancia de información se calcula para determinar si se debe incluir una sonda dirigida a ese sitio CpG. La ganancia de información se calcula para las muestras con un determinado tipo de cáncer en comparación con todas las demás muestras. Por ejemplo, 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). “CT” es una variable aleatoria binaria que indica si el cáncer es de un tipo particular (p. ej., cáncer de pulmón o cáncer distinto de pulmón). Se puede calcular la información mutua con respecto a “CT” dada “AF”. Es decir, cuántos bits de información sobre el tipo de cáncer (pulmonar frente a no pulmonar en el ejemplo) se obtienen si se sabe si hay un fragmento anómalo que se superpone a un sitio CpG particular. Esto se puede utilizar para clasificar los CpG basándose en su grado de especificidad para un tipo de cáncer en particular (p. ej., TOO). Este procedimiento se repite para una pluralidad de tipos de cáncer. Por ejemplo, si una región particular está comúnmente metilada diferencialmente solo en el cáncer de pulmón (y no en otros tipos de cáncer o no cáncer), los CpG en esa región tenderían a tener altas ganancias de información para el cáncer de pulmón. Para cada tipo de cáncer, los sitios CpG se clasifican según esta métrica de ganancia de información y, a continuación, se añaden con avidez a un panel hasta que se agote la dotación de tamaño para ese tipo de cáncer.

Se puede realizar una filtración adicional para seleccionar regiones genómicas objetivo que tengan regiones genómicas fuera de la diana inferiores a un valor umbral. Por ejemplo, una región genómica se selecciona solo cuando hay menos de 15, 10 u 8 regiones genómicas fuera de la diana. En otros casos, la filtración se realiza para eliminar las regiones genómicas cuando la secuencia de las regiones genómicas objetivo aparece más de 5, 10, 15, 20, 25 o 30, 31 veces en un genoma. Se puede realizar una filtración adicional para seleccionar las regiones genómicas objetivo cuando una secuencia, homóloga en un 90 %, 95 %, 98 % o 99 % a las regiones genómicas objetivo, aparece menos de 15, 10 u 8 veces en un genoma, o para eliminar las regiones genómicas objetivo cuando la secuencia, homóloga en un 90 %, 95 %, 98 % o 99 % a las regiones genómicas objetivo, aparece más de 5, 10, 15, 20, 25 o 30 veces en un genoma. 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 realizaciones, se demostró que era necesaria una superposición fragmento-sonda de al menos 45 pb para lograr una cantidad no despreciable de extracción (aunque este número puede ser diferente dependiendo de los detalles 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 lo tanto, la eficiencia de la extracción. Por lo 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 lo tanto, en una realización, se califica el número de dichas regiones. Las mejores sondas tienen una puntuación de 1, lo que significa que coinciden en un solo lugar (la región diana prevista). Se aceptan las sondas con una puntuación baja (por ejemplo, menos de 5 o 10), pero se descartan las sondas por encima de esa puntuación. Pueden usarse otros valores de corte para muestras específicas.

En varias realizaciones, las regiones genómicas objetivo seleccionadas pueden ubicarse en varias posiciones en un genoma, que incluyen, aunque no de forma limitativa, exones, intrones, regiones intergénicas y otras partes. En algunas realizaciones, pueden añadirse sondas dirigidas a regiones genómicas no humanas, tales como aquellas dirigidas a regiones genómicas virales.

Ejemplo 7: Selección de las regiones genómicas humanas utilizadas para los bins.

En algunas realizaciones de la presente descripción, cada bin de la pluralidad de bins se extrae de un panel de regiones genómicas que se diseña para la selección dirigida de patrones de metilación específicos de cáncer. En algunas realizaciones, cada una de dichas regiones genómicas se extrae de la tabla 2 de la solicitud de patente internacional n.º PCT/US2020/015082, titulada “Detecting Cancer, Cancer Tissue or Origin, or Cancer Type”, presentada el 24 de enero de 2020, que se reproduce a continuación:

Lista	Regiones genómicas diana	Id. de sec. n.º		Tamaño del panel (Mb)
		Primera	Última	
1	34844	1	34844	6,43
2	67431	34845	102275	12,14
3	94955	102276	197230	17,72
4	23941	197231	221171	4,63
5	56624	221172	277795	16,42
6	52850	277796	330645	10,45
7	14284	330646	344929	8,48
8	1370	344930	346299	0,39

9	2842	346300	349141	0,79
10	7483	349142	356624	1,94
11	12328	356625	368952	3,08
12	14725	368953	383677	3,65
13	3814	383678	387491	0,62
14	7730	387492	395221	1,26
15	19424	395222	414645	3,23
16	38061	414646	452706	6,58

Las Id. de sec. n.º 452.706-483.478 del documento PCT/US2020/015082 proporcionan información adicional sobre ciertas regiones genómicas objetivo hipermetiladas o hipometiladas. Estos registros de la Id. de sec. n.º identifican regiones genómicas diana que pueden metilarse diferencialmente en muestras de pares específicos de tipos de cáncer. Las regiones genómicas objetivo de las Id. de sec. n.º 452.706-483.478 del documento PCT/US2020/015082 se extraen de la lista 6 del documento PCT/US2020/015082. Muchas de las mismas regiones genómicas objetivo también se encuentran en las listas 1-5 y 7-16 del documento PCT/US2020/015082. La entrada para cada Id. de sec. indica la ubicación cromosómica de la región genómica objetivo en relación con hg19, si los fragmentos de ADNlc que se enriquecerán de la región están hipermetilados o hipometilados, la secuencia de una cadena de ADN de la región genómica objetivo y el par o pares de tipos de cáncer que están diferencialmente metilados en esa región genómica. Debido a que el estado de metilación de algunas regiones genómicas objetivo distingue más de un par de tipos de cáncer, cada entrada identifica un primer tipo de cáncer, como se indica en la tabla 3 del documento PCT/US2020/015082, incluido el listado de secuencias citado en la presente memoria y uno o más segundos tipos de cáncer.

En algunas realizaciones, la pluralidad de bins de la presente descripción incluye un bin separado para cada una de las 200, 500, 1000, 5000, 10.000, 15.000, 20.000, 30.000, 40.000 o 50.000 regiones genómicas objetivo en cualquiera de las listas 1-16 del documento PCT/US2020/015082.

En algunas realizaciones, la pluralidad de bins de la presente descripción incluye un bin separado para cada uno de al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas objetivo en cualquiera de las listas 1-16 del documento PCT/US2020/015082.

En algunas realizaciones, la pluralidad de bins de la presente descripción incluye un bin separado para cada uno de 200, 500, 1000, 5000, 10.000, 15.000, 20.000, 30.000, 40.000 o 50.000 regiones genómicas objetivo en cualquier combinación de las listas 1-16 del documento PCT/US2020/015082.

En algunas realizaciones, la pluralidad de bins de la presente descripción incluye un bin separado para cada uno de al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas objetivo en cualquier combinación de las listas 1-16 del documento PCT/US2020/015082.

En algunas realizaciones, la pluralidad de bins de la presente descripción incluye un bin separado para cada uno de al menos 200, 500, 1000, 5000, 10.000, 15.000, 20.000, 30.000, 40.000 o 50.000 regiones genómicas objetivo en una cualquiera de las listas 1-3 del documento PCT/US2020/015082.

En algunas realizaciones, la pluralidad de bins de la presente descripción incluye un bin separado para cada uno de al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas objetivo en una cualquiera de las listas 1-3 del documento PCT/US2020/015082.

En algunas realizaciones, la pluralidad de bins de la presente descripción incluye un bin separado para al menos 200, 500, 1000, 5000, 10.000, 15.000, 20.000, 30.000, 40.000 o 50.000 regiones genómicas objetivo en una cualquiera de las listas 13-16 del documento PCT/US2020/015082.

En algunas realizaciones, la pluralidad de bins de la presente descripción incluye un bin separado para cada uno de al menos el 10 %, el 20 %, el 25 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas objetivo en una cualquiera de las listas 13-16 del documento PCT/US2020/015082.

En algunas realizaciones, la pluralidad de bins de la presente descripción incluye un bin separado para cada uno de al menos 200, 500, 1000, 5000, 10.000, 15.000, 20.000, 30.000, 40.000 o 50.000 regiones genómicas objetivo seleccionadas de la lista 12 del documento PCT/US2020/015082.

En algunas realizaciones, la pluralidad de bins de la presente descripción incluye un bin separado para cada uno de al menos el 10 %, el 20 %, el 25 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas objetivo de la lista 12 del documento PCT/US2020/015082.

En algunas realizaciones, la pluralidad de bins de la presente descripción incluye un bin separado para cada uno de al menos 200, 500, 1000, 5000, 10.000, 15.000, 20.000, 30.000, 40.000 o 50.000 regiones genómicas objetivo seleccionadas de una cualquiera de las listas 8-11 del documento PCT/US2020/015082.

5 En algunas realizaciones, la pluralidad de bins de la presente descripción incluye un bin separado para cada uno de al menos el 10 %, el 20 %, el 25 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas objetivo en una cualquiera de las listas 8-11 del documento PCT/US2020/015082.

10 En algunas realizaciones, la pluralidad de bins de la presente descripción incluye un bin separado para cada uno de al menos 200, 500, 1000, 5000, 10.000, 15.000, 20.000, 30.000, 40.000 o 50.000 regiones genómicas objetivo seleccionadas de la lista 4 del documento PCT/US2020/015082.

15 En algunas realizaciones, la pluralidad de bins de la presente descripción incluye un bin separado para cada uno de al menos el 10 %, el 20 %, el 25 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas objetivo de la lista 4 del documento PCT/US2020/015082.

Ejemplo 8: *Regiones genómicas humanas seleccionadas adicionales utilizadas para los bins.*

20 En algunas realizaciones de la presente descripción, cada bin de la pluralidad de bins se extrae de un panel de regiones genómicas que se diseña para la selección dirigida de patrones de metilación específicos de cáncer. En algunas realizaciones, cada una de estas regiones genómicas se extrae de la Tabla 2 de la solicitud de patente internacional n.º PCT/US2019/053509, publicada como WO2020/669350A1, titulada “Methylated Markers and Targeted Methylation Probe Panel”, presentada el 27 de septiembre de 2019, que se reproduce a continuación:

Nombre	Lista	Rango de la Id. de sec. n.º en el documento WO2020/069350
Panel 1 de ensayo	Lista 1	1-1644
Panel 2 de ensayo	Lista 2	1645-5270
Panel 3 de ensayo	Lista 3	5271-16837
Panel 3A de ensayo	Lista 4	16838-25984
Panel 4 de ensayo	Lista 5	25985-46929
Panel 4A de ensayo	Lista 6	46930-67335
Panel 5 de ensayo	Lista 7	67336-101617
Panel 6 de ensayo	Lista 8	101618-131822

El listado de secuencias del documento WO2020/669350A1 incluye la siguiente información: (1) Id. de sec. n.º, (2) un identificador de secuencia que identifica (a) un cromosoma o cóntigo en el que se ubica el sitio CpG y (b) una posición de inicio y terminación de la región, (3) la secuencia correspondiente a (2) y (4) si la región se incluyó basándose en su puntuación de hipermetilación o hipometilación. Los números de cromosomas y las posiciones de inicio y terminación se proporcionan en relación con un genoma de referencia humana conocido, GRCh37/hgl9. La secuencia de GRCh37/hgl9 está disponible en el Centro Nacional de Información Biotecnológica (NCBI, por sus siglas en inglés), el Consorcio de Referencia del Genoma y el navegador de genomas proporcionado por el Instituto de Genómica de Santa Cruz.

Generalmente, un bin puede abarcar cualquiera de los sitios CpG incluidos dentro de los bins de inicio/terminación de cualquiera de las regiones objetivo incluidas en las Listas 1-8 del documento WO2020/069350.

50 En algunas realizaciones, la pluralidad de bins de la presente descripción incluye un bin separado para cada uno de 200, 500, 1000, 5000, 10.000, 15.000, 20.000, 30.000, 40.000 o 50.000 regiones genómicas objetivo en una cualquiera de las listas 1-8 del documento WO2020/069350.

55 En algunas realizaciones, la pluralidad de bins de la presente descripción incluye un bin separado para cada uno de al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas objetivo en cualquiera de las listas 1-8 del documento WO2020/069350.

60 En algunas realizaciones, la pluralidad de bins de la presente descripción incluye un bin separado para cada uno de 200, 500, 1000, 5000, 10.000, 15.000, 20.000, 30.000, 40.000 o 50.000 regiones genómicas objetivo en cualquier combinación de las listas 1-8 del documento WO2020/069350.

En algunas realizaciones, la pluralidad de bins de la presente descripción incluye un bin separado para cada uno de al menos el 20 %, el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas objetivo en cualquier combinación de las listas 1-8 del documento WO2020/069350.

65 Ejemplo 9: *Regiones genómicas humanas seleccionadas adicionales utilizadas para los bins.*

En algunas realizaciones de la presente descripción, cada bin de la pluralidad de bins se extrae de un panel de regiones genómicas que se diseña para la selección dirigida de patrones de metilación específicos de cáncer. En algunas realizaciones, cada uno de dichos bins corresponde a una región genómica en cualquiera de las tablas 1-24 de la solicitud de patente internacional n.º PCT/US2019/025358, publicada como WO2019/195268A2, titulada “Methylated Markers and Targeted Methylation Probe Panels”, presentada el 2 de abril de 2019.

En algunas realizaciones, cada bin de la presente descripción se mapea en una región genómica enumerada en las tablas 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23 y/o 24 del documento WO2019/195268A2.

En algunas realizaciones, la totalidad de la pluralidad de los bins de la presente descripción se configura en conjunto para mapearse en al menos el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas de una o más de las tablas 1-24 del documento WO2019/195268A2. En algunas de tales realizaciones, cada bin de la pluralidad de bins se mapea en una sola región genómica correspondiente única de cualquiera de las tablas 1-24 del documento WO2019/195268A2. En algunas de tales realizaciones, un bin de la pluralidad de mapas de bin de la presente descripción se mapea en una, dos, tres, cuatro, cinco, seis, siete, ocho, nueve o diez regiones genómicas correspondientes únicas en cualquier combinación de las tablas 1-24 del documento WO2019/195268A2.

En algunas de tales realizaciones, cada bin de la pluralidad de bins de la presente descripción se mapea en una sola región genómica correspondiente única de cualquiera de las tablas 2-10 o 16-24 del documento WO2019/195268A2. En algunas de tales realizaciones, un bin de la pluralidad de bins se mapea en una, dos, tres, cuatro, cinco, seis, siete, ocho, nueve o diez regiones genómicas correspondientes únicas en cualquier combinación de las tablas 2-10 o 16-24 del documento WO2019/195268A2.

En algunas realizaciones, la totalidad de la pluralidad de bins de la presente descripción en conjunto se configura para mapearse en al menos el 30 %, el 40 %, el 50 %, el 60 %, el 70 %, el 80 %, el 90 % o el 95 % de las regiones genómicas de las tablas 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23 y/o 24 del documento WO2019/195268A2.

Ejemplo 10: Protocolo para obtener información de metilación a partir de lecturas de secuencia de fragmentos en una muestra biológica.

La Figura 19 es un diagrama de flujo que describe un proceso 1900 de secuenciación de fragmentos, según una realización de la presente descripción.

En el paso 1902, el ácido nucleico (p. ej., ADN o ARN) se extrae de una muestra biológica correspondiente de un sujeto respectivo. En la presente descripción, el ADN y el ARN pueden usarse indistintamente a menos que se indique lo contrario. Es decir, las realizaciones descritas en la presente memoria pueden ser aplicables para secuencias de ácidos nucleicos de tipo ADN y ARN. Sin embargo, para mayor claridad y explicación, los ejemplos descritos en la presente memoria pueden centrarse en el ADN. La muestra biológica puede incluir moléculas de ácido nucleico obtenidas de cualquier subconjunto del genoma humano, incluido el genoma completo. La muestra biológica puede incluir sangre, plasma, suero, orina, materia fecal, saliva, otros tipos de fluidos corporales o cualquier combinación de los mismos. En algunas realizaciones, los métodos para extraer una muestra de sangre (p. ej., con jeringa o pinchazo en el dedo) pueden ser menos invasivos que los procedimientos para obtener una biopsia de tejido, que pueden requerir cirugía. La muestra extraída puede comprender ADNlc y/o ADNtc. Si un sujeto presenta un estado patológico, tal como cáncer, los ácidos nucleicos libres de células (p. ej., ADNlc) en una muestra extraída del sujeto generalmente incluyen un nivel detectable de los ácidos nucleicos que pueden usarse para evaluar un estado patológico.

En el paso 1904, los ácidos nucleicos extraídos (p. ej., incluidos los fragmentos de ADNlc) se tratan para convertir citosinas no metiladas en uracilos. En algunas realizaciones, el método 1900 usa un tratamiento con bisulfito de las muestras que convierte las citosinas no metiladas en uracilos sin convertir las citosinas metiladas. Por ejemplo, para la conversión con bisulfito se utiliza un kit comercial tal como el kit EZ DNA Methylation™-Gold, EZ DNA Methylation™-Direct o EZ DNA Methylation™-Lightning (comercializado por Zymo Research Corp [Irvine, CA]). En otra realización, la conversión de citosinas no metiladas en uracilos se logra mediante el uso de una reacción enzimática. Por ejemplo, la conversión puede utilizar un kit disponible comercialmente para la conversión de citosinas no metiladas en uracilos, tal como APOBEC-Seq (NEBiolabs, Ipswich, MA).

En el paso 1906, se prepara una biblioteca de secuenciación. En algunas realizaciones, la preparación incluye al menos dos pasos. En un primer paso, se añade un adaptador de ADNmc al extremo 3'-OH de una molécula de ADNmc convertida con bisulfito utilizando una reacción de ligación de ADNmc. En algunas realizaciones, la reacción de ligación de ADNmc usa Circligase II (Epicentre) para ligar el adaptador de ADNmc al extremo 3'-OH de una molécula de ADNmc convertida con bisulfito, en donde el extremo 5' del adaptador está fosforilado y el ADNmc convertido con bisulfito se ha desfosforilado (es decir, el extremo 3' tiene un grupo hidroxilo). En otra realización, la reacción de ligación de ADNmc usa la ligasa termoestable 5' AppDNA/RNA (comercializada por 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 realización, la reacción de ligación de ADNmc usa una ligasa de ARN T4 (comercializada por 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 utiliza en una reacción de extensión de cebador para formar una molécula de ADN bicatenario convertida con bisulfito. Opcionalmente, en algunas realizaciones, 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. A continuación, el ADN bicatenario convertido con bisulfito se amplifica para añadir adaptadores de secuenciación. Por ejemplo, la amplificación por PCR utilizando un cebador directo que incluye una secuencia P5 y un cebador inverso que incluye una secuencia P7 se utiliza 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, por sus siglas en inglés) a las moléculas de ácido nucleico (p. ej., moléculas de ADN) mediante ligación del adaptador. Los UMI son secuencias cortas de ácido nucleico (p. ej., 4-10 pares de bases) que se añaden a los extremos de los fragmentos de ADN durante el ligación del adaptador. En algunas realizaciones, los 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 ligación 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 un paso opcional, los ácidos nucleicos (p. ej., fragmentos) pueden hibridarse. Las sondas de hibridación (también denominadas en la presente memoria “sondas”) pueden usarse para seleccionar como objetivo y extraer fragmentos de ácido nucleico que proporcionen información sobre estados patológicos. Para un flujo de trabajo determinado, las sondas pueden diseñarse para hibridarse con una cadena objetivo (complementaria) de ADN o ARN. La cadena objetivo 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 c, 100 c o 1000 c de pares de bases. Además, las sondas pueden cubrir porciones superpuestas de una región objetivo.

En una etapa opcional, los fragmentos de ácido nucleico hibridados se capturan y pueden enriquecerse, p. ej., amplificarse mediante PCR. En algunas realizaciones, las secuencias de ADN objetivo pueden enriquecerse a partir de la biblioteca. Esto se usa, por ejemplo, donde se realiza un ensayo de panel objetivo en las muestras. 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, biotinilarse) para facilitar el aislamiento de ácidos nucleicos objetivo hibridados con sondas utilizando una superficie recubierta con estreptavidina (p. ej., perlas recubiertas con estreptavidina).

En el paso, se generan lecturas de secuencia a partir de la muestra de ácido nucleico, p. ej., secuencias enriquecidas. Los datos de secuenciación se pueden adquirir a partir de las secuencias de ADN enriquecidas mediante medios conocidos en la técnica. Por ejemplo, el método puede incluir técnicas de secuenciación de próxima generación (NGS, por sus siglas en inglés) 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 ligación (secuenciación SOLiD), secuenciación de nanoporos (Oxford Nanopore Technologies) o secuenciación de extremos emparejados. En algunas realizaciones, la secuenciación masivamente paralela se realiza mediante secuenciación por síntesis con terminadores de colorante reversibles.

En el paso, un procesador de secuencias puede generar información de metilación utilizando las lecturas de secuencia. A continuación, se puede generar un vector de estado de metilación utilizando la información de metilación determinada a partir de las lecturas de secuencia. La Figura 20 es una ilustración del proceso de secuenciación de una molécula de ADNlc para obtener un vector de estado de metilación, según una realización. Como ejemplo, se recibió un fragmento de ADNlc en 2012 que, en este ejemplo, contiene tres sitios CpG. Como se muestra, el primer y tercer sitios de CpG del fragmento (molécula) 2012 de ADNlc están metilados 2014. Durante el paso de tratamiento, la molécula 2012 de ADNlc se convierte para generar una molécula 2022 de ADNlc convertida. Durante el tratamiento 2015, el segundo sitio CpG que no estaba metilado tenía su citosina convertida en uracilo. Sin embargo, el primer y el tercer sitio de CpG no se convierten.

Después de la conversión, se prepara una biblioteca de secuenciación generando una lectura de secuencia. La lectura de secuencia se alinea con un genoma de referencia. El genoma de referencia proporciona el contexto de la posición en un genoma humano de la que proviene el fragmento ADNlc. En este ejemplo simplificado, el sistema de análisis alinea la lectura de secuencia de modo que los tres sitios CpG se correlacionan con los sitios CpG 23, 24 y 25 (identificadores de referencia arbitrarios usados por

comodidad de uso de la descripción). Por lo tanto, los sistemas y métodos descritos generan información tanto sobre el estado de metilación de todos los sitios CpG en el fragmento (molécula) 2012 de ADNlc como sobre la posición en el genoma humano en la que se mapean los sitios CpG. Como se muestra, los sitios CpG en la lectura 2042 de secuencia que estaban metilados se leen como citosinas. En este ejemplo, las citosinas aparecen en la lectura 2042 de secuencia solo en el primer y el tercer sitio CpG, lo que permite inferir que el primer y el tercer sitio CpG en la molécula de ADNlc original estaban metilados. Mientras que 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 la molécula de ADNlc original. Con estos dos datos, el estado y la ubicación de metilación, los sistemas y métodos descritos generan un vector 2052 de estado de metilación para el fragmento ADNlc 2012. En este ejemplo, el vector 2052 de estado de metilación resultante es $\langle M_{23}, U_{24}, M_{25} \rangle$, 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.

CONCLUSIÓN

Se pueden proporcionar varias instancias para los componentes, operaciones o estructuras descritas en la presente memoria como una única instancia. Finalmente, los límites entre varios componentes, operaciones y almacenes de datos son algo arbitrarios, y las operaciones particulares se ilustran en el contexto de configuraciones ilustrativas específicas. Se prevén otras asignaciones de funcionalidad que pueden estar dentro del alcance de la(s) implementación(es). Generalmente, las estructuras y la funcionalidad presentadas como componentes separados en las configuraciones de ejemplo pueden implementarse como una estructura o un componente combinados. Similarmente, las estructuras y la funcionalidad presentadas como un único componente pueden implementarse como componentes separados. Estas y otras variaciones, modificaciones, adiciones y mejoras entran dentro del alcance de la(s) implementación(es).

También se entenderá que, aunque los términos primero, segundo, etc. pueden usarse en la presente memoria para describir varios elementos, estos elementos no deben limitarse a esos términos. Estos términos solo se utilizan para distinguir un elemento de otro. Por ejemplo, un primer sujeto podría denominarse segundo sujeto y, similarmente, un segundo sujeto podría denominarse primer sujeto, sin apartarse del alcance de la presente descripción. El primer sujeto y el segundo sujeto son ambos sujetos, pero no son el mismo sujeto.

La terminología usada en la presente descripción tiene el propósito de describir únicamente realizaciones particulares y no pretende limitar la invención. Como se utiliza en la descripción de la invención y en las reivindicaciones adjuntas, las formas singulares “un”, “una” y “el/la” pretenden incluir también las formas plurales, a menos que el contexto indique claramente lo contrario. También se entenderá que el término “y/o”, tal como se utiliza en la presente memoria, se refiere a y abarca todas y cada una de las combinaciones posibles de uno o más de los artículos listados asociados. Se entenderá además que los términos “comprende” y/o “que comprende”, cuando se usan en esta especificación, especifican la presencia de las características, enteros, pasos, operaciones, elementos y/o componentes establecidos, pero no excluyen la presencia o adición de una o más características, enteros, pasos, operaciones, elementos, componentes y/o grupos de los mismos.

Como se utiliza en la presente memoria, el término “si” puede interpretarse en el sentido de “cuándo” o “tras” o “en respuesta a la determinación” o “en respuesta a la detección”, según el contexto. Similarmente, la frase “si se determina” o “si se detecta [una condición o evento declarado]” puede interpretarse en el sentido de “al determinar” o “en respuesta a la determinación” o “al detectar (la condición o evento declarado)” o “en respuesta a la detección (la condición o evento declarado)”, según el contexto.

La descripción anterior incluyó ejemplos de sistemas, métodos, técnicas, secuencias de instrucciones y productos de programas de máquinas informáticas que incorporan implementaciones ilustrativas. Con fines explicativos, se expusieron numerosos detalles específicos para proporcionar una comprensión de las diversas implementaciones del objeto de la invención. Sin embargo, resultará evidente para los expertos en la técnica que las implementaciones del objeto de la invención pueden ponerse en práctica sin estos detalles específicos. En general, no se han mostrado en detalle instancias de instrucciones, protocolos, estructuras y técnicas conocidas.

La descripción anterior, con fines explicativos, se ha descrito con referencia a implementaciones específicas. Sin embargo, los análisis ilustrativos anteriores no pretenden ser exhaustivos ni limitar las implementaciones a las formas precisas descritas. Son posibles muchas modificaciones y variaciones en vista de las enseñanzas anteriores. Las implementaciones se eligieron y describieron para explicar mejor los principios y sus aplicaciones prácticas, para permitir de este modo a otros expertos en la técnica utilizar mejor las implementaciones y las diversas implementaciones con diversas modificaciones que sean adecuadas para el uso particular contemplado.

REIVINDICACIONES

1. Un sistema informático para entrenar a uno o más clasificadores para discriminar entre cada condición de cáncer de un conjunto de condiciones de cáncer, comprendiendo el conjunto de condiciones de cáncer dos o más condiciones de cáncer, comprendiendo el sistema informático:
 - al menos un procesador; y
 - una memoria que almacena al menos un programa para la ejecución por parte del al menos un procesador, comprendiendo el al menos un programa instrucciones para:
 - (A) proporcionar una primera pluralidad de funciones de extracción de características basadas en un primer conjunto de datos, en donde
 - el primer conjunto de datos comprende, para cada sujeto respectivo de una primera pluralidad de sujetos de una especie, la primera información de genotipo correspondiente que comprende (i) una primera pluralidad correspondiente de valores de bin, cada valor de bin respectivo de la primera pluralidad correspondiente de valores de bin para un bin correspondiente de una pluralidad de bins y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer, en donde cada bin de la pluralidad de bins representa una porción de un genoma de referencia de la especie, y en donde la primera pluralidad correspondiente de valores de bin de cada sujeto respectivo de la primera pluralidad de sujetos se obtiene de una muestra biológica correspondiente del sujeto respectivo, que comprende un primer tipo de tejido, utilizando una primera pluralidad correspondiente de lecturas de secuencia determinadas mediante un primer método de secuenciación de ácido nucleico, y
 - la primera pluralidad de funciones de extracción de características se obtiene aplicando una primera técnica de extracción de características a los valores de bin respectivos de los sujetos respectivos del primer conjunto de datos, identificando de este modo la primera pluralidad de funciones de extracción de características,
 - en donde cada función de extracción de características de la primera pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal de los valores de bin de la totalidad o un subconjunto de la pluralidad de bins, y
 - la primera pluralidad de funciones de extracción de características discrimina colectivamente a los sujetos respectivos de la primera pluralidad de sujetos según tengan una condición de cáncer dentro del conjunto de condiciones de cáncer basándose en los valores de bin respectivos para los sujetos respectivos;
 - caracterizándose el al menos un programa por comprender además instrucciones para:
 - (B) obtener un segundo conjunto de datos que comprende, para cada sujeto respectivo de una segunda pluralidad de sujetos de la especie, la segunda información de genotipo correspondiente que comprende (i) una segunda pluralidad correspondiente de valores de bin, representando cada valor de bin respectivo de la segunda pluralidad correspondiente de valores de bin un bin correspondiente en la pluralidad de bins y (ii) una indicación de una condición de cáncer del sujeto respectivo del conjunto de condiciones de cáncer, y en donde la correspondiente segunda pluralidad de valores de bin de cada sujeto respectivo de la segunda pluralidad de sujetos se obtiene de una muestra biológica correspondiente del sujeto respectivo, que comprende un segundo tipo de tejido, utilizando una segunda pluralidad correspondiente de lecturas de secuencia determinadas mediante un segundo método de secuenciación de ácido nucleico, y en donde
 - al menos el primer método de secuenciación de ácido nucleico difiere del segundo método de secuenciación de ácido nucleico o el primer tipo de tejido difiere del segundo tipo de tejido;
 - (C) transformar, basándose en al menos cada función de extracción de características respectiva de la primera pluralidad de funciones de extracción de características, la segunda pluralidad respectiva de valores de bin de cada sujeto correspondiente de la segunda pluralidad de sujetos con respecto a la función de extracción de características respectiva, contribuyendo de este modo a un segundo conjunto de datos transformado con una pluralidad respectiva de valores de características para cada sujeto correspondiente; y
 - (D) utilizar el segundo conjunto de datos transformado y la indicación de la condición de cáncer de los sujetos respectivos de la segunda pluralidad de sujetos para entrenar a un primer clasificador del uno o más clasificadores en al menos un par de condiciones de cáncer del conjunto de condiciones de cáncer.
 2. El sistema informático de la reivindicación 1, en donde cada valor de bin respectivo de la primera pluralidad de valores de bin o la segunda pluralidad de valores de bin representa un número de lecturas de secuencia medidas a partir de ácidos nucleicos libres de células en la muestra biológica correspondiente que se asocian con el bin respectivo.

3. El sistema informático de una cualquiera de las reivindicaciones 1-2, en donde cada valor de bin respectivo de la primera pluralidad de valores de bin o la segunda pluralidad de valores de bin representa un número de lecturas de secuencia asociadas con el bin respectivo de la pluralidad de bins o un patrón de metilación respectivo medido en la muestra biológica correspondiente para el bin respectivo de la pluralidad de bins.

4. El sistema informático de una cualquiera de las reivindicaciones 1-3, en donde la transformación (C) se basa en cada función de extracción de características respectiva de la primera pluralidad de funciones de extracción de características y una segunda pluralidad de funciones de extracción de características, en donde la segunda pluralidad de funciones de extracción de características se obtiene mediante:

la aplicación de una segunda técnica de extracción de características a los valores de bin respectivos de los sujetos respectivos del segundo conjunto de datos, identificando de este modo la segunda pluralidad de funciones de extracción de características; en donde

cada función de extracción de características de la segunda pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal de los valores de bin de la totalidad o un subconjunto de la pluralidad de bins, y

la segunda pluralidad de funciones de extracción de características discrimina colectivamente a los sujetos respectivos de la segunda pluralidad de sujetos según tengan una condición de cáncer dentro del conjunto de condiciones de cáncer basándose en los valores de bin respectivos para los sujetos respectivos.

5. El sistema informático de una cualquiera de las reivindicaciones 1-4, en donde

el conjunto de condiciones de cáncer comprende tres o más condiciones de cáncer,

para cada condición de cáncer respectiva del conjunto de condiciones de cáncer, hay dos o más sujetos de la primera pluralidad de sujetos que tienen la condición de cáncer respectiva, y

para cada par respectivo de condiciones de cáncer del conjunto de condiciones de cáncer, la aplicación de la primera técnica de extracción de características se realiza aplicando una instancia de la primera técnica de extracción de características a los valores de bin respectivos de los sujetos respectivos del primer conjunto de datos que tienen una primera condición de cáncer o una segunda condición de cáncer del par respectivo de condiciones de cáncer, en donde cada instancia de la primera técnica de extracción de características contribuye con un subconjunto correspondiente de la primera pluralidad de funciones de extracción de características a la primera pluralidad de funciones de extracción de características.

6. El sistema informático de la reivindicación 5, en donde el al menos un programa comprende además instrucciones para:

aplicar una segunda técnica de extracción de características a los valores de bin respectivos de los sujetos respectivos del segundo conjunto de datos, identificando de este modo una segunda pluralidad de funciones de extracción de características; en donde

cada función de extracción de características de la segunda pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal de los valores de bin de la totalidad o un subconjunto de la pluralidad de bins, y

la segunda pluralidad de funciones de extracción de características discrimina colectivamente a los sujetos respectivos de la segunda pluralidad de sujetos según tengan una condición de cáncer dentro del conjunto de condiciones de cáncer basándose en los valores de bin respectivos para los sujetos respectivos; y aplicar la segunda pluralidad de funciones de extracción de características a los valores de bin respectivos de los sujetos respectivos del segundo conjunto de datos para obtener una pluralidad respectiva de segundos valores de características para cada sujeto correspondiente del segundo conjunto de datos, y

en donde el segundo conjunto de datos transformado comprende además la pluralidad respectiva de segundos valores de característica.

7. El sistema informático de una cualquiera de las reivindicaciones 5-6, en donde

la aplicación de la primera técnica de extracción de características comprende aplicar un algoritmo de reducción de dimensiones a los sujetos del primer conjunto de datos que tienen una primera condición de cáncer o una segunda condición de cáncer del par respectivo de condiciones de cáncer, identificando de este modo el subconjunto correspondiente de la primera pluralidad de funciones de extracción de características.

8. El sistema informático de la reivindicación 7, en donde

la aplicación de la segunda técnica de extracción de características comprende aplicar el algoritmo de reducción de dimensiones a los sujetos del segundo conjunto de datos que tienen una primera condición de cáncer o una segunda condición de cáncer del par respectivo de condiciones de cáncer, identificando de este modo el subconjunto correspondiente de la segunda pluralidad de funciones de extracción de características.

9. El sistema informático de una cualquiera de las reivindicaciones 7-8, en donde el algoritmo de reducción de dimensiones se selecciona del grupo que consiste en un algoritmo de análisis de componentes principales, un algoritmo de análisis factorial, un mapeo de Sammon, un análisis de componentes curvilíneos, un algoritmo de incrustación de vecinos estocásticos (SNE), un algoritmo de Isomap, un algoritmo de despliegue de varianza máxima, un algoritmo de incrustación lineal local, un algoritmo de t-SNE, un algoritmo de factorización matricial no negativa, un algoritmo de análisis de componentes principales de kernel, un algoritmo de análisis de componentes principales de kernel basado en gráficos, un algoritmo de análisis discriminante lineal, un algoritmo de análisis discriminante generalizado, un algoritmo de aproximación y proyección de variedad uniforme (UMAP), un algoritmo LargeVIS, un algoritmo Laplacian Eigenmap y un algoritmo de análisis discriminante lineal de Fisher.

10. El sistema informático de una cualquiera de las reivindicaciones 1-9, en donde cada función de extracción de características de la primera pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal del valor de bin de la totalidad o un subconjunto de la pluralidad de bins en forma de un peso independiente para cada bin respectivo de la pluralidad de bins o el subconjunto de la pluralidad de bins.

11. El sistema informático de una cualquiera de las reivindicaciones 4, 6, 7 o 10, en donde cada función de extracción de características de la segunda pluralidad de funciones de extracción de características codifica independientemente una función lineal o no lineal del valor de bin de la totalidad o un subconjunto de la pluralidad de bins en forma de un peso independiente para cada bin respectivo de la pluralidad de bins o el subconjunto de la pluralidad de bins.

12. El sistema informático de la reivindicación 1, en donde

la primera técnica de extracción de características es una primera red neuronal convolucional que comprende una primera pluralidad de capas convolucionales,

cada capa convolucional respectiva en la primera pluralidad de capas convolucionales se asocia con un vector de peso aprendido que se obtiene mediante la retropropagación en la primera red neuronal convolucional utilizando los valores de bin respectivos y las indicaciones respectivas de la condición de cáncer de los sujetos respectivos del primer conjunto de datos, y

el vector de peso aprendido respectivo de cada capa convolucional de un subconjunto de la primera pluralidad de capas convolucionales representa colectivamente la primera pluralidad de funciones de extracción de características, y

la transformación (C) comprende introducir la segunda pluralidad correspondiente de valores de bin de un sujeto respectivo de la segunda pluralidad de sujetos en una segunda red convolucional que comprende el subconjunto de la primera pluralidad de capas convolucionales, en donde un vector de peso de cada capa convolucional respectiva de la segunda red neuronal convolucional se inicializa y, en algunas realizaciones, se congela en valores del vector de peso aprendido de la capa convolucional correspondiente en la primera red neuronal convolucional.

13. El sistema informático de una cualquiera de las reivindicaciones 1-12, en donde el al menos un programa comprende además:

(i) instrucciones para escalar un primer valor de bin respectivo para cada bin respectivo de la pluralidad de bins para cada sujeto respectivo de la primera pluralidad de sujetos mediante:

la realización de una transformación logarítmica del primer valor de bin respectivo, formando de este modo un primer valor de bin transformado logarítmicamente para el bin respectivo,

la resta de un valor medio del primer valor de bin transformado logarítmicamente respectivo de la primera pluralidad de sujetos del primer valor de bin transformado logarítmicamente del bin respectivo, formando de este modo un primer valor de bin normalizado para el bin respectivo, y

la división posterior del primer valor de bin normalizado respectivo para el bin respectivo por una desviación estándar del primer valor de bin normalizado en la primera pluralidad de sujetos, escalando de este modo el

primer valor de bin para cada bin respectivo de la pluralidad de bins para cada sujeto respectivo de la primera pluralidad de sujetos;

y/o

(ii) instrucciones para escalar un segundo valor de bin respectivo para cada bin respectivo de la pluralidad de bins para cada sujeto respectivo de la segunda pluralidad de sujetos mediante:

la realización de una transformación logarítmica del segundo valor de bin respectivo, formando de este modo un segundo valor de bin transformado logarítmicamente para el bin respectivo,

la resta de un valor medio del segundo valor de bin transformado logarítmicamente respectivo en la segunda pluralidad de sujetos del segundo valor de bin transformado logarítmicamente del bin respectivo, formando de este modo un segundo valor de bin normalizado para el bin respectivo, y

la división posterior del segundo valor de bin normalizado respectivo para el bin respectivo por una desviación estándar del segundo valor de bin normalizado en la segunda pluralidad de sujetos, escalando de este modo el segundo valor de bin para cada bin respectivo de la pluralidad de bins para cada sujeto respectivo de la segunda pluralidad de sujetos.

14. El sistema informático de una cualquiera de las reivindicaciones 1-13, en donde

cada lectura de secuencia respectiva usada para formar la primera pluralidad de valores de bin de un sujeto respectivo de la primera pluralidad de sujetos incluye (i) una primera porción que puede mapearse en el genoma de la especie y (ii) una segunda porción, y

las lecturas de secuencia usadas para formar la primera pluralidad de valores de bin de un sujeto respectivo de la primera pluralidad de sujetos se filtran de modo que solo las lecturas de secuencia cuya primera porción es inferior a 160 nucleótidos se usan para formar los valores de bin.

15. El sistema informático de la reivindicación 1, en donde el primer método de secuenciación de ácido nucleico o el segundo método de secuenciación de ácido nucleico es la secuenciación del genoma completo, la secuenciación por panel dirigida o la secuenciación con bisulfito del genoma completo.

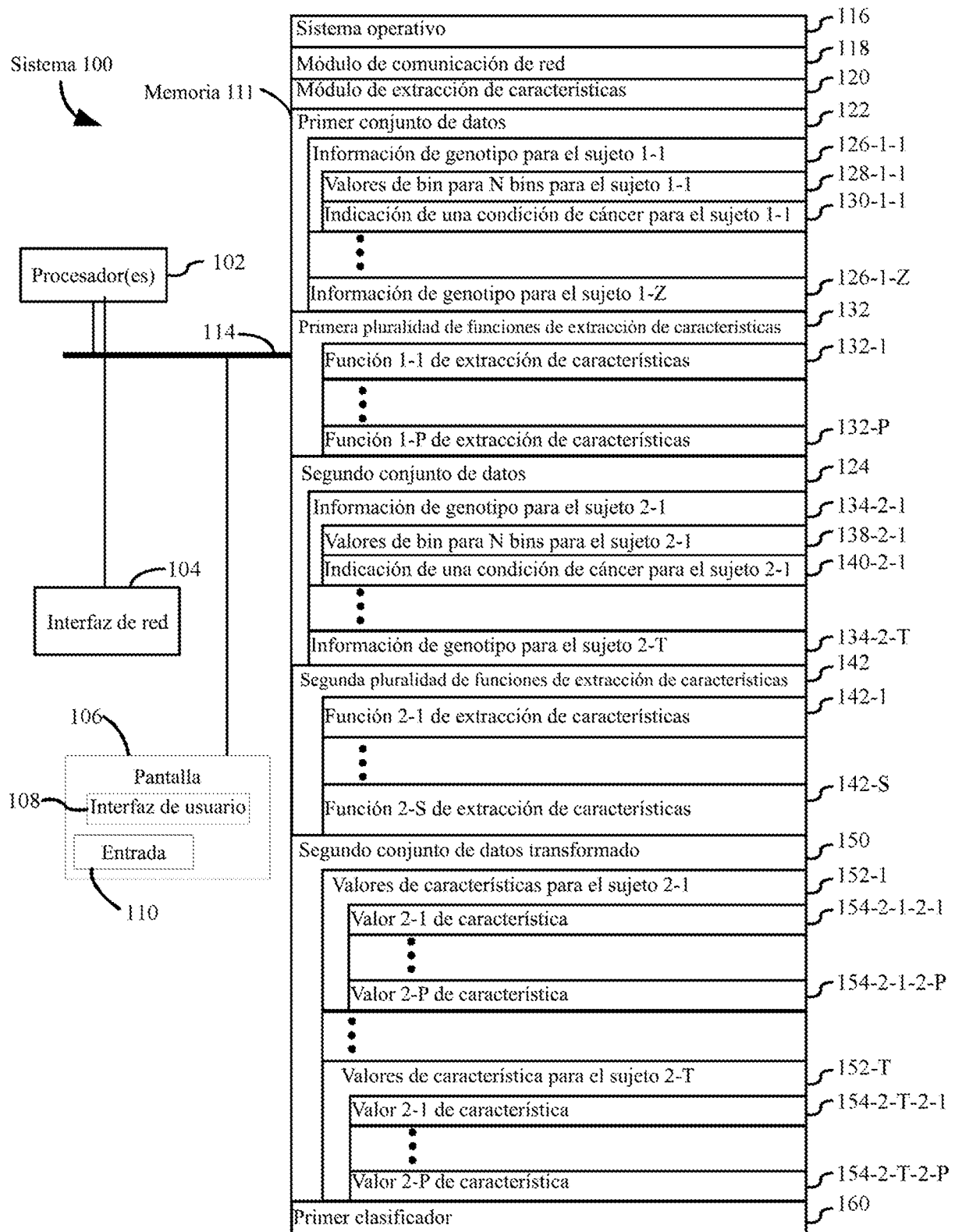


Figura 1

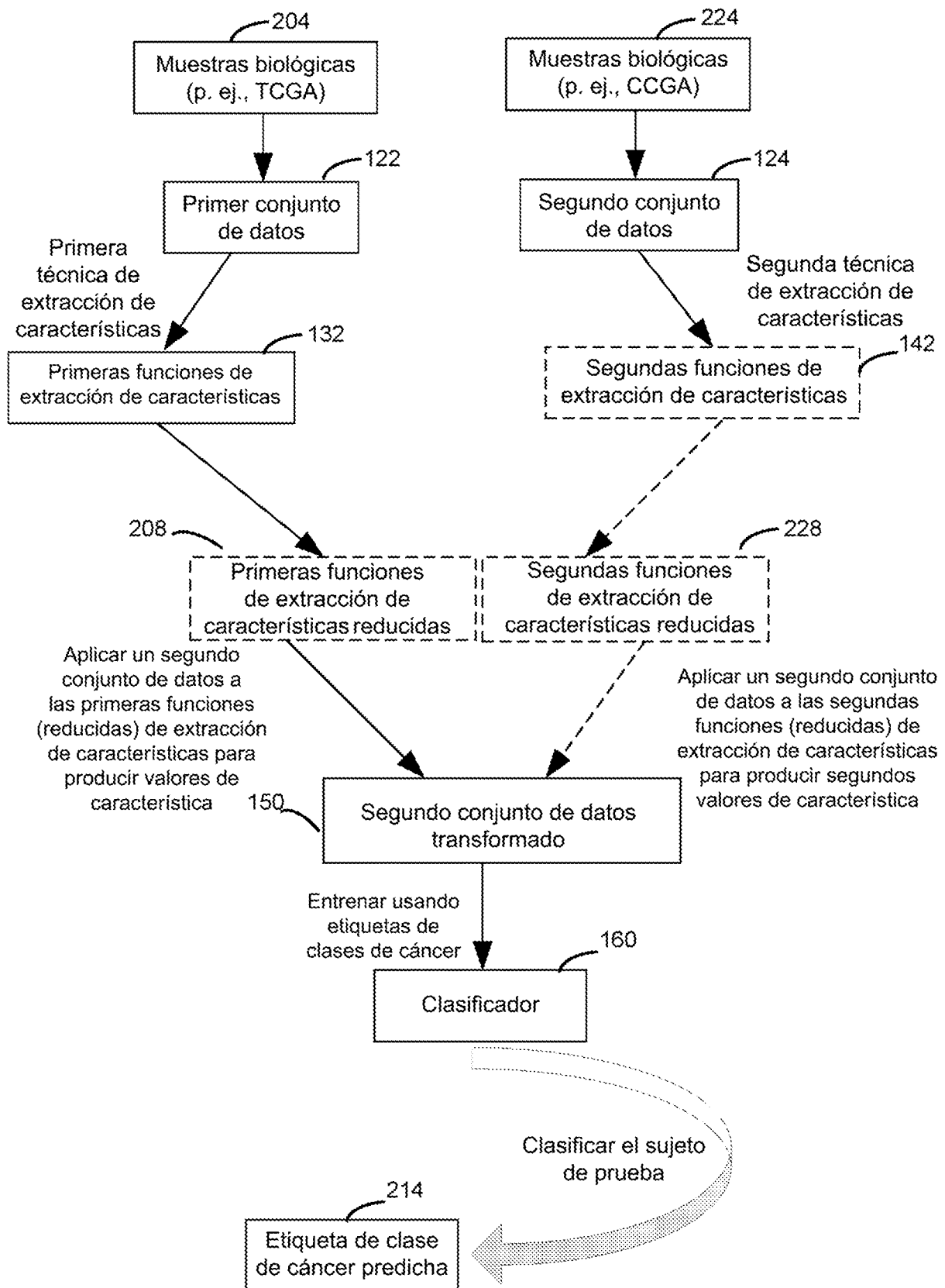


Figura 2

Primer conjunto de datos (p. ej., TCGA)

		Bins				
Sujetos	Etiqueta de clase	Valor de bin 1	Valor de bin 2	Valor de bin 3	...	Valor de bin N
	Sujeto 1-1	L_{1-1}	cnt_{1-1-2}	cnt_{1-1-3}	...	cnt_{1-1-N}
	Sujeto 1-2	L_{1-2}	cnt_{1-2-2}	cnt_{1-2-3}	...	cnt_{1-2-N}
	...	...	...	...	...	...
	Sujeto 1-Z	L_{1-Z}	cnt_{1-Z-2}	cnt_{1-Z-3}	...	cnt_{1-Z-N}

Segundo conjunto de datos (p. ej., CCGA)

		Bins				
Sujetos	Etiqueta de clase	Valor de bin 1	Valor de bin 2	Valor de bin 3	...	Valor de bin N
	Sujeto 2-1	L_{2-1}	cnt_{2-1-2}	cnt_{2-1-3}	...	cnt_{2-1-N}
	Sujeto 2-2	L_{2-2}	cnt_{2-2-2}	cnt_{2-2-3}	...	cnt_{2-2-N}
	...	...	...	...	...	...
	Sujeto 2-T	L_{2-T}	cnt_{2-T-2}	cnt_{2-T-3}	...	cnt_{2-T-N}

Figura 3

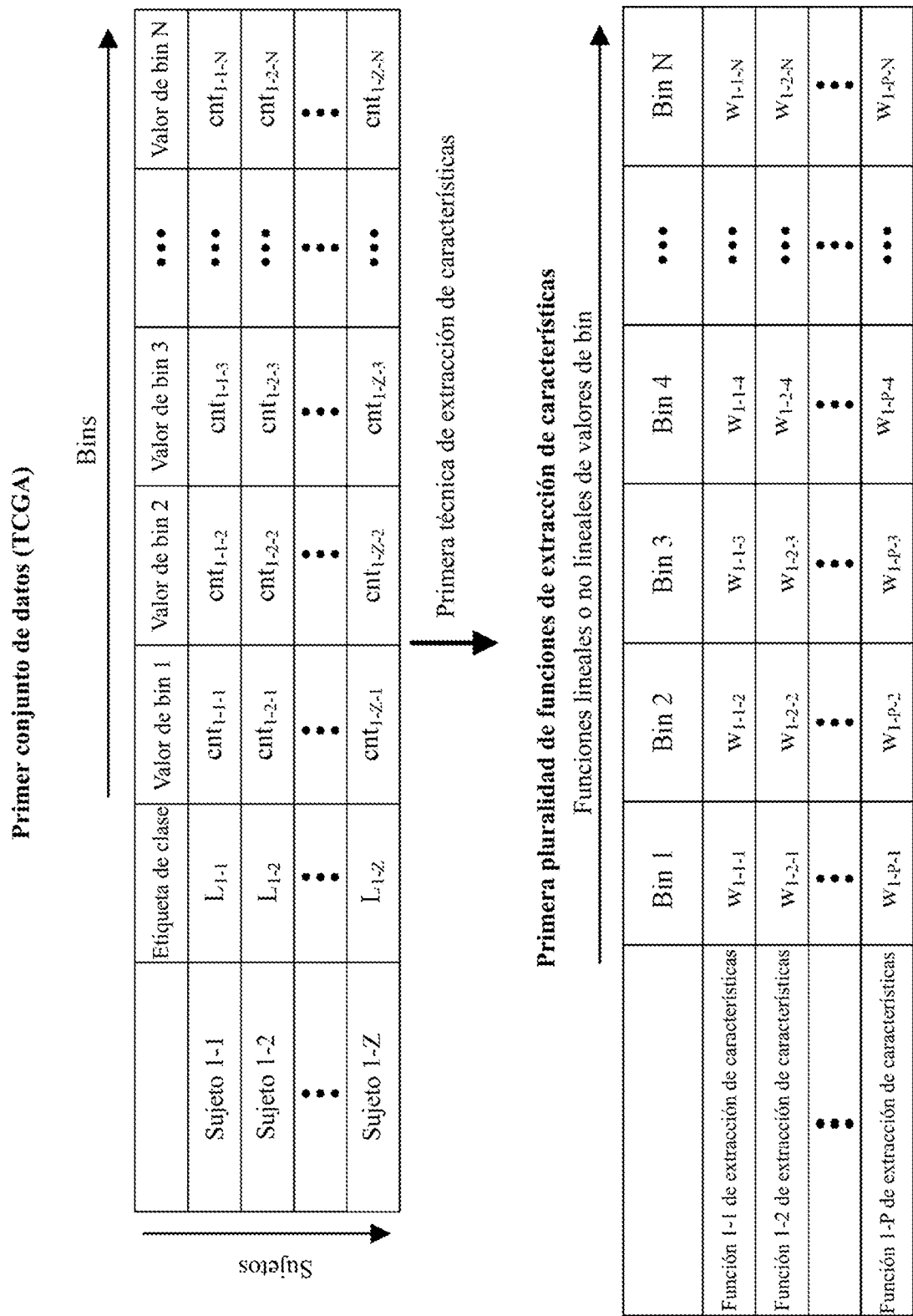


Figura 4

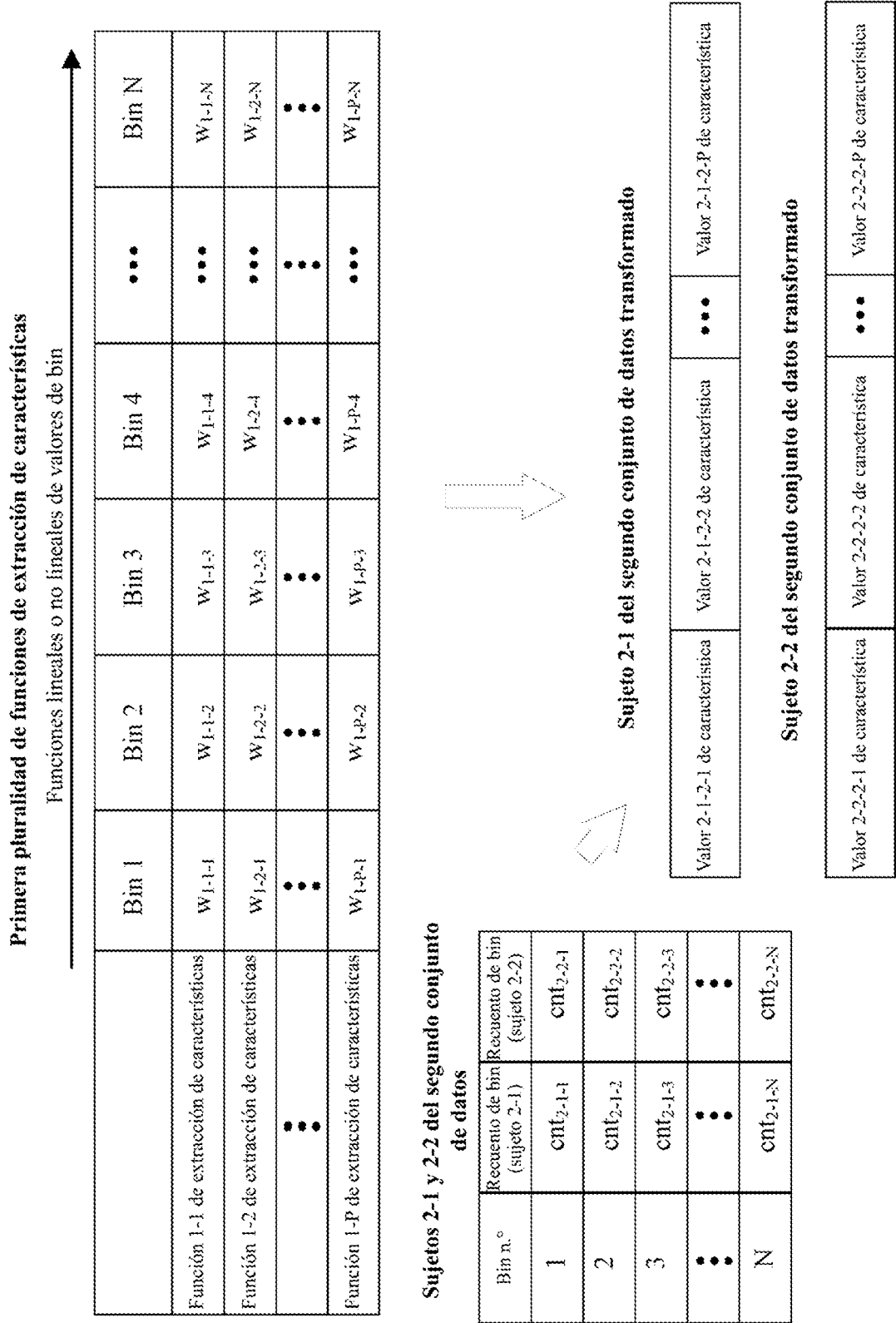


Figura 5

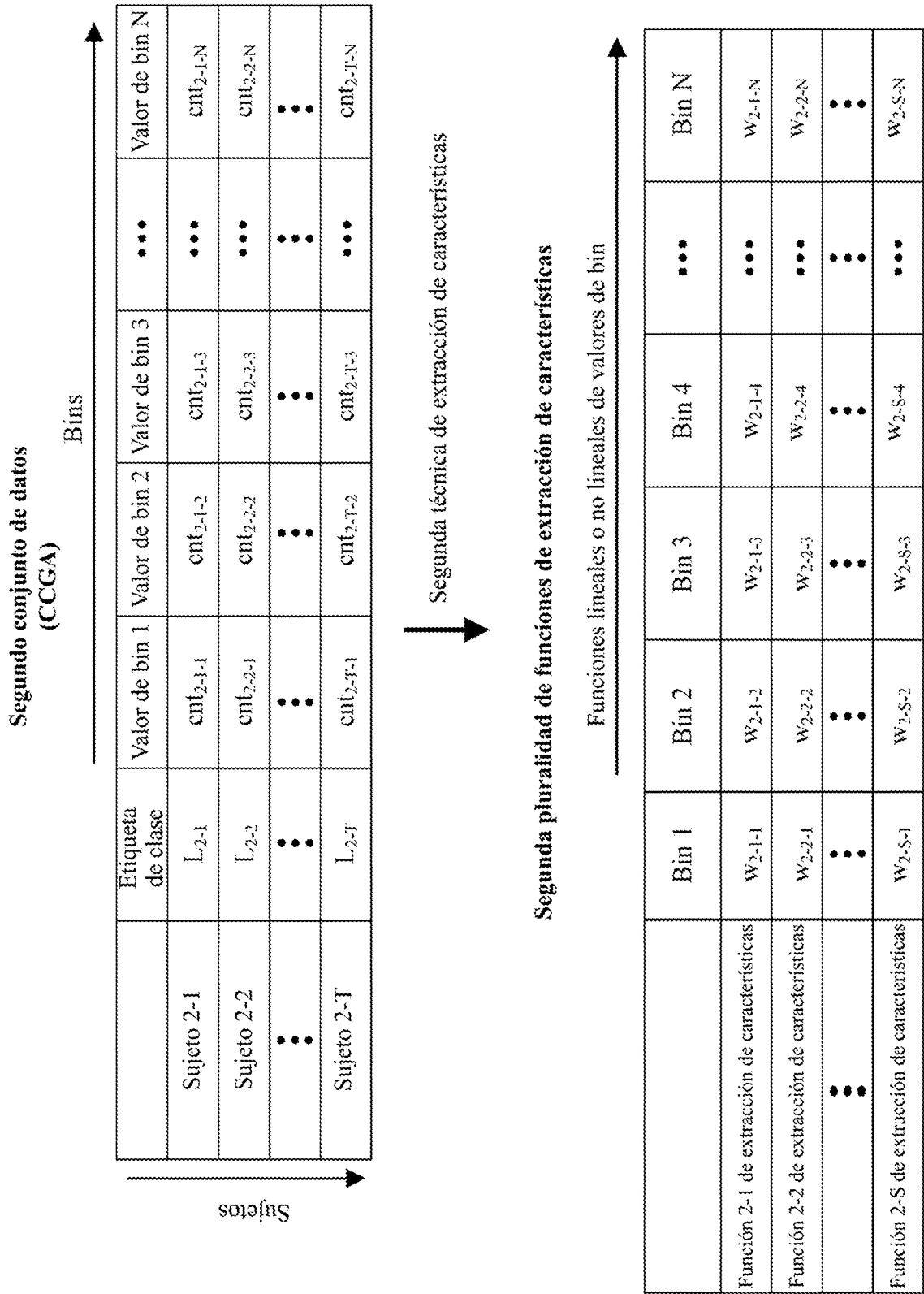


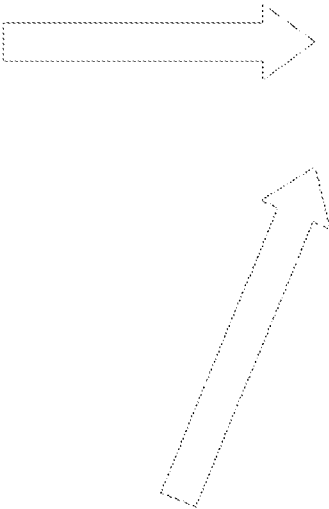
Figura 6

Segunda pluralidad de funciones de extracción de características

Funciones lineales o no lineales de valores de bin						
	Bin 1	Bin 2	Bin 3	Bin 4	...	Bin N
Función 2-1 de extracción de características	w_{2-1-1}	w_{2-1-2}	w_{2-1-3}	w_{2-1-4}	...	w_{2-1-N}
Función 2-2 de extracción de características	w_{2-2-1}	w_{2-2-2}	w_{2-2-3}	w_{2-2-4}	...	w_{2-2-N}
...	...	...	...	...	...	...
Función 2-S de extracción de características	w_{2-S-1}	w_{2-S-2}	w_{2-S-3}	w_{2-S-4}	...	w_{2-S-N}

Sujeto 2-1 del segundo conjunto de datos

Bin n.º	Recuento de bin (sujeto 2-1)
1	cnt_{2-1-1}
2	cnt_{2-1-2}
3	cnt_{2-1-3}
...	...
N	cnt_{2-1-N}



Sujeto 2-1 del segundo conjunto de datos transformado y sujeto 2-1

Valor 2-1-2-1 de característica	...	Valor 2-1-2-P de característica	Segundo valor 2-1-2'-1 de característica	...	Segundo valor 2-1-2'-S de característica
---------------------------------	-----	---------------------------------	--	-----	--

Figura 7

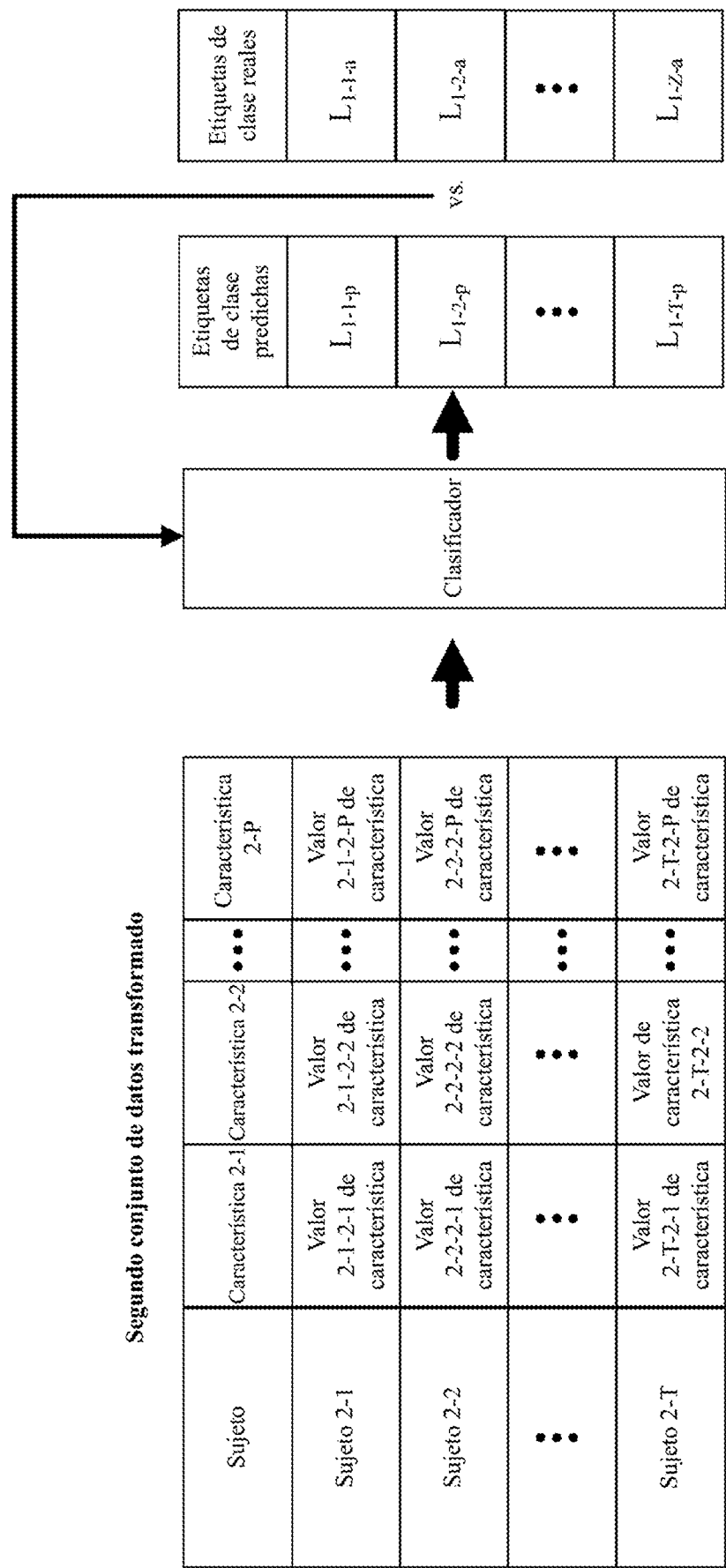


Figura 8

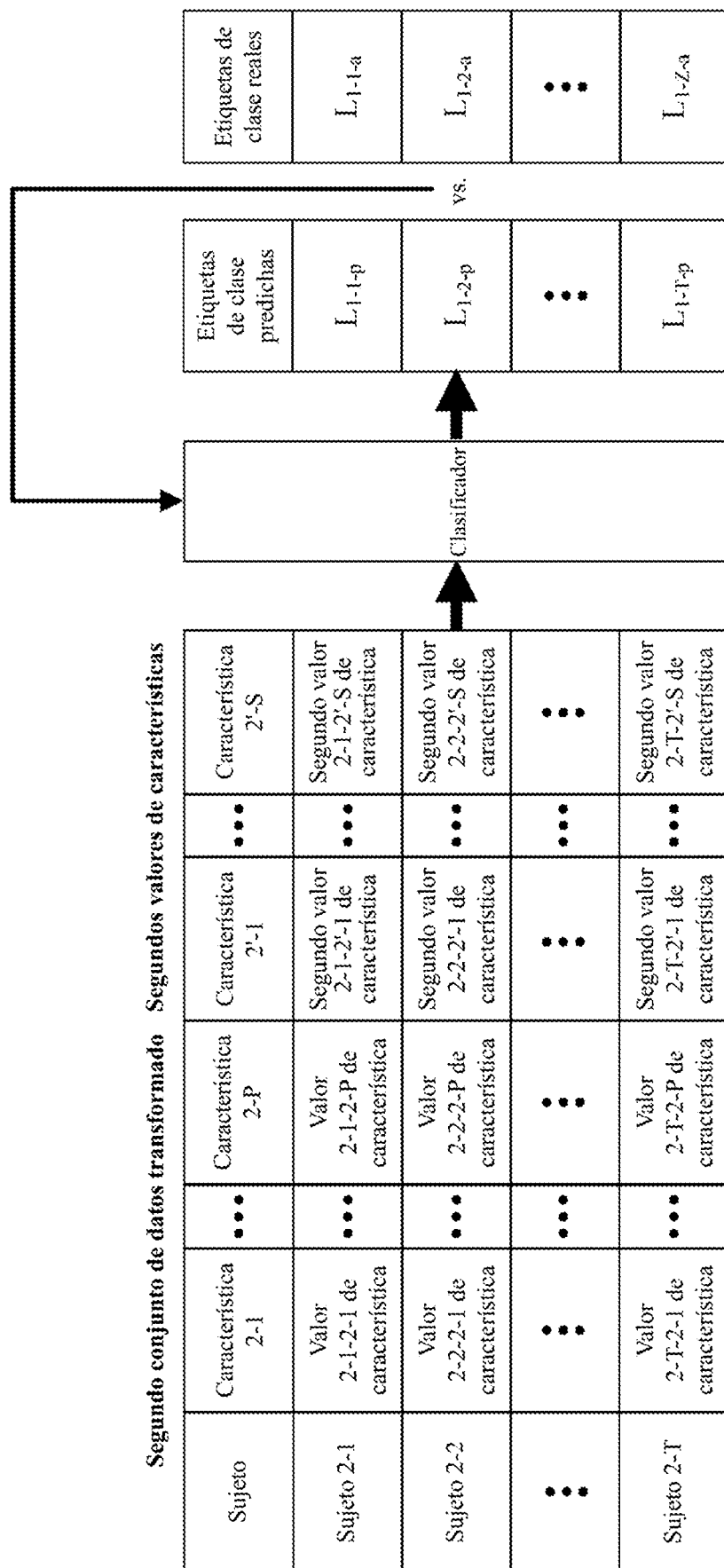


Figura 9

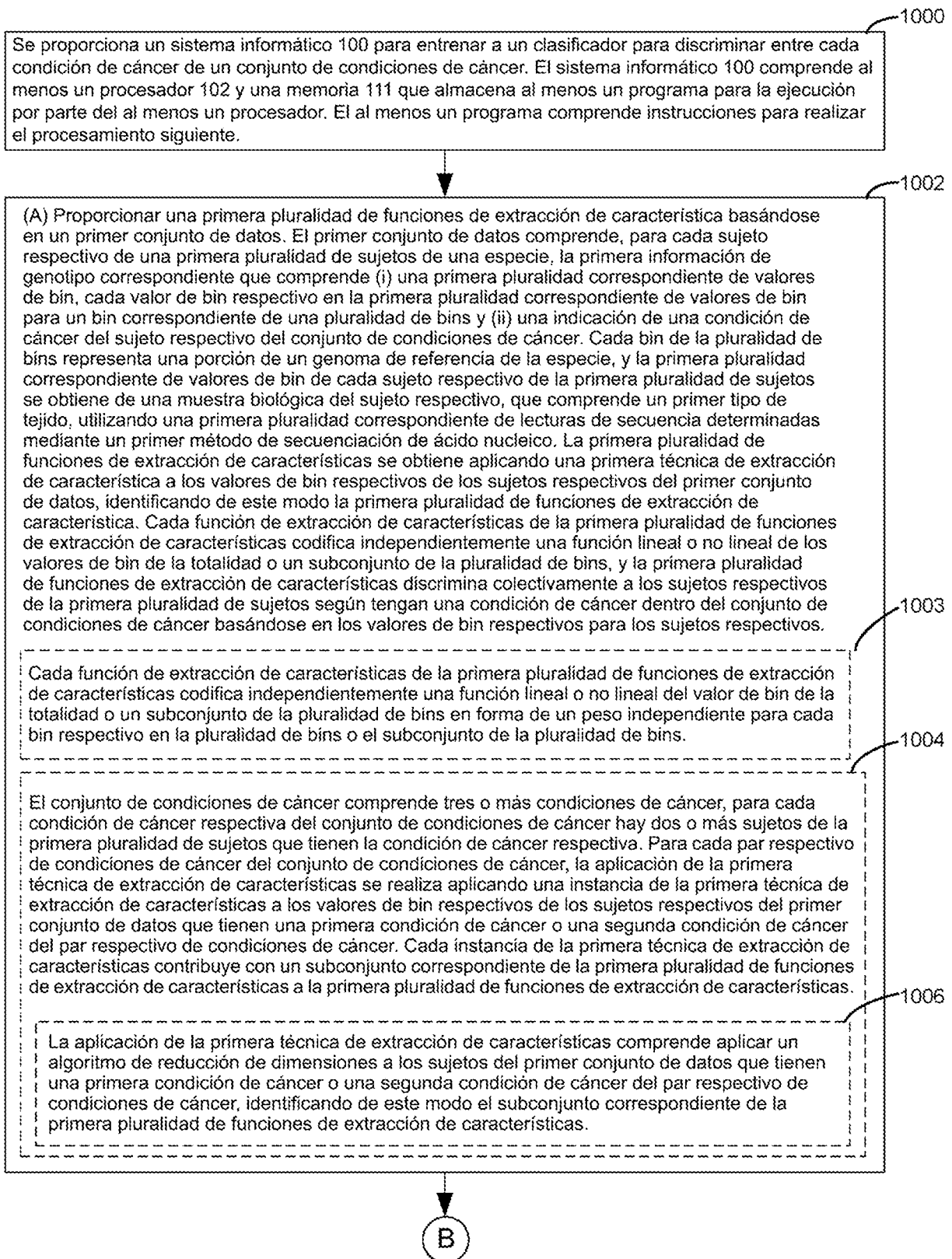


Figura 10A

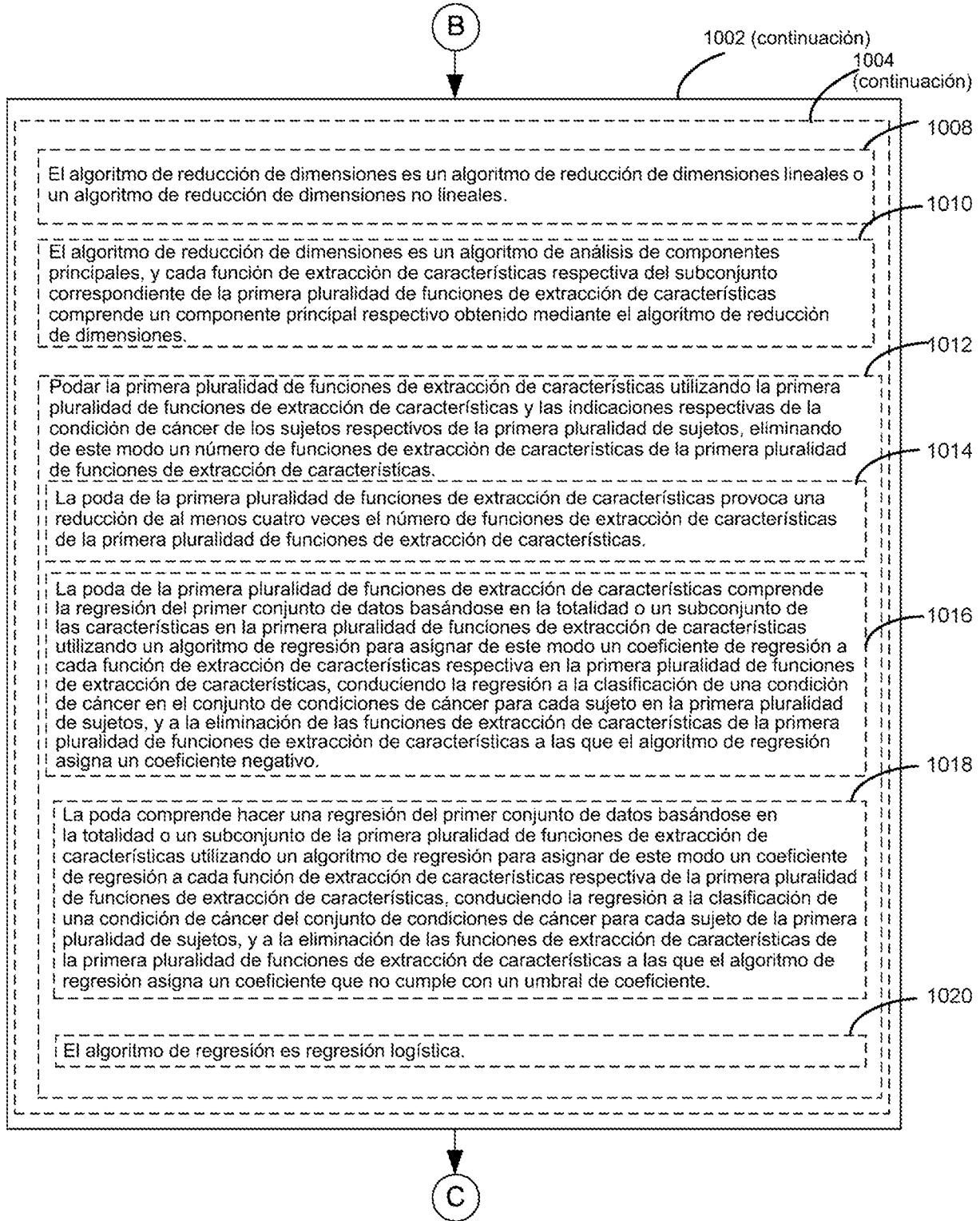


Figura 10B

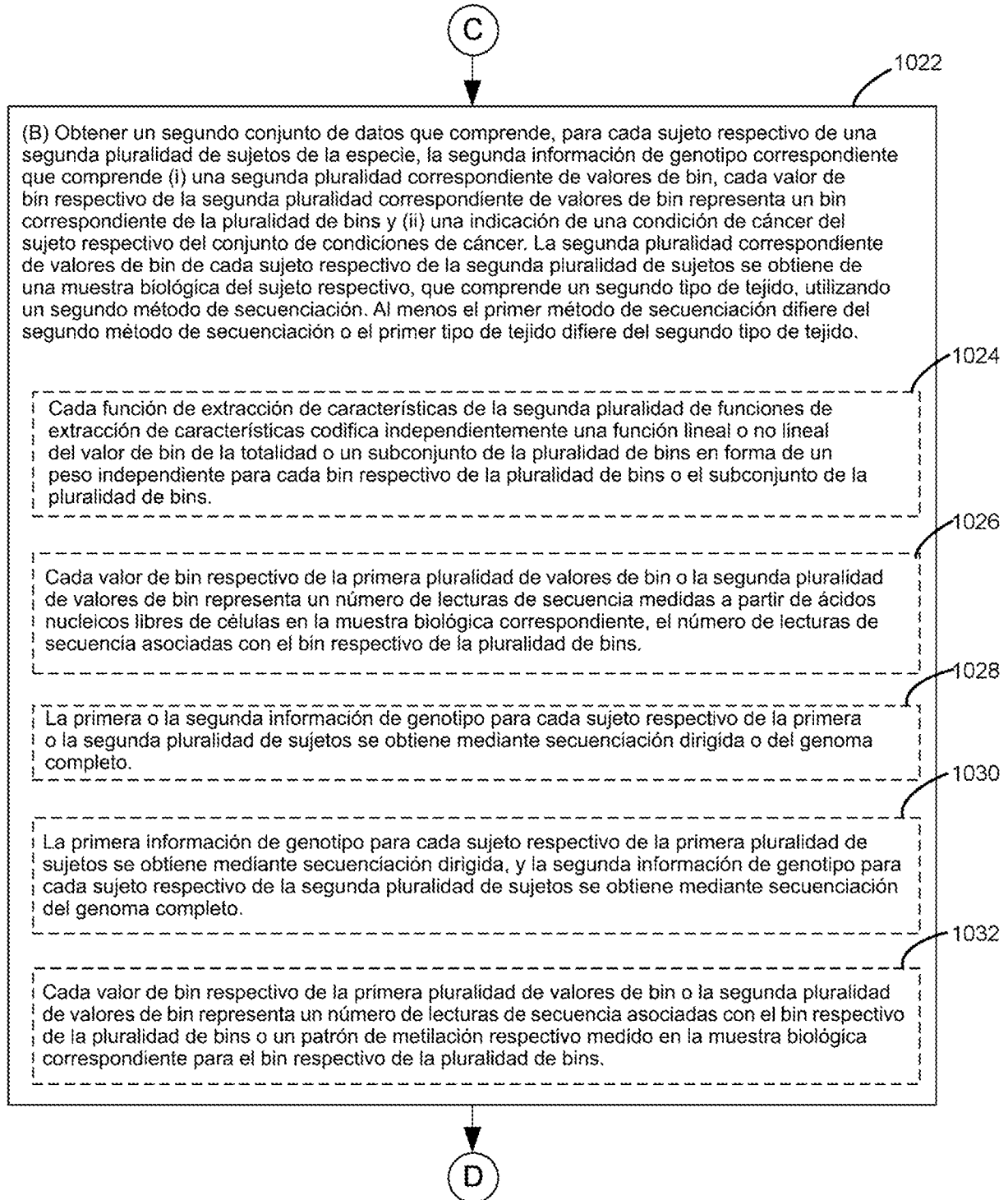


Figura 10C

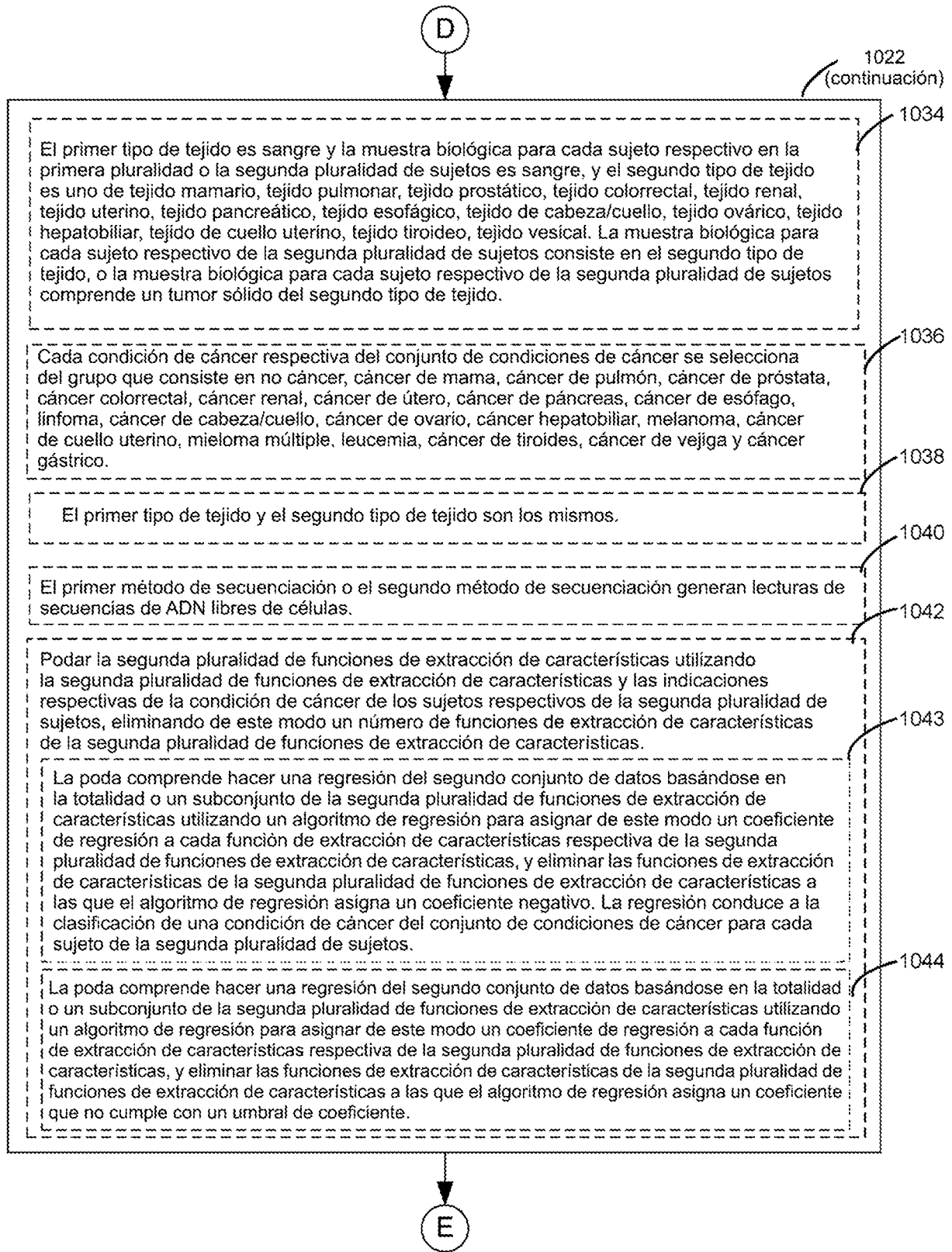


Figura 10D

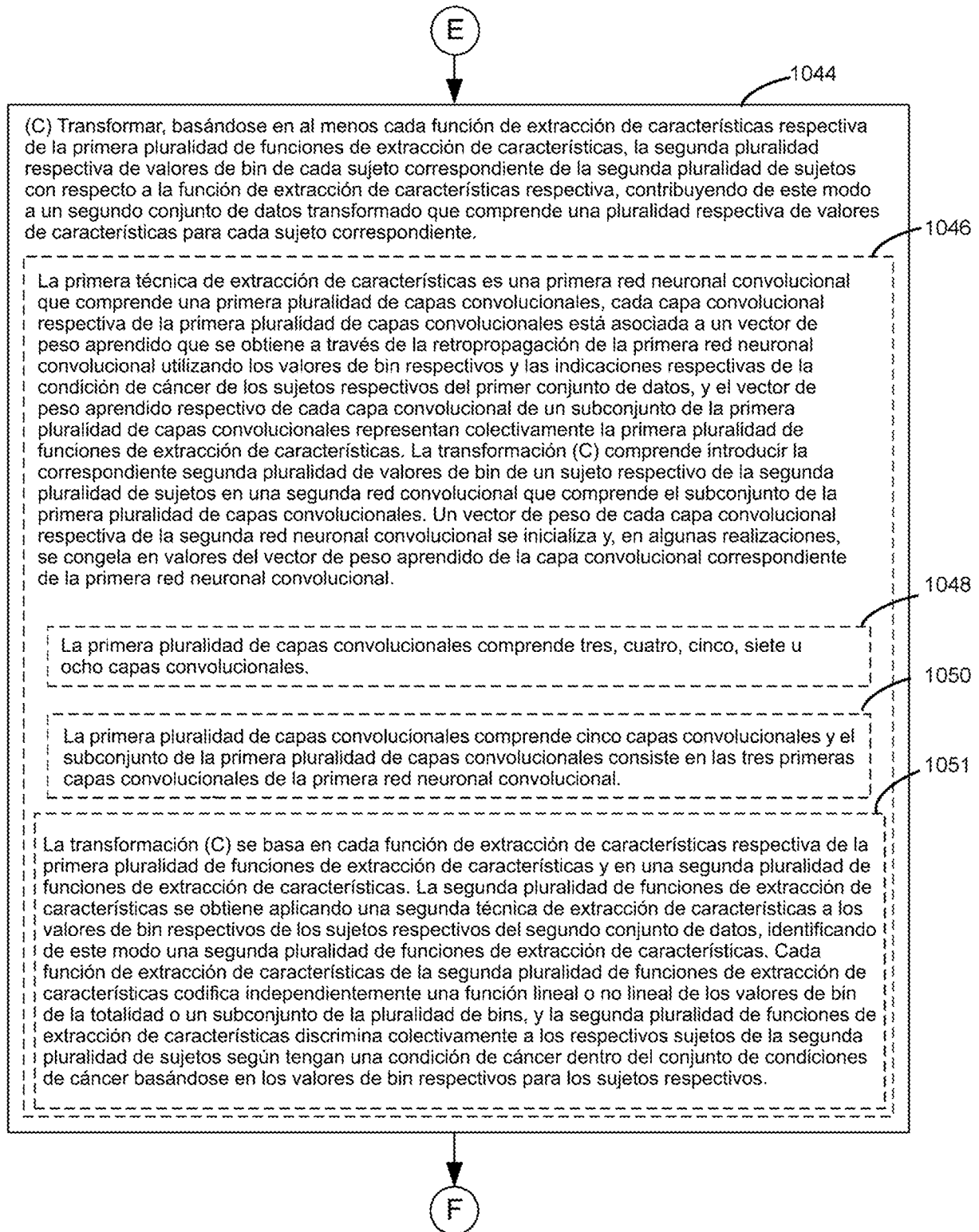


Figura 10E

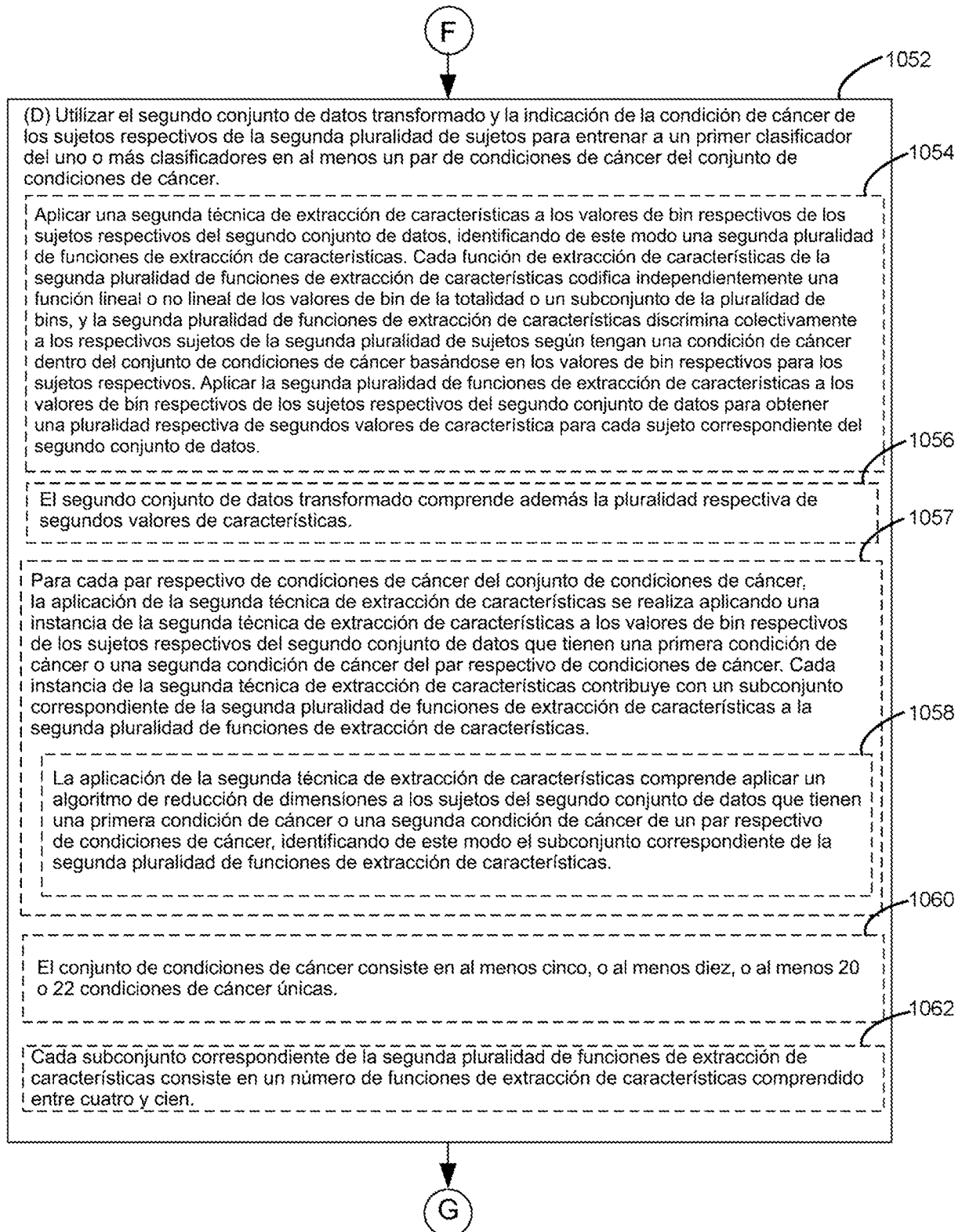


Figura 10F

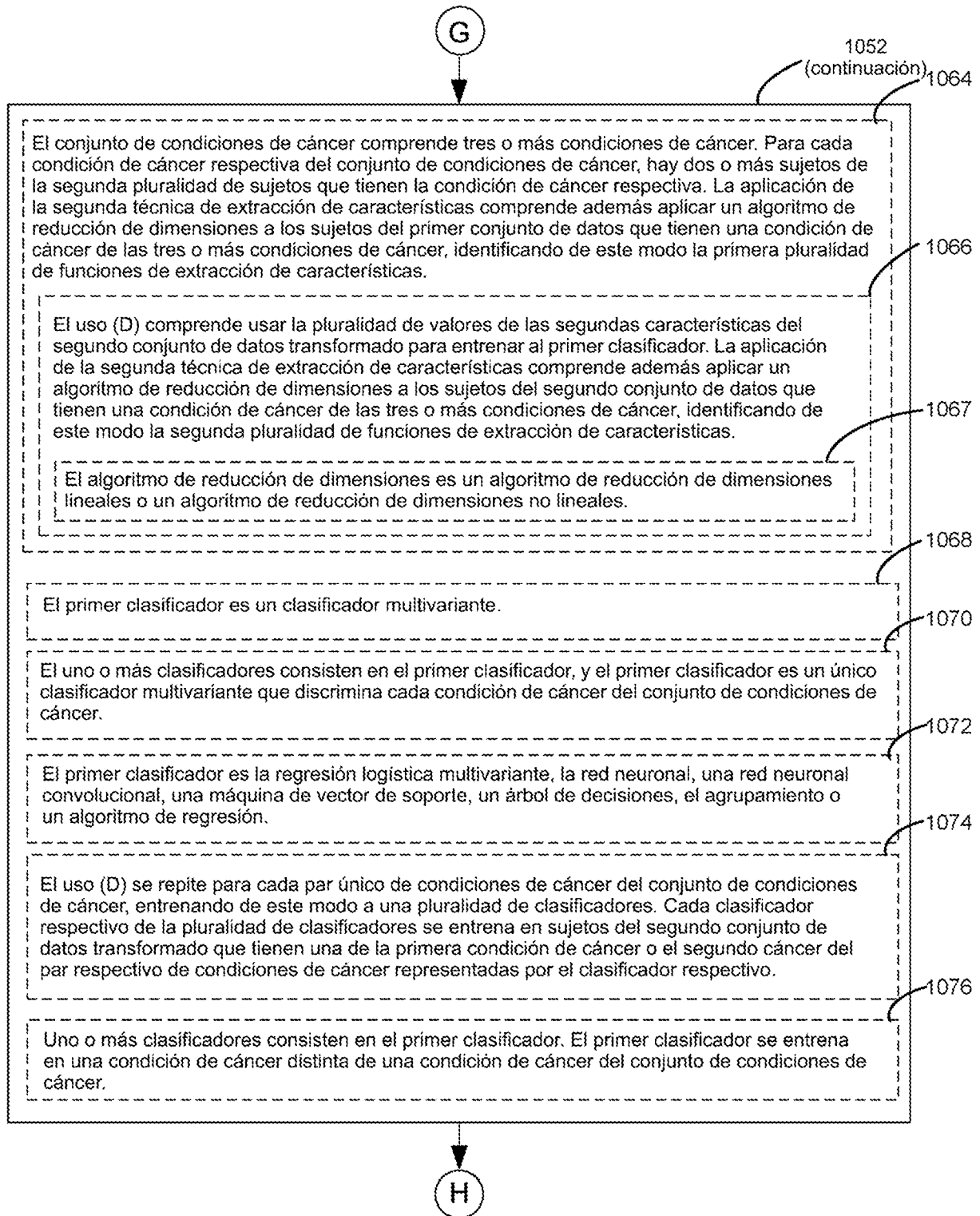


Figura 10G

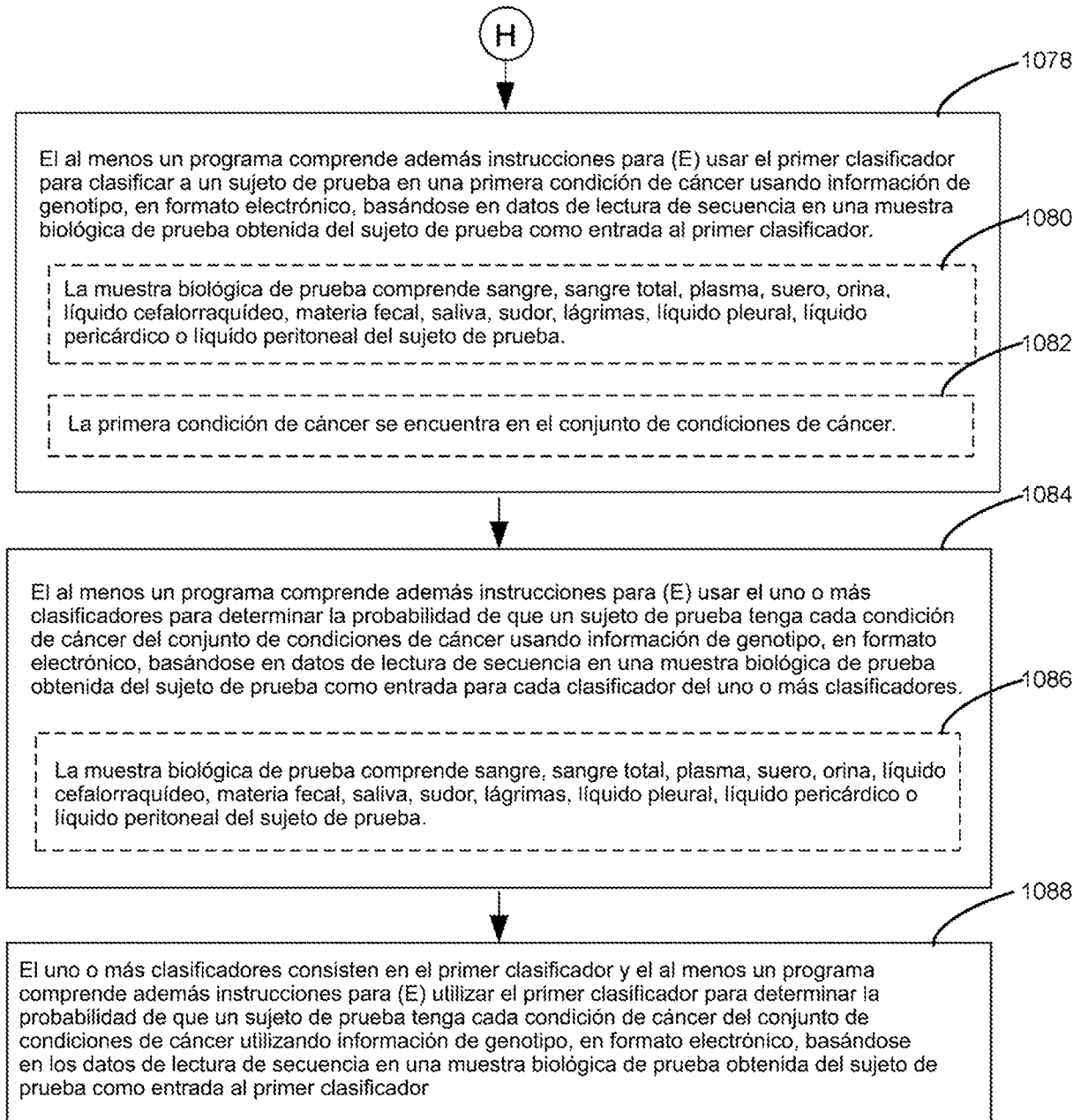


Figura 10H

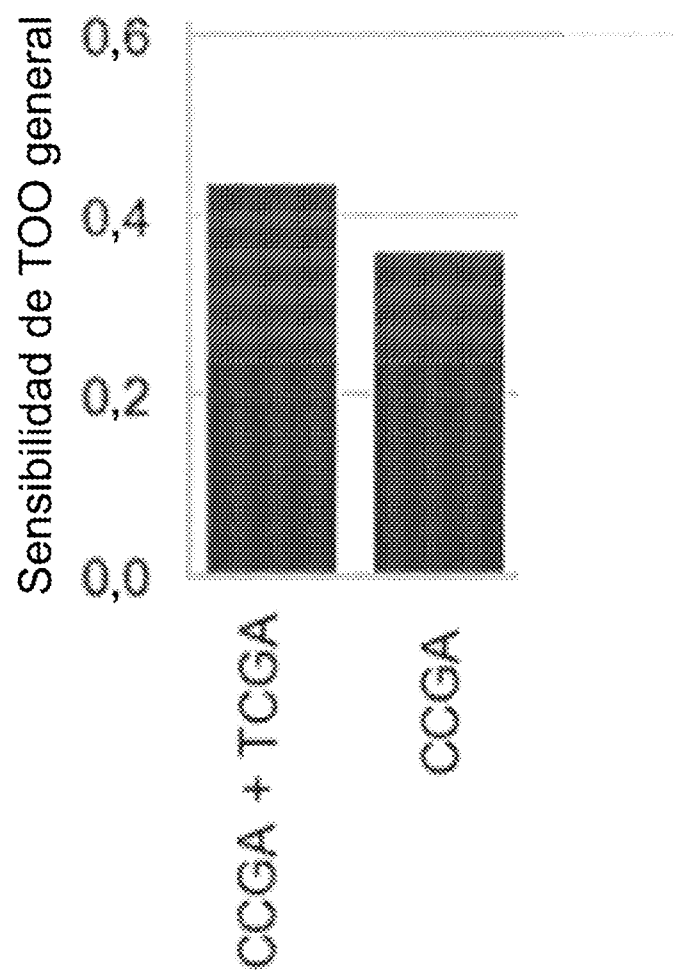


Figura 11

Primocat	De ovario	0,5	0,25
	Cabeza/cuello	0,67	0,33
	Linfoma	0,2	0,33
	Mieloma múltiple	0,64	0,55
	Colorrectal	0,52	0,3
	Páncreas	0,33	0,2
	Pulmón	0,52	0,55
	Hepatobiliar	0,33	0,11
	Mama	0,48	0,43
	Leucemia	0,25	0,25
	Esofágico	0	0,2
	Otros	0,11	0
		CCGA + TCGA (equilibrado)	CCGA (equilibrado)

Figura 12

Valor predictivo positivo

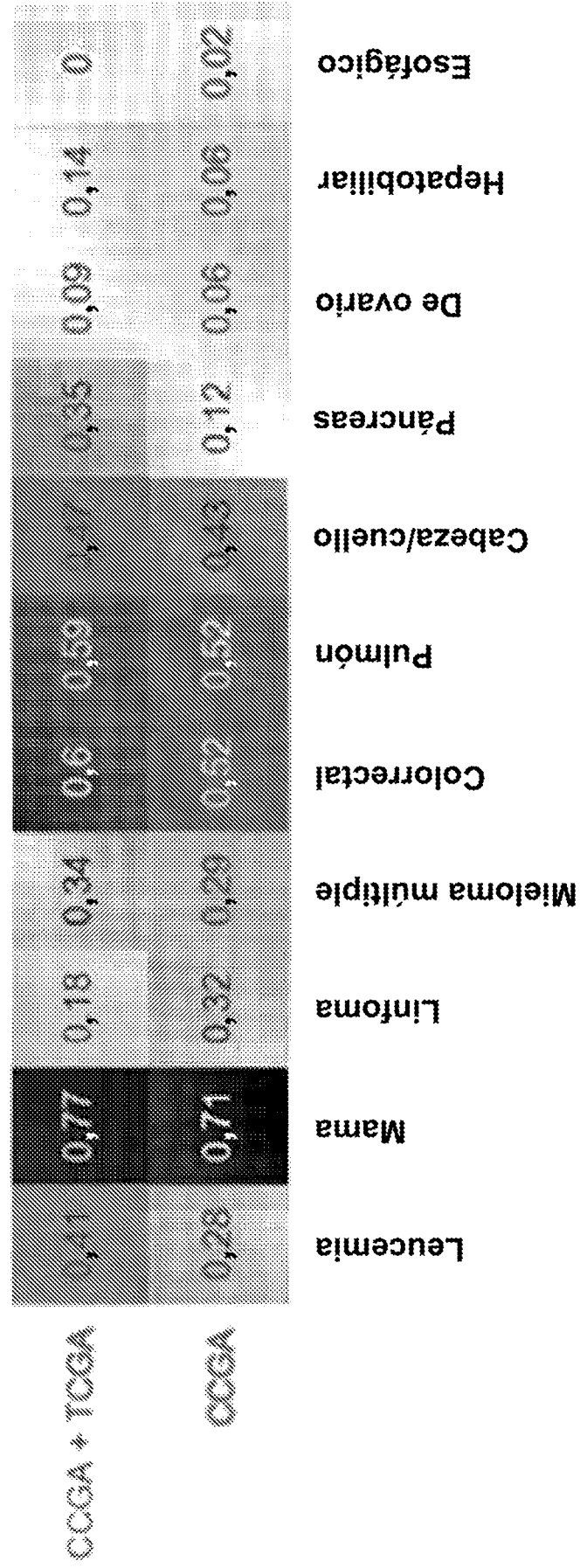


Figura 13

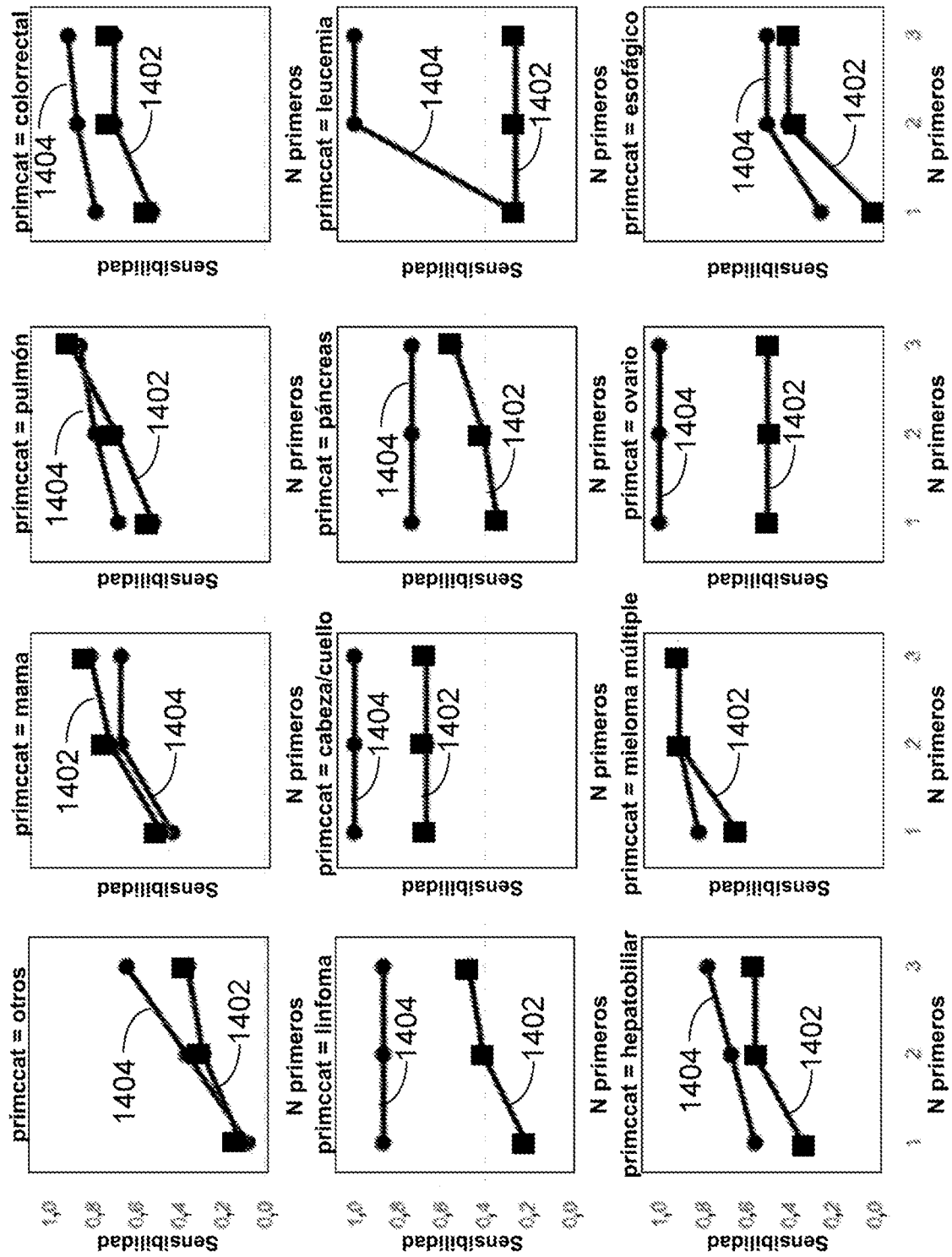


Figura 14

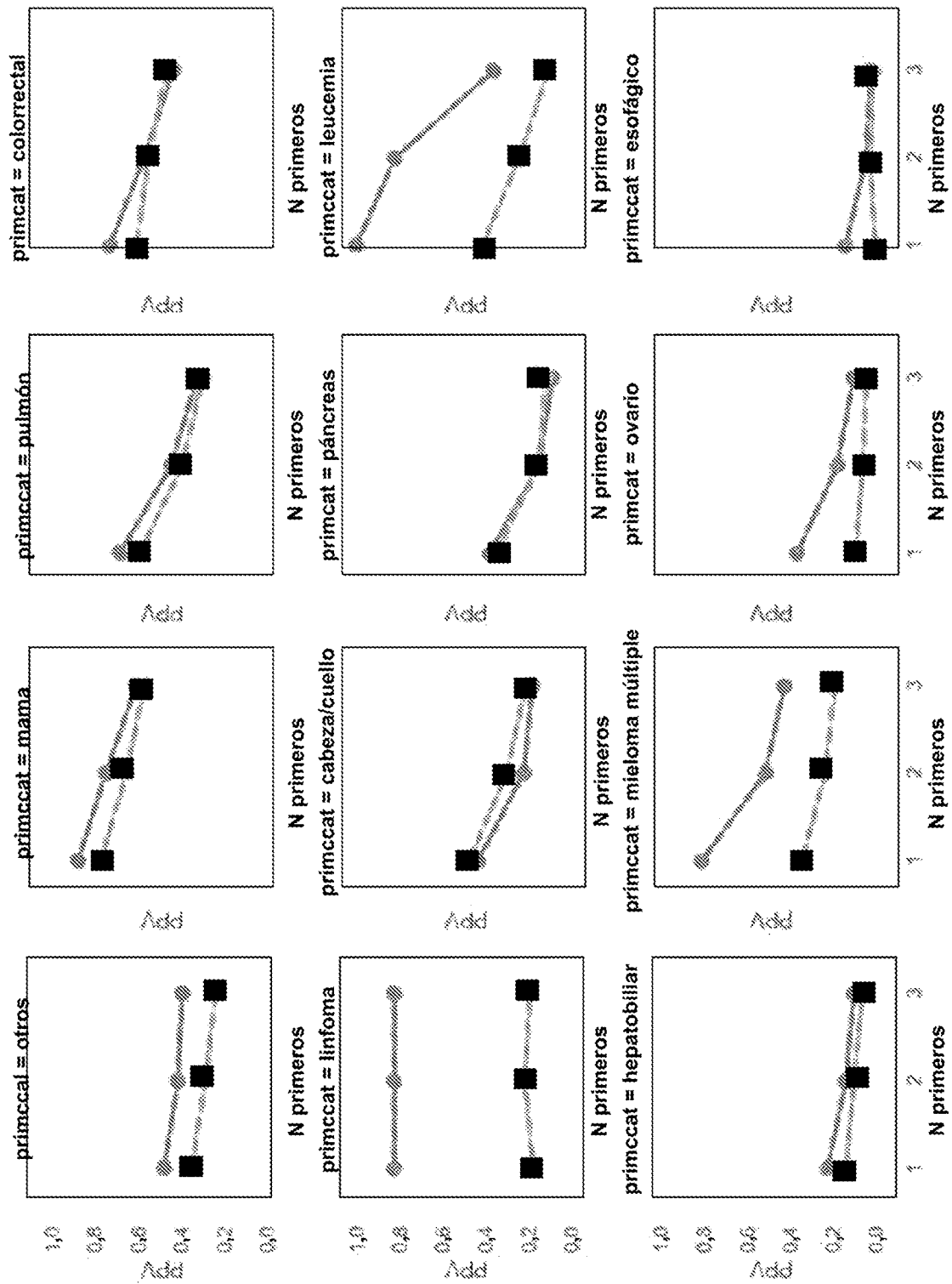


Figura 15

Aplicación a cáncer/no cáncer:

El aprendizaje por transferencia casi duplica la sensibilidad con una especificidad perfecta

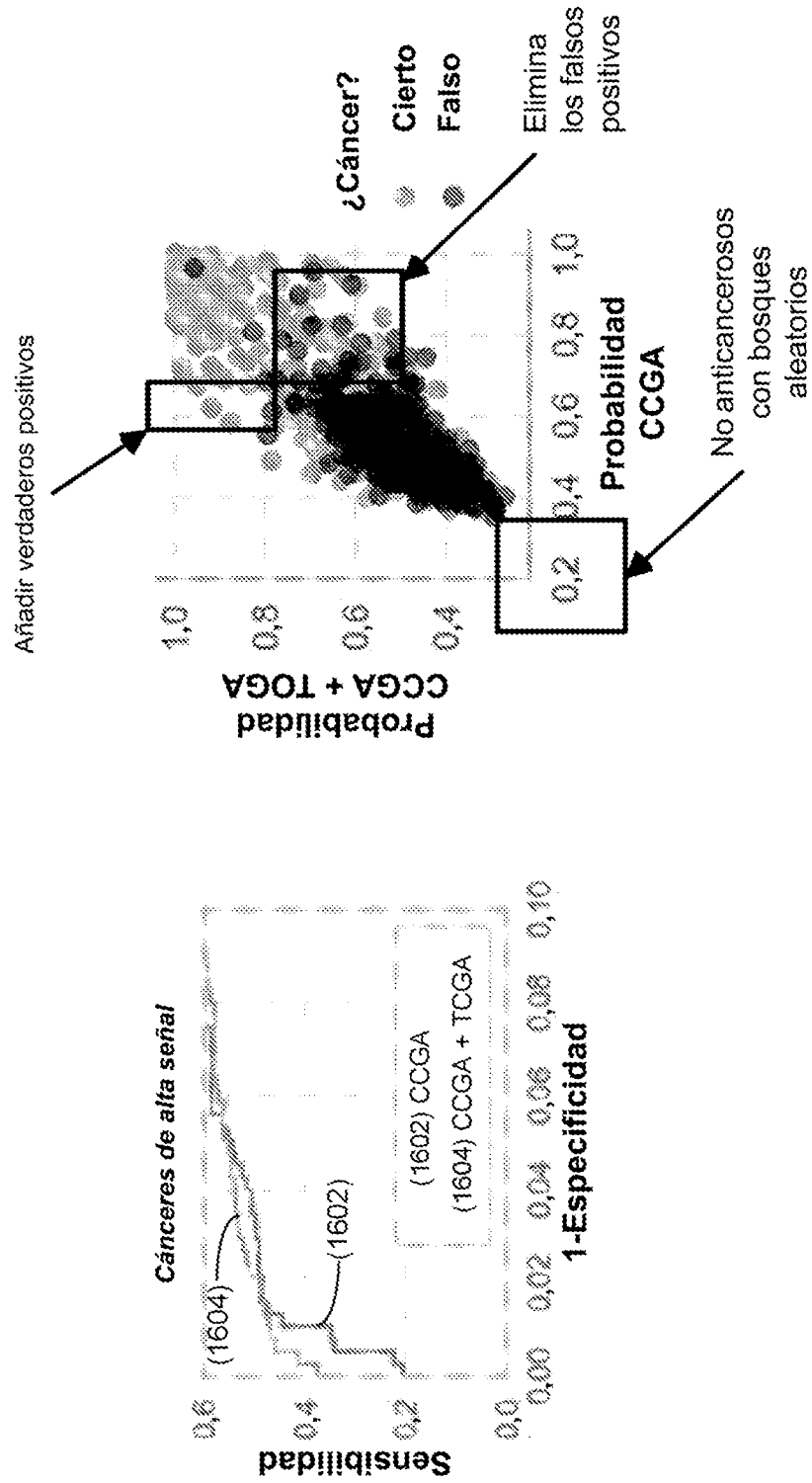


Figura 16

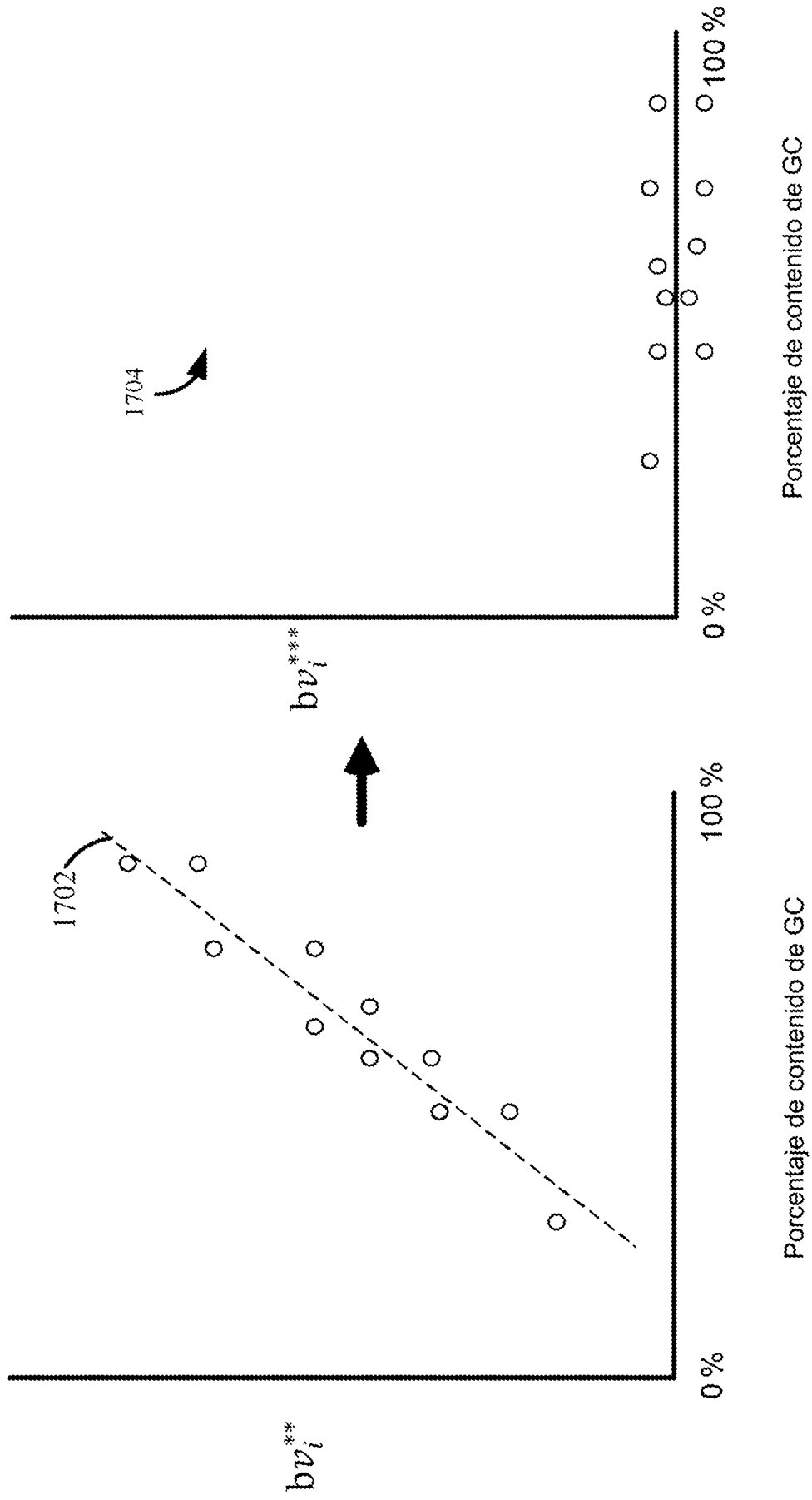


Figura 17

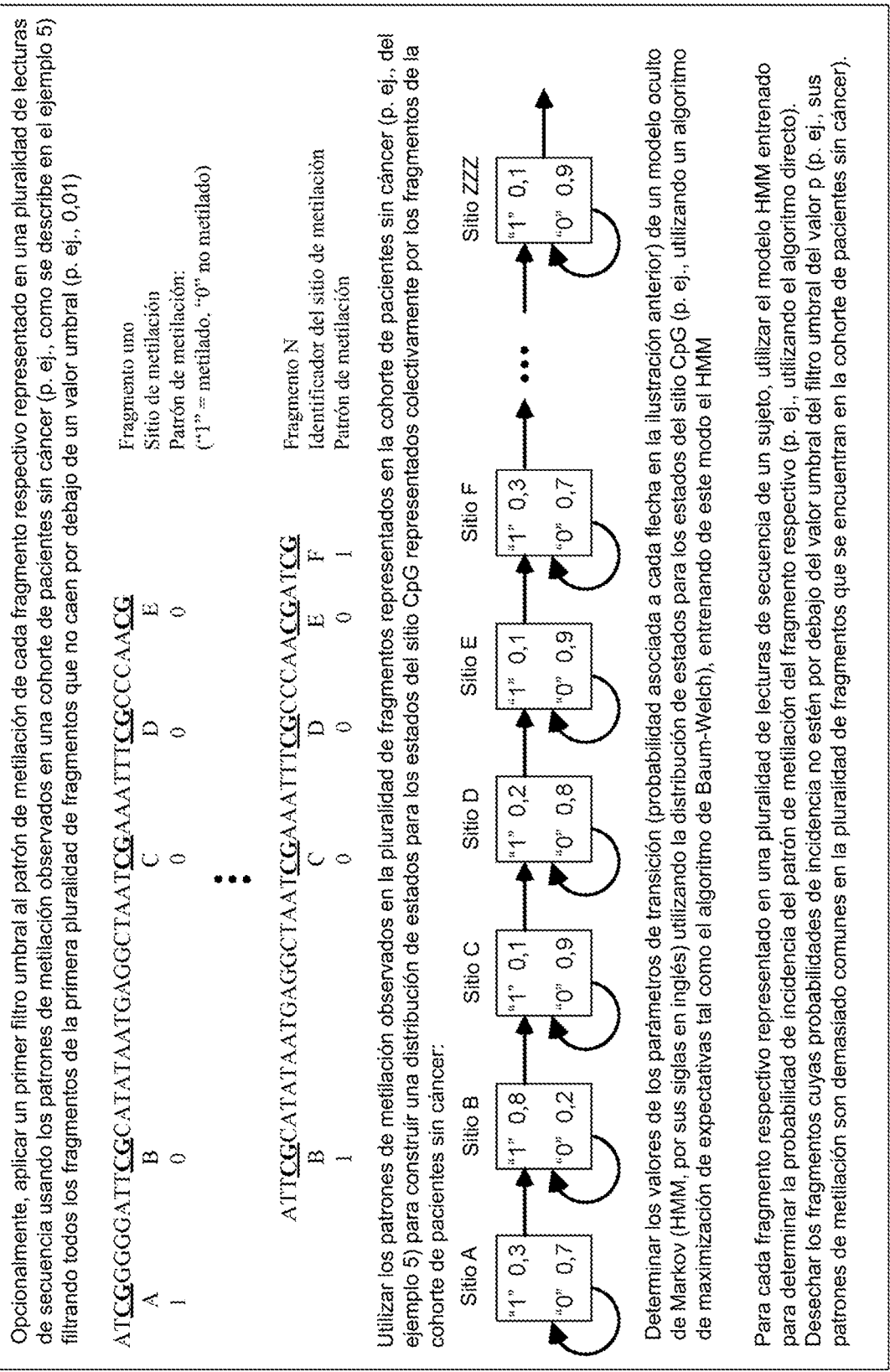


Figura 18

1900

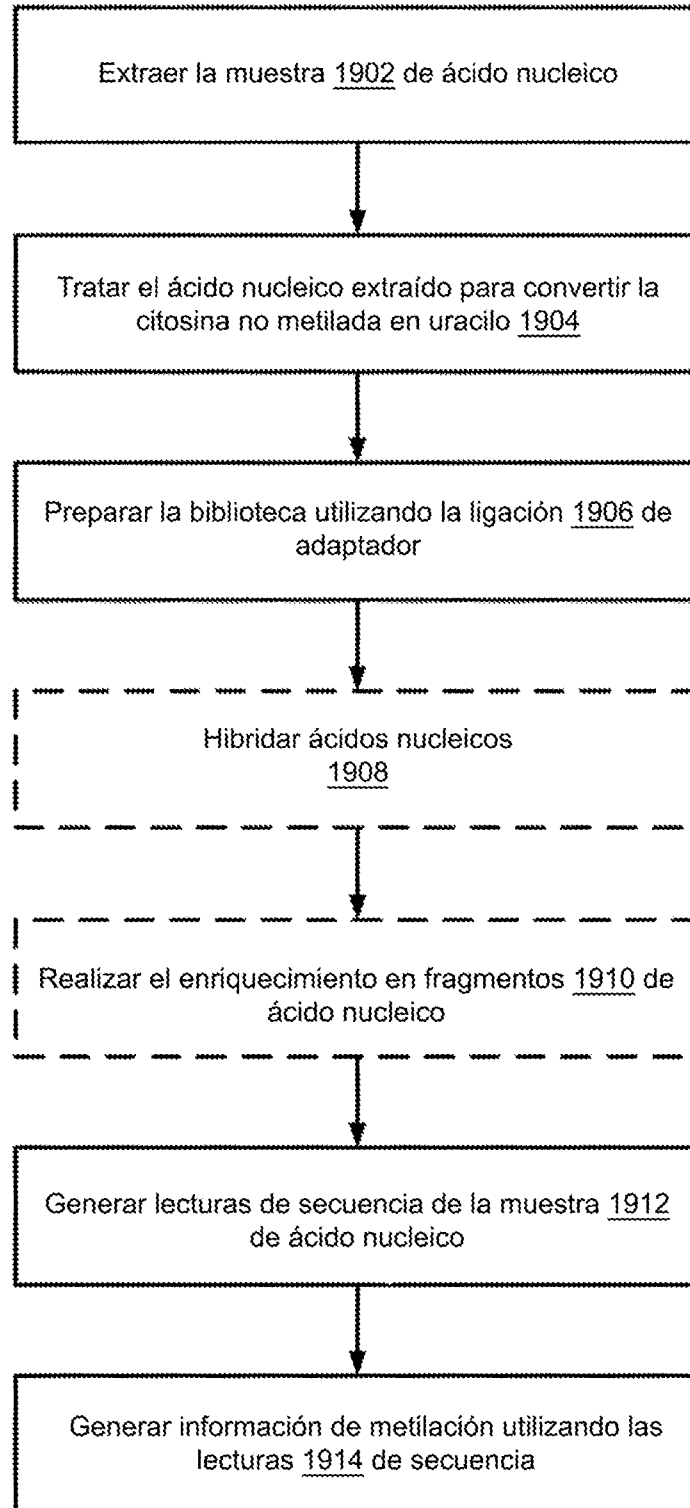


Figura 19

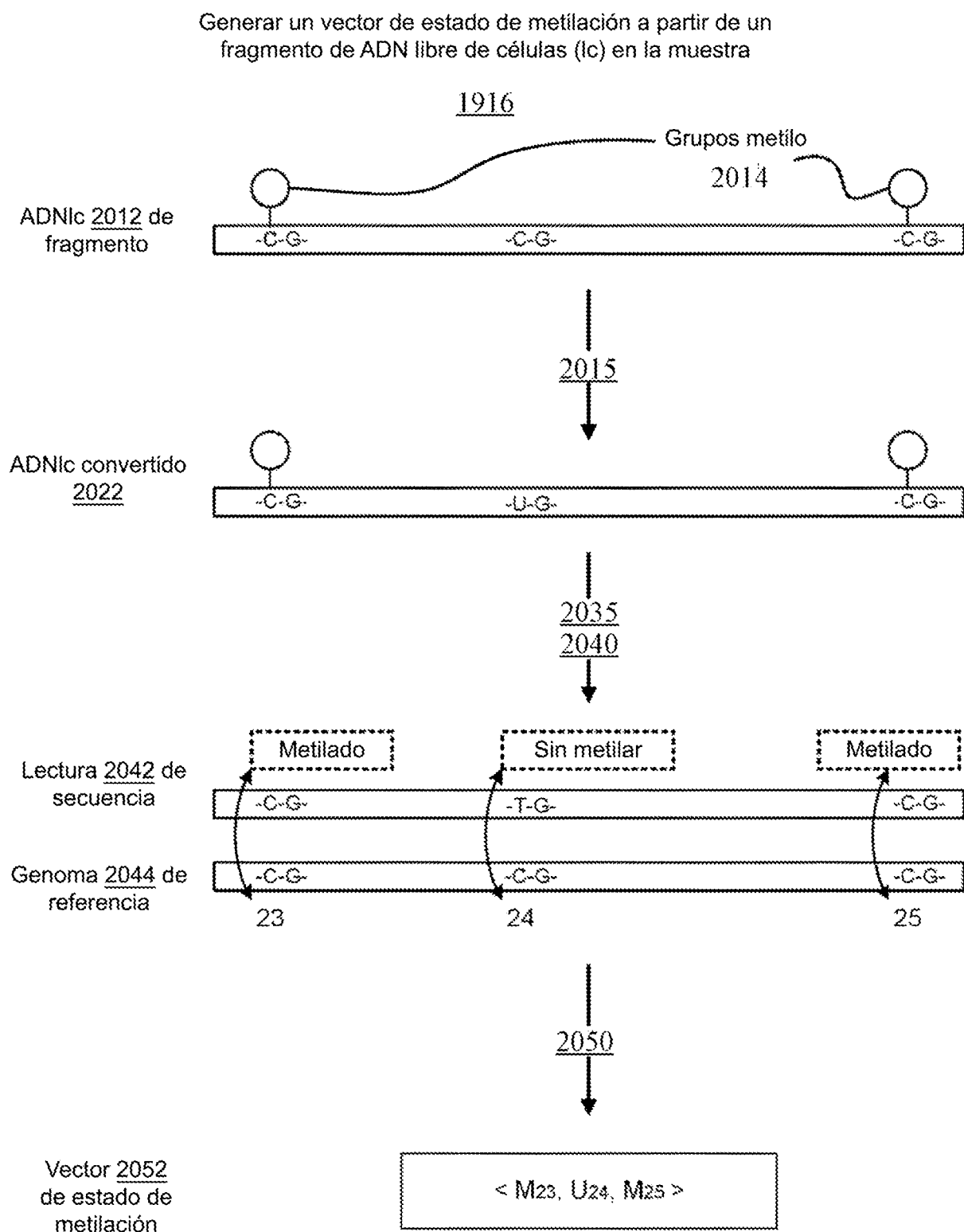


Figura 20

Generar una estructura de datos para un grupo
2100 de control

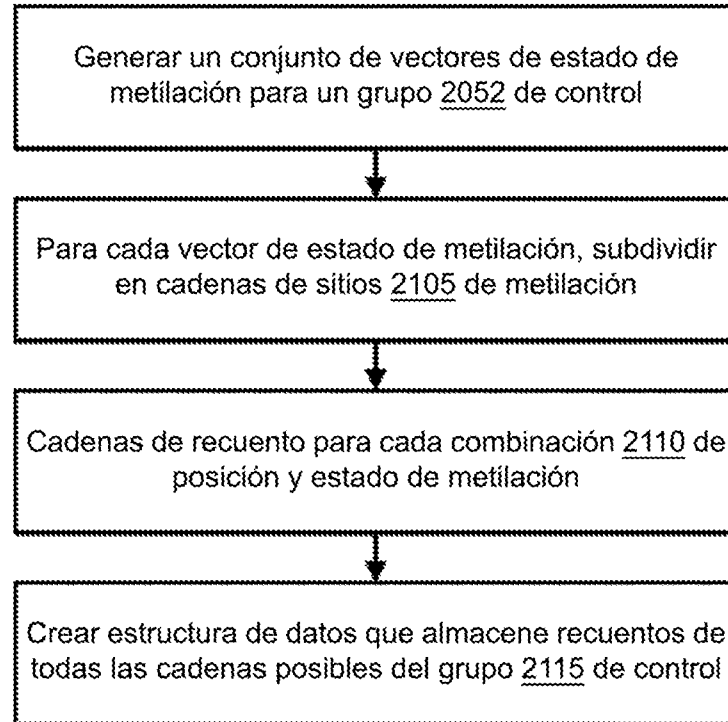


Figura 21

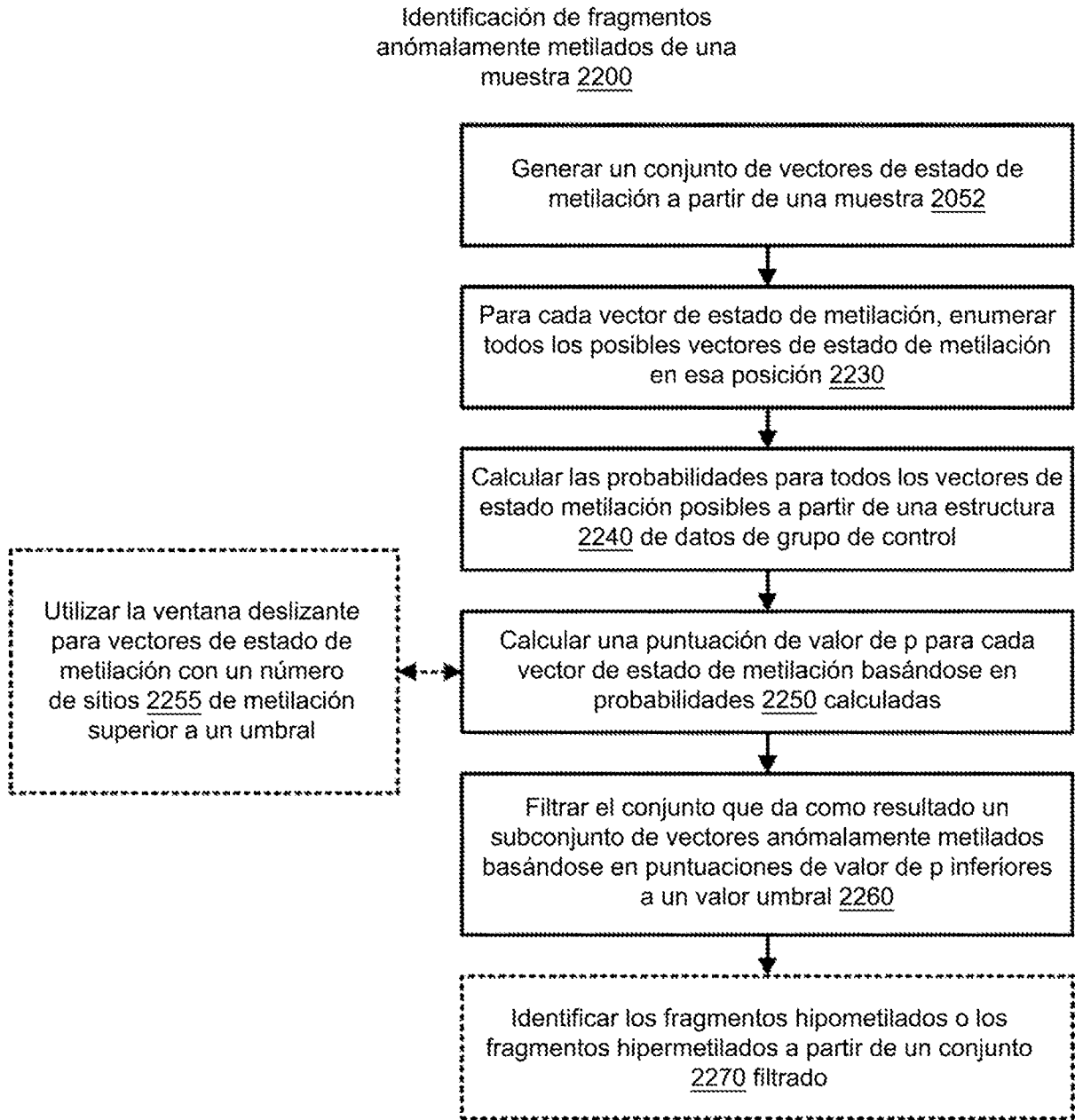


Figura 22

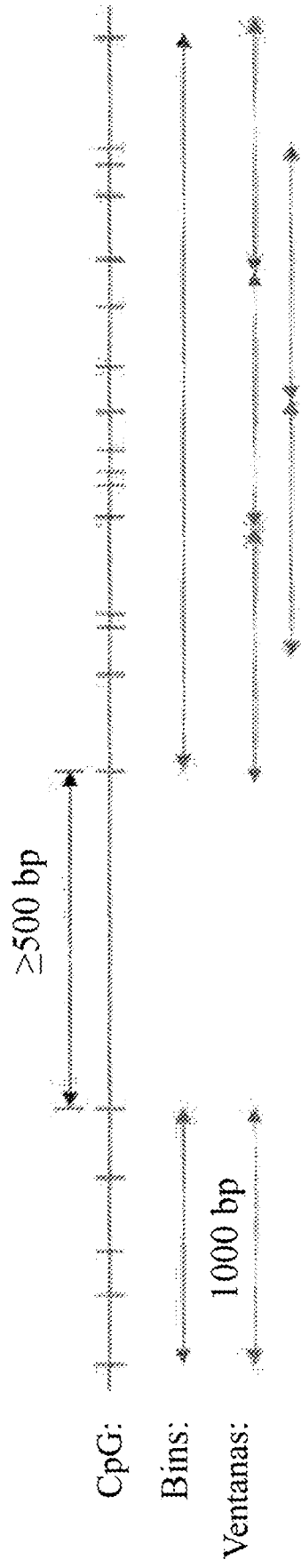


Figura 23