

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **3 013 495**

51 Int. Cl.:

C12Q 1/6806 (2008.01)

C12Q 1/6886 (2008.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **31.01.2020** **PCT/US2020/016120**

87 Fecha y número de publicación internacional: **06.08.2020** **WO20160414**

96 Fecha de presentación y número de la solicitud europea: **31.01.2020** **E 20708962 (4)**

97 Fecha y número de publicación de la concesión europea: **15.01.2025** **EP 3918089**

54 Título: **Método para el aislamiento y secuenciación de ADN libre de células**

30 Prioridad:

31.01.2019 US 201962799637 P

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

14.04.2025

73 Titular/es:

GUARDANT HEALTH, INC. (100.00%)
3100 Hanover Street
Palo Alto, CA 94304, US

72 Inventor/es:

KENNEDY, ANDREW;
JAIMOVICH, ARIEL;
SCHULTZ, MATTHEW y
GREENLEAF, WILLIAM J.

74 Agente/Representante:

IZQUIERDO BLANCO, María Alicia

ES 3 013 495 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

Método para el aislamiento y secuenciación de ADN libre de células

5 ANTECEDENTES

El cáncer es responsable de millones de muertes al año en todo el mundo. La detección precoz del cáncer puede dar como resultado resultados mejorados, ya que el cáncer en estadios tempranos tiende a ser más susceptible al tratamiento.

El crecimiento celular indebidamente controlado es un sello distintivo del cáncer que generalmente resulta de una acumulación de cambios genéticos y epigenéticos, como variaciones en el número de copias (CNV), variaciones de un solo nucleótido (SNV), fusiones de genes, inserciones y/o deleciones (indeles), las variaciones epigenéticas incluyen la 5-metilación de la citosina (5-metilcitosina) y la asociación del ADN con proteínas de la cromatina y factores de transcripción.

Las biopsias representan un enfoque tradicional para detectar o diagnosticar el cáncer en el que se extraen células o tejidos de un sitio posible de cáncer y se analizan en busca de características fenotípicas y/o genotípicas relevantes. Las biopsias tienen el inconveniente de ser invasivas.

La detección del cáncer basada en el análisis de fluidos corporales ("biopsias líquidas"), como la sangre, es una alternativa intrigante basada en la observación de que el ADN de las células cancerosas se libera en los fluidos corporales. Una biopsia líquida no es invasiva (quizá sólo requiera una extracción de sangre). Sin embargo, ha sido un desafío desarrollar métodos precisos y sensibles para analizar el material de la biopsia líquida dada la baja concentración y heterogeneidad del ADN libre de células. Aislar las fracciones de ADN libre de células útiles para análisis posteriores en procedimientos de biopsia líquida es una parte importante de este proceso. Por consiguiente, hay una necesidad de métodos y composiciones mejorados para aislar el ADN libre de células, por ejemplo, para su uso en biopsias líquidas.

El documento US 2018/0305738 A1 describe "Métodos de unión de adaptadores a ácidos nucleicos de muestra".

SUMARIO

La invención se expone en el conjunto de reivindicaciones adjunto. La presente divulgación se basa en parte en la siguiente realización. Puede ser beneficioso aislar ADN libre de células para capturar dos conjuntos de regiones diana, un conjunto de regiones diana de secuencia variable y un conjunto de regiones diana epigenéticas, en donde el rendimiento de captura del conjunto de regiones diana de secuencia variable es mayor que el rendimiento de captura del conjunto de regiones diana epigenéticas. En todas las realizaciones descritas en la presente que incluyen un conjunto de regiones diana de secuencia variable y un conjunto de regiones diana epigenéticas, el conjunto de regiones diana de secuencia variable comprende regiones que no están presentes en el conjunto de regiones diana epigenéticas y viceversa, aunque en algunos casos una fracción de las regiones puede solaparse (por ejemplo, una fracción de posiciones genómicas puede estar representada en ambos conjuntos de regiones diana). La diferencia en el rendimiento de la captura puede permitir una determinación profunda y, por lo tanto, más precisa de la secuencia en el conjunto de regiones diana de secuencia variable y una cobertura superficial y amplia en el conjunto de regiones diana epigenéticas, por ejemplo, durante la secuenciación concurrente, como en la misma célula de secuenciación o en el mismo grupo de material que se va a secuenciar.

El conjunto de regiones diana epigenéticas puede analizarse de varias maneras, que incluyen métodos que no dependen de un alto grado de precisión en la determinación de la secuencia de nucleótidos específicos dentro de una diana. Los ejemplos incluyen la determinación de la metilación y/o la distribución y tamaños de los fragmentos, que pueden indicar estructuras de cromatina normales o aberrantes en las células de las que se obtuvieron los fragmentos. Tales análisis pueden realizarse mediante secuenciación y requieren menos datos (por ejemplo, número de lecturas de secuencia o profundidad de la cobertura de secuenciación) que la determinación de la presencia o ausencia de una mutación de secuencia, como una sustitución, inserción o delección de bases.

Al contrario que los enfoques descritos en la presente, aislar un conjunto de regiones diana epigenéticas y un conjunto de regiones diana de secuencia variable con el mismo rendimiento de captura llevaría a la generación innecesaria de datos redundantes para el conjunto de regiones diana epigenéticas y/o proporcionaría menos precisión de la deseable en las determinaciones del genotipo de los miembros del conjunto de regiones diana de secuencia variable.

La presente divulgación está dirigida a satisfacer la necesidad de un aislamiento mejorado de ADN libre de células y/o proporcionar otros beneficios. Por consiguiente, se proporcionan las siguientes realizaciones ejemplares.

En otro aspecto, la presente divulgación proporciona un método para determinar una probabilidad de que un sujeto tenga cáncer que comprende: a) dividir ADNcf obtenido de un sujeto de prueba en dos o más fracciones basadas en el nivel de metilación usando un dominio de unión a metilo, y, posteriormente, en cada fracción; b) capturar una pluralidad de conjuntos de regiones diana del ADNcf, en donde: I) la pluralidad de conjuntos de regiones diana comprende un conjunto de regiones diana de secuencia variable y un conjunto de regiones diana epigenéticas, con lo que se produce un conjunto capturado de moléculas de ADNcf; II) las moléculas de ADNcf correspondientes al conjunto de regiones diana de secuencia variable se capturan con un mayor rendimiento de captura en el conjunto capturado de moléculas de ADNcf que las moléculas de ADNcf correspondientes al conjunto de regiones diana epigenéticas; y III) el conjunto de regiones diana epigenéticas se selecciona entre una o más de las siguientes: i) un conjunto de regiones diana variables de hipermetilación; ii) un conjunto de regiones diana variables de hipometilación; y/o iii) un conjunto de regiones diana variables de fragmentación, en donde el conjunto de regiones diana variables de fragmentación comprende regiones de sitios de inicio de transcripción y/o regiones de unión a CTCF; opcionalmente, en donde el conjunto de regiones diana epigenéticas comprende además un conjunto de regiones diana de control de metilación; c) secuenciar las moléculas de ADNcf capturadas, en donde las moléculas de ADNcf capturadas del conjunto de regiones diana de secuencia variable se secuencian con una profundidad de secuenciación mayor que las moléculas de ADNcf capturadas del conjunto de regiones diana epigenéticas; d) obtener una pluralidad de lecturas de secuencias generadas por un secuenciador de ácidos nucleicos a partir de la secuenciación de las moléculas de ADNcf capturadas; e) mapear la pluralidad de lecturas de secuencias a una o más secuencias de referencia para generar lecturas de secuencias mapeadas; f) procesar las lecturas de secuencias mapeadas correspondientes al conjunto de regiones diana de secuencia variable y al conjunto de regiones diana epigenéticas para determinar la probabilidad de que el sujeto padezca cáncer.

En algunas realizaciones, las moléculas de ADNcf capturadas del conjunto de regiones diana de secuencia variable se secuencian con una profundidad de secuenciación por lo menos 2 veces mayor que las moléculas de ADNcf capturadas del conjunto de regiones diana epigenéticas. En algunas realizaciones, las moléculas de ADNcf capturadas del conjunto de regiones diana de secuencia variable se secuencian con una profundidad por lo menos 3 veces mayor que las moléculas de ADNcf capturadas del conjunto de regiones diana epigenéticas. En algunas realizaciones, las moléculas de ADNcf capturadas del conjunto de regiones diana de secuencia variable se secuencian con una profundidad de secuenciación entre 4 y 10 veces mayor que las moléculas de ADNcf capturadas del conjunto de regiones diana epigenéticas. En algunas realizaciones, las moléculas de ADNcf capturadas del conjunto de regiones diana de secuencia variable se secuencian con una profundidad de secuenciación entre 4 y 100 veces mayor que las moléculas de ADNcf capturadas del conjunto de regiones diana epigenéticas.

En algunas realizaciones, la amplificación de ADNcf comprende los pasos de ligar adaptadores que contienen códigos de barras al ADNcf. En algunas realizaciones, la amplificación de ADNcf comprende los pasos de ligar adaptadores que contienen códigos de barras al ADNcf.

En algunas realizaciones, capturar la pluralidad de conjuntos de regiones diana de ADNcf comprende poner en contacto del ADNcf con sondas de unión a diana específicas para el conjunto de regiones diana de secuencia variable y sondas de unión a diana específicas para el conjunto de regiones diana epigenéticas. En algunas realizaciones, las sondas de unión a diana específicas para el conjunto de regiones diana de secuencia variable están presentes en una concentración más alta que las sondas de unión a diana específicas para el conjunto de regiones diana epigenéticas. En algunas realizaciones, las sondas de unión a diana específicas para el conjunto de regiones diana de secuencia variable están presentes en una concentración por lo menos 2 veces mayor que la de las sondas de unión a diana específicas para el conjunto de regiones diana epigenéticas. En algunas realizaciones, las sondas de unión a diana específicas para el conjunto de regiones diana de secuencia variable están presentes en una concentración por lo menos 4 o 5 veces mayor que la de las sondas de unión a diana específicas para el conjunto de regiones diana epigenéticas. En algunas realizaciones, las sondas de unión a diana específicas para el conjunto de regiones diana de secuencia variable tienen una afinidad de unión a diana mayor que la de las sondas de unión a diana específicas para el conjunto de regiones diana epigenéticas.

El ADNcf obtenido del sujeto de prueba se divide en por lo menos 2 fracciones sobre la base del nivel de metilación, y los pasos posteriores del método se realizan en cada fracción.

En algunas realizaciones, el paso de división comprende poner en contacto el ADNcf recogido con un reactivo de unión a metilo inmovilizado en un soporte sólido.

En algunas realizaciones, el rendimiento de captura de las sondas de unión a diana específicas para el conjunto de regiones diana de secuencia variable es por lo menos 4 o 5 veces mayor que el rendimiento de captura de las sondas de unión a diana específicas para el conjunto de regiones diana epigenéticas.

En algunas realizaciones, hay por lo menos 10 regiones en el conjunto de regiones diana de secuencia variable y por lo menos 100 regiones en el conjunto de regiones diana epigenéticas.

En algunas realizaciones, las sondas están presentes en una única solución. En algunas realizaciones, las

sondas comprenden una fracción de captura.

En otro aspecto, la presente divulgación proporciona un sistema que comprende una interfaz de comunicación que recibe, a través de una red de comunicación, una pluralidad de lecturas de secuencia generadas por un secuenciador de ácido nucleico a partir de la secuenciación de un conjunto capturado de moléculas de ADNcf, en donde el conjunto capturado de moléculas de ADNcf se obtiene de acuerdo con los pasos (a) y (b) del método descrito anteriormente, en donde las moléculas de ADNcf capturadas correspondientes a la región diana de secuencia variable se secuencian a una mayor profundidad de secuenciación que las moléculas de ADNcf capturadas correspondientes al conjunto de regiones diana epigenéticas; y un controlador que comprende o es capaz de acceder a medios legibles por ordenador que comprenden instrucciones ejecutables por ordenador no transitorias que, cuando son ejecutadas por al menos un procesador electrónico, realizan un método que comprende: (i) recibir, a través de la red de comunicación, las lecturas de secuencia generadas por el secuenciador de ácidos nucleicos; (ii) mapear la pluralidad de lecturas de secuencia con una o más secuencias de referencia para generar lecturas de secuencia mapeadas; (iii) procesar las lecturas de secuencia mapeadas correspondientes al conjunto de regiones diana de secuencia variable y al conjunto de regiones diana epigenéticas para determinar la probabilidad de que el sujeto tenga cáncer.

En algunas realizaciones, la profundidad de secuenciación correspondiente al conjunto de regiones diana de secuencia variable es por lo menos 2 veces mayor que la profundidad de secuenciación correspondiente al conjunto de regiones diana epigenéticas. En algunas realizaciones, la profundidad de secuenciación correspondiente al conjunto de regiones diana de secuencia variable es por lo menos 3 veces mayor que la profundidad de secuenciación correspondiente al conjunto de regiones diana epigenéticas. En algunas realizaciones, la profundidad de la secuenciación correspondiente al conjunto de regiones diana de secuencia variable es de 4 a 10 veces mayor que la profundidad de la secuenciación correspondiente al conjunto de regiones diana epigenéticas. En algunas realizaciones, la profundidad de secuenciación correspondiente al conjunto de regiones diana de secuencia variable es de 4 a 100 veces mayor que la profundidad de secuenciación correspondiente al conjunto de regiones diana epigenéticas.

En algunas realizaciones, las moléculas de ADNcf capturadas del conjunto de regiones diana de secuencia variable se agrupan con las moléculas de ADNcf capturadas del conjunto de regiones diana epigenéticas antes de la secuenciación. En algunas realizaciones, las moléculas de ADNcf capturadas del conjunto de regiones diana de secuencia variable y las moléculas de ADNcf capturadas del conjunto de regiones diana epigenéticas se secuencian en la misma célula de secuenciación.

En algunas realizaciones, el conjunto de regiones diana epigenéticas comprende un conjunto de regiones diana variables de hipermetilación. En algunas realizaciones, el conjunto de regiones diana epigenéticas comprende un conjunto de regiones diana variables de hipometilación. En algunas realizaciones, el conjunto de regiones diana epigenéticas comprende además un conjunto de regiones diana de control de metilación. En algunas realizaciones, el conjunto de regiones diana epigenéticas comprende un conjunto de regiones diana variables de fragmentación. En algunas realizaciones, el conjunto de regiones diana variables de fragmentación comprende regiones de sitios de inicio de transcripción. En algunas realizaciones, el conjunto de regiones diana variables de fragmentación comprende regiones de unión a CTCF.

En algunas realizaciones, la huella del conjunto de regiones diana epigenéticas es por lo menos 2 veces mayor que el tamaño del conjunto de regiones diana de secuencia variable. En algunas realizaciones, la huella del conjunto de regiones diana epigenéticas es por lo menos 10 veces mayor que el tamaño del conjunto de regiones diana de secuencia variable.

En algunas realizaciones, la huella del conjunto de regiones objetivo de secuencia variable es de por lo menos 25 kB o 50 kB.

En algunas realizaciones, el ADNcf capturado comprende etiquetas de secuencia. En algunas realizaciones, las etiquetas de secuencia comprenden códigos de barras. En algunas realizaciones, la concentración de las regiones diana de secuencia variable es por lo menos 2 veces mayor que la concentración de las regiones diana epigenéticas. En algunas realizaciones, la concentración de las regiones diana de secuencia variable es por lo menos 4 o 5 veces mayor que la concentración de las regiones diana epigenéticas. En algunas realizaciones, las concentraciones son concentraciones de masa por volumen normalizadas para los tamaños de huella de las regiones diana.

En algunas realizaciones, las regiones diana epigenéticas comprenden una, dos, tres o cuatro de las regiones diana variables de hipermetilación; regiones diana variables de hipometilación; regiones de sitios de inicio de transcripción; y regiones de unión a CTCF; opcionalmente en donde las regiones diana epigenéticas comprenden además regiones diana de control de metilación.

En algunas realizaciones, la composición se produce de acuerdo con un método divulgado en otra parte de la presente. En algunas realizaciones, la captura se lleva a cabo en un único recipiente.

En algunas realizaciones, los resultados de los sistemas y/o métodos divulgados en la presente se usan como entrada para generar un informe. El informe puede estar en papel o en formato electrónico. Por ejemplo, en tal informe puede mostrarse la información en, y/o la información derivada de, la información de la secuencia según lo determinado mediante los métodos o los sistemas divulgados en la presente. En algunas realizaciones, esta información es el estado del cáncer del sujeto, según lo determinado por los métodos o sistemas divulgados en la presente. Los métodos o sistemas divulgados en la presente pueden comprender además un paso de comunicar el informe a un tercero, como el sujeto del que procede la muestra o un profesional sanitario.

Los varios pasos de los métodos divulgados en la presente, o los pasos llevados a cabo mediante los sistemas divulgados en la presente, pueden llevarse a cabo al mismo tiempo o en momentos diferentes, y/o en la misma ubicación geográfica o en diferentes ubicaciones geográficas, por ejemplo, países. Los varios pasos de los métodos divulgados en la presente pueden ser realizadas por la misma persona o por personas diferentes.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

Los dibujos acompañantes, que se incorporan y forman parte de esta memoria descriptiva, ilustran ciertas realizaciones y, junto con la descripción escrita, sirven para explicar ciertos principios de los métodos, medios legibles por ordenador y sistemas divulgados en la presente. La descripción proporcionada en la presente se entiende mejor cuando se lee junto con los dibujos acompañantes que se incluyen a modo de ejemplo y no de limitación. Se entenderá que en todos los dibujos números de referencia similares identifican componentes similares, a menos que el contexto indique lo contrario. También se entenderá que algunas o todas las figuras pueden ser representaciones esquemáticas con propósitos de ilustración y no representan necesariamente los tamaños o ubicaciones relativos reales de los elementos mostrados.

La FIG. 1 muestra una visión general de la metodología de división.

La FIG. 2 es un diagrama esquemático de un ejemplo de un sistema adecuado para su uso con algunas realizaciones de la divulgación.

La FIG. 3 muestra la sensibilidad de detección de cánceres de diferentes estadios usando una o ambas regiones diana epigenéticas como de regiones diana de secuencia variable en una prueba de biopsia líquida como se describe en el Ejemplo ii.

La FIG. 4 muestra la supervivencia libre de recurrencia a lo largo del tiempo para los sujetos en los que se detectó o no ADNct como se describe en el Ejemplo iii.

DESCRIPCIÓN DETALLADA

A continuación se hará referencia en detalle a ciertas realizaciones de la invención. Aunque la invención se describirá junto con dichas realizaciones, se entenderá que no se pretende limitar la invención a dichas realizaciones. Por el contrario, se pretende que la invención cubra todas las realizaciones que caen bajo el alcance de las reivindicaciones adjuntas.

Antes de describir en detalle las presentes enseñanzas, debe entenderse que la divulgación no se limita a composiciones o pasos de proceso específicos, ya que éstos pueden variar. Debe tenerse en cuenta que, como se usa en esta memoria descriptiva y en las reivindicaciones adjuntas, la forma singular "un", "uno" y "el" incluyen referencias plurales a menos que el contexto dicte claramente lo contrario. Así, por ejemplo, la referencia a "un ácido nucleico" incluye una pluralidad de ácidos nucleicos, la referencia a "una célula" incluye una pluralidad de células, y similares.

Los intervalos numéricos incluyen los números que definen el intervalo. Los valores medidos y medibles se entienden como aproximados, teniendo en cuenta los dígitos significativos y el error asociado a la medición. Además, no se pretende que el uso de "comprender", "comprende", "que comprende", "contener", "contiene", "que contiene", "incluir", "incluye" e "incluyendo" sea limitativo. Debe entenderse que tanto la descripción general anterior como la descripción detallada son solamente ejemplares y explicativas y no son restrictivas de las enseñanzas.

A menos que se indique específicamente en la memoria descriptiva anterior, las realizaciones en la memoria descriptiva que mencionan "que comprenden" varios componentes también se contemplan como "que consisten en" o "que consisten esencialmente en" los componentes mencionados; las realizaciones de la memoria descriptiva que mencionan "que consisten en" varios componentes también se contemplan como "que comprenden" o "que consisten esencialmente en" los componentes mencionados; y las realizaciones de la memoria descriptiva que mencionan "que consisten esencialmente en" varios componentes también se contemplan como "que consisten en" o "que comprenden" los componentes mencionados (esta intercambiabilidad no se aplica al uso de estos términos en las reivindicaciones).

Los encabezamientos de sección usados en la presente tienen propósitos organizativos y no deben interpretarse de ninguna manera como una limitación de la materia divulgada. En caso de que cualquier documento u otro material contradiga cualquier contenido explícito de esta memoria descriptiva, incluyendo las definiciones,

prevalecerá esta memoria descriptiva.

I. DEFINICIONES

"ADN libre de células", "moléculas de ADNcf" o simplemente "ADNcf" incluyen moléculas de ADN que se encuentran en un sujeto en forma extracelular (por ejemplo, en la sangre, el suero, el plasma u otros fluidos corporales como la linfa, el líquido cefalorraquídeo, la orina o el esputo) e incluye el ADN que no está contenido dentro o que no está unido de otro modo a una célula. Aunque el ADN existía originalmente en una célula o células de un gran organismo biológico complejo, por ejemplo, un mamífero, el ADN se ha liberado de la célula o células a un fluido que se encuentra en el organismo. Típicamente, el ADNcf puede obtenerse mediante la obtención de una muestra del fluido sin la necesidad de realizar un paso de lisis celular in vitro y también incluye la eliminación de las células presentes en el fluido (por ejemplo, centrifugación de la sangre para eliminar las células).

El "rendimiento de captura" de una colección de sondas para un conjunto de regiones diana dado se refiere a la cantidad (por ejemplo, cantidad relativa a otro conjunto de regiones diana o una cantidad absoluta) de ácido nucleico correspondiente al conjunto de regiones diana que la colección captura en condiciones típicas. Las condiciones de captura típicas ejemplares son una incubación del ácido nucleico de la muestra y las sondas a 65°C durante 10-18 horas en un volumen de reacción pequeño (aproximadamente 20 µl) que contenga tampón de hibridación estricto. El rendimiento de captura puede expresarse en términos absolutos o, para una pluralidad de colecciones de sondas, en términos relativos. Cuando se comparan los rendimientos de captura para una pluralidad de conjuntos de regiones diana, se normalizan para el tamaño de la huella del conjunto de regiones diana (por ejemplo, en una base kilobases). Por tanto, por ejemplo, si los tamaños de la huella de la primera y la segunda regiones diana son de 50 kb y 500 kb, respectivamente (lo que da un factor de normalización de 0,1), entonces el ADN correspondiente al primer conjunto de regiones diana se captura con un rendimiento mayor que el del ADN correspondiente al segundo conjunto de regiones diana cuando la concentración de masa por volumen del ADN capturado correspondiente al primer conjunto de regiones diana es más de 0,1 veces la concentración de masa por volumen del ADN capturado correspondiente al segundo conjunto de regiones diana. Como ejemplo adicional, usando los mismos tamaños de huella, si el ADN capturado correspondiente al primer conjunto de regiones diana tiene una concentración de masa por volumen de 0,2 veces la concentración de masa por volumen del ADN capturado correspondiente al segundo conjunto de regiones diana, entonces el ADN correspondiente al primer conjunto de regiones diana se capturó con un rendimiento de captura dos veces mayor que el ADN correspondiente al segundo conjunto de regiones diana.

"Capturar" o "enriquecer" uno o más ácidos nucleicos diana se refiere a aislar o separar preferencialmente el uno o más ácidos nucleicos diana de los ácidos nucleicos no diana.

Un "conjunto capturado" de ácidos nucleicos se refiere a ácidos nucleicos que han sido sometidos a captura.

Un "conjunto de regiones diana" o "regiones diana" se refiere a una pluralidad de loci genómicos o una pluralidad de regiones genómicas objeto de captura y/u objetivo de un conjunto de sondas (por ejemplo, mediante complementariedad de secuencias).

"Correspondiente a un conjunto de regiones diana" significa que un ácido nucleico, como el ADNcf, procedente de un locus del conjunto de regiones diana o se une específicamente a una o más sondas del conjunto de regiones diana.

"Se une específicamente" en el contexto de una sonda u otro oligonucleótido y una secuencia diana significa que, en condiciones de hibridación adecuadas, el oligonucleótido o la sonda hibrida con su secuencia diana, o réplicas de la misma, para formar un híbrido sonda:diana estable, a la vez que se minimiza la formación de híbridos estables sonda:no diana. Por tanto, una sonda hibrida con una secuencia diana o una réplica de la misma en una medida suficientemente mayor que con una secuencia no diana, para permitir la captura o detección de la secuencia diana. Las condiciones de hibridación adecuadas son bien conocidas en la técnica, pueden predecirse basándose en la composición de la secuencia, o pueden determinarse usando métodos de prueba rutinarios (véase, por ejemplo, Sambrook et al., Molecular Cloning, A Laboratory Manual, 2ª ed. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989) en §§ 1.90-1.91, 7.37-7.57, 9.47-9.51 y 11.47-11.57, en particular §§ 9.50-9.51, 11.12-11.13, 11.45-11.47 y 11.55-11.57).

"Conjunto de regiones diana de secuencia variable" se refiere a un conjunto de regiones diana que pueden presentar cambios en la secuencia como sustituciones, inserciones, deleciones de nucleótidos o fusiones o transposiciones de genes en células neoplásicas (por ejemplo, células tumorales y células cancerosas).

"Conjunto de regiones diana epigenéticas" se refiere a un conjunto de regiones diana que pueden manifestar modificaciones no de secuencia en células neoplásicas (por ejemplo, células tumorales y células cancerosas) y células no tumorales (por ejemplo, células inmunitarias, células del microambiente tumoral). El conjunto de regiones diana epigenéticas se selecciona entre una o más de i) un conjunto de regiones diana variables de hipermetilación; ii) un conjunto de regiones diana variables de hipometilación; y/o iii) un conjunto de regiones diana variables de

fragmentación, en donde el conjunto de regiones diana variables de fragmentación comprende regiones de sitios de inicio de transcripción y/o regiones de unión a CTCF; opcionalmente, en donde el conjunto de regiones diana epigenéticas comprende además un conjunto de regiones diana de control de metilación. Estas modificaciones no cambian la secuencia del ADN. Ejemplos de cambios de modificaciones no de secuencia incluyen, pero no se limitan a, cambios en la metilación (aumentos o disminuciones), distribución de nucleosomas, unión a CTCF, sitios de inicio de transcripción, regiones de unión a proteínas reguladoras y cualquier otra proteína que pueda unirse al ADN. La detección de un cambio en el número de copias mediante secuenciación o una secuencia fusionada que mapea para más de un locus en un genoma de referencia tiende a ser más similar a la detección de los cambios epigenéticos ejemplares comentados anteriormente que la detección de sustituciones, inserciones o deleciones de nucleótidos, por ejemplo, en que las amplificaciones focales y/o las fusiones génicas pueden detectarse a una profundidad de secuenciación relativamente baja porque su detección no depende de la precisión de las llamadas de bases en una o unas pocas posiciones individuales. En la presente, los términos "epigenético" y "epigenómico" se usan indistintamente.

Un ADN tumoral circulante o ADNct es un componente del ADNcf que se originó a partir de una célula tumoral o célula cancerosa. En algunas realizaciones, el ADNcf comprende ADN originado a partir de células normales y ADN originado a partir de células tumorales (es decir, ADNct). Las células tumorales son células neoplásicas que se originaron a partir de un tumor, independientemente de si permanecen en el tumor o se separan del tumor (como en los casos, por ejemplo, de células cancerosas metastásicas y células tumorales circulantes).

El término "hipermetilación" se refiere a un nivel o grado de metilación aumentados de molécula o moléculas de ácido nucleico con respecto a otras moléculas de ácido nucleico dentro de una población (por ejemplo, muestra) de moléculas de ácido nucleico. En algunas realizaciones, el ADN hipermetilado puede incluir moléculas de ADN que comprenden por lo menos 1 residuo metilado, por lo menos 2 residuos metilados, por lo menos 3 residuos metilados, por lo menos 5 residuos metilados, por lo menos 10 residuos metilados, por lo menos 20 residuos metilados, por lo menos 25 residuos metilados o por lo menos 30 residuos metilados.

El término "hipometilación" se refiere a un nivel o grado disminuido de metilación de molécula o moléculas de ácido nucleico con respecto a las otras moléculas de ácido nucleico dentro de una población (por ejemplo, muestra) de moléculas de ácido nucleico. En algunas realizaciones, el ADN hipometilado incluye moléculas de ADN no metiladas. En algunas realizaciones, el ADN hipometilado puede incluir moléculas de ADN que comprenden 0 residuos metilados, como máximo 1 residuo metilado, como máximo 2 residuos metilados, como máximo 3 residuos metilados, como máximo 4 residuos metilados, o como máximo 5 residuos metilados.

Los términos "o una combinación de los mismos" y "o combinaciones de los mismos" como se usan en la presente se refieren a todas y cada una de las permutaciones y combinaciones de los términos enumerados que preceden al término. Por ejemplo, "A, B, C, o combinaciones de los mismos", se pretende que incluya por lo menos uno de: A, B, C, AB, AC, BC, o ABC, y si el orden es importante en un contexto particular, también BA, CA, CB, ACB, CBA, BCA, BAC, o CAB. Continuando con este ejemplo, se incluyen expresamente las combinaciones que contienen repeticiones de uno o más elementos o términos, como BB, AAA, AAB, BBC, AAABCCCC, CBBAAA, CABABB, y demás. El experto en la técnica entenderá que típicamente no hay límite en el número de elementos o términos en cualquier combinación, a menos que sea evidente lo contrario por el contexto.

"O" se usa en sentido inclusivo, es decir, equivalente a "y/o", a menos que el contexto requiera lo contrario.

II. MÉTODOS EJEMPLARES

En la presente se proporcionan métodos para aislar ADN libre de células (ADNcf) y/o identificar la presencia de ADN producido por un tumor (o células neoplásicas, o células cancerosas).

En algunas realizaciones, los métodos comprenden la captura de ADNcf obtenido de un sujeto de prueba para una pluralidad de conjuntos de regiones diana. Las regiones diana comprenden regiones diana epigenéticas, que pueden mostrar diferencias en los niveles de metilación y/o patrones de fragmentación dependiendo de si se originaron en un tumor o en células sanas. Las regiones diana también comprenden regiones diana de secuencia variable, que pueden mostrar diferencias en la secuencia dependiendo de si se originaron en un tumor o en células sanas. El paso de captura produce un conjunto capturado de moléculas de ADNcf, y las moléculas de ADNcf correspondientes al conjunto de regiones diana de secuencia variable se capturan con un mayor rendimiento de captura en el conjunto capturado de moléculas de ADNcf que las moléculas de ADNcf correspondientes al conjunto de regiones diana epigenéticas.

En algunas realizaciones, los métodos comprenden poner en contacto ADNcf obtenido de un sujeto de prueba con un conjunto de sondas específicas de la diana, en donde el conjunto de sondas específicas de la diana está configurado para capturar ADNcf correspondiente al conjunto de regiones diana de secuencia variable con un mayor rendimiento de captura que el ADNcf correspondiente al conjunto de regiones diana epigenéticas.

Puede ser beneficioso capturar ADNcf correspondiente al conjunto de regiones diana de secuencia variable con un mayor rendimiento de captura que el ADNcf correspondiente al conjunto de regiones diana epigenéticas porque puede ser necesaria una mayor profundidad de secuenciación para analizar las regiones diana de secuencia variable con suficiente confianza o precisión que la necesaria para analizar las regiones diana epigenéticas. La mayor profundidad de secuenciación puede dar como resultado más lecturas por molécula de ADN y puede facilitarse capturando más moléculas únicas por región. El volumen de datos necesario para determinar los patrones de fragmentación (por ejemplo, para comprobar la perturbación de los sitios de inicio de la transcripción o los sitios de unión a CTCF) o la abundancia de fragmentos (por ejemplo, en divisiones hipermetiladas e hipometiladas) es generalmente menor que el volumen de datos necesario para determinar la presencia o ausencia de mutaciones de secuencia relacionadas con el cáncer. La captura de los conjuntos de regiones diana en diferentes rendimientos puede facilitar la secuenciación de las regiones diana a diferentes profundidades de secuenciación en la misma serie de secuenciación (por ejemplo, usando una mezcla agrupada y/o en la misma celda de secuenciación).

En varias realizaciones, los métodos comprenden además secuenciar el ADNcf capturado, por ejemplo, a diferentes grados de profundidad de secuenciación para los conjuntos de regiones diana epigenéticas y de secuencia variable, de acuerdo con el análisis anterior.

1. Paso de captura; amplificación; adaptadores; códigos de barras

En algunas realizaciones, los métodos divulgados comprenden un paso de captura de uno o más conjuntos de regiones diana de ADN, como ADNcf. La captura puede realizarse usando cualquier enfoque adecuado conocido en la técnica.

En algunas realizaciones, la captura comprende poner en contacto el ADN a capturar con un conjunto de sondas específicas de la diana. El conjunto de sondas específicas de la diana puede tener cualquiera de las características descritas en la presente para conjuntos de sondas específicas de la diana, incluyendo pero no limitadas a, las realizaciones expuestas anteriormente y las secciones relativas a sondas a continuación.

El paso de captura puede realizarse usando condiciones adecuadas para la hibridación específica de ácidos nucleicos, que generalmente dependen en cierta medida de las características de las sondas, como la longitud, la composición de bases, etc. Los expertos en la técnica estarán familiarizados con las condiciones apropiadas dado el conocimiento general en la técnica referente a la hibridación de ácidos nucleicos. En algunas realizaciones, se forman complejos de sondas específicas de la diana y ADN.

En algunas realizaciones, los complejos de sondas específicas de la diana y ADN se separan del ADN no unido a las sondas específicas de la diana. Por ejemplo, cuando las sondas específicas de la diana están unidas covalentemente o no covalentemente a un soporte sólido, puede usarse un paso de lavado o aspiración para separar el material no unido. Alternativamente, cuando los complejos tienen propiedades cromatográficas distintas del material no unido (por ejemplo, cuando las sondas comprenden un ligando que se une a una resina cromatográfica), puede usarse cromatografía.

Como se analiza en detalle en otra parte de la presente, el conjunto de sondas específicas de la diana puede comprender una pluralidad de conjuntos como sondas para un conjunto de regiones diana de secuencia variable y sondas para un conjunto de regiones diana epigenéticas. En algunas de tales realizaciones, el paso de captura se lleva a cabo con las sondas para el conjunto de regiones diana de secuencia variable y las sondas para el conjunto de regiones diana epigenéticas en el mismo recipiente al mismo tiempo, por ejemplo, las sondas para los conjuntos de regiones diana de secuencia variable y epigenéticas están en la misma composición. Este enfoque proporciona un flujo de trabajo relativamente racionalizado. En algunas realizaciones, la concentración de las sondas para el conjunto de regiones diana de secuencia variable es mayor que la concentración de las sondas para el conjunto de regiones diana epigenéticas.

Alternativamente, el paso de captura se realiza con el conjunto de sondas de regiones diana de secuencia variable en un primer recipiente y con el conjunto de sondas de regiones diana epigenéticas en un segundo recipiente, o el paso de contacto se realiza con el conjunto de sondas de regiones diana de secuencia variable en un primer momento y en un primer recipiente y con el conjunto de sondas de regiones diana epigenéticas en un segundo momento antes o después del primer momento. Este enfoque permite la preparación de la primera y la segunda composiciones separadas que comprenden ADN capturado correspondiente al conjunto de regiones diana de secuencia variable y ADN capturado correspondiente al conjunto de regiones diana epigenéticas. Las composiciones pueden procesarse por separado según se desee (por ejemplo, para fraccionar sobre la base de la metilación, como se describe en otra parte de la presente) y recombinarse en proporciones adecuadas para proporcionar material para su procesamiento y análisis adicionales, como la secuenciación.

En algunas realizaciones, el ADN se amplifica. En algunas realizaciones, la amplificación se realiza antes del paso de captura. En algunas realizaciones, la amplificación se realiza después del paso de captura. Los métodos para la amplificación inespecífica del ADN son conocidos en la técnica. Véase, por ejemplo, Smallwood et al, Nat. Methods

11: 817-820 (2014). Por ejemplo, pueden usarse cebadores aleatorios que tengan secuencias adaptadoras en sus extremos 5' y bases aleatorias en el extremo 3'. Habitualmente tienen 6 bases aleatorias, pero pueden tener entre 4 y 9 bases. Este enfoque es adecuado para la amplificación de células con baja cantidad entrada/individuales y/o la secuenciación por bisulfito.

En algunas realizaciones, en el ADN se incluyen adaptadores. Esto puede hacerse simultáneamente con un procedimiento de amplificación, por ejemplo, proporcionando los adaptadores en una parte 5' de un cebador, por ejemplo, como se ha descrito anteriormente. Alternativamente, los adaptadores pueden añadirse mediante otros enfoques, como la ligadura.

En algunas realizaciones, en el ADN se incluyen etiquetas, que pueden ser o incluir códigos de barras. Las etiquetas pueden facilitar la identificación del origen de un ácido nucleico. Por ejemplo, los códigos de barras pueden usarse para permitir la identificación del origen (por ejemplo, el sujeto) de donde procede el ADN tras la agrupación de una pluralidad de muestras para secuenciación paralela. Esto puede hacerse concurrentemente con un procedimiento de amplificación, por ejemplo, proporcionando los códigos de barras en una parte 5' de un cebador, por ejemplo, como se ha descrito anteriormente. En algunas realizaciones, los adaptadores y las etiquetas/códigos de barras se proporcionan mediante el mismo cebador o conjunto de cebadores. Por ejemplo, el código de barras puede estar situado 3' del adaptador y 5' de la parte de hibridación con la diana del cebador. Alternativamente, los códigos de barras pueden añadirse mediante otros enfoques, como la ligación, opcionalmente junto con los adaptadores en el mismo sustrato de ligación.

Detalles adicionales referentes a la amplificación, etiquetas y códigos de barras se analizan en la sección "Características Generales de los Métodos" a continuación, que pueden combinarse en la medida de lo posible con cualquiera de las realizaciones anteriores y las realizaciones expuestas en la sección de introducción y sumario.

2. Conjunto capturado

En algunas realizaciones, se proporciona un conjunto capturado de ADN (por ejemplo, ADNcf). Con respecto a los métodos divulgados, el conjunto capturado de ADN puede ser proporcionado, por ejemplo, después de los pasos de captura, y/o de separación como se describe en la presente. El conjunto capturado puede comprender ADN correspondiente a un conjunto de regiones diana de secuencia variable y un conjunto de regiones diana epigenéticas. En algunas realizaciones, la cantidad de ADN de región diana de secuencia variable capturado es mayor que la cantidad de ADN de región diana epigenética capturado, cuando se normaliza para la diferencia en el tamaño de las regiones diana (tamaño de huella).

Alternativamente, pueden proporcionarse el primer y el segundo conjuntos capturados, que comprenden, respectivamente, ADN correspondiente a un conjunto de regiones diana de secuencia variable y ADN correspondiente a un conjunto de regiones diana epigenéticas. El primer y el segundo conjuntos capturados pueden combinarse para proporcionar un conjunto capturado combinado.

En un conjunto capturado que comprende ADN correspondiente al conjunto de regiones diana de secuencia variable y al conjunto de regiones diana epigenéticas, incluyendo un conjunto capturado combinado como se ha analizado anteriormente, el ADN correspondiente al conjunto de regiones diana de secuencia variable puede estar presente en una concentración mayor que el ADN correspondiente al conjunto de regiones diana epigenéticas, por ejemplo, una concentración de 1,1 a 1,2 veces más, una concentración de 1,2 a 1,4 veces más, una concentración de 1,4 a 1,6 veces más, una concentración de 1,6 a 1,8 veces más, una concentración de 1,8 a 2,0 veces más, una concentración de 2,0 a 2,2 veces más, una concentración, de 2,2 a 2,4 veces más, una concentración de 2,4 a 2,6 veces más, una concentración, de 2,6 a 2,8 veces más, una concentración de 2,8 a 3,0 veces más, una concentración de 3,0 a 3,5 veces más, una concentración de 3,5 a 4 veces más, una concentración de 4,0 a 4,5 veces más, una concentración de 4,5 a 5 veces más, una concentración de 5,0 a 5,5 veces más, una concentración de 5,5 a 6 veces más, una concentración de 6,0 a 6,5 veces más, una concentración de 6,5 a 7,0 veces más, una concentración de 7,0 a 7,5 veces más, una concentración de 7,5 a 8,0 veces más, una concentración de 8,0 a 8,5 veces más, una concentración de 8,5 a 9,0 veces más, una concentración de 9,0 a 9,5 veces más, una concentración de 9,5 a 10,0 veces más, una concentración de 10 a 11 veces más, una concentración de 11 a 12 veces más, una concentración de 12 a 13 veces más, una concentración de 13 a 14 veces más, una concentración de 14 a 15 veces más, una concentración de 15 a 16 veces más, una concentración de 16 a 17 veces más, una concentración de 17 a 18 veces más, una concentración de 18 a 19 veces más, o una concentración de 19 a 20 veces más. El grado de diferencia en las concentraciones tiene en cuenta la normalización para los tamaños de huella de las regiones diana, como se analiza en la sección de definiciones.

a. Conjunto de regiones diana epigenéticas

El conjunto de regiones diana epigenéticas puede comprender uno o más tipos de regiones diana susceptibles de diferenciar el ADN de células neoplásicas (por ejemplo, tumorales o cancerosas) y de células sanas, por ejemplo, células circulantes no neoplásicas. El conjunto de regiones diana epigenéticas se selecciona entre una o

más de:

- i) un conjunto de regiones diana variables de hipermetilación;
- ii) un conjunto de regiones diana variables de hipometilación; y/o
- iii) un conjunto de regiones diana variables de fragmentación, en donde el conjunto de regiones diana variables de fragmentación comprende regiones de sitios de inicio de transcripción y/o regiones de unión a CTCF;

opcionalmente, en donde el conjunto de regiones diana epigenéticas comprende además un conjunto de regiones diana de control de metilación.

En la presente se analizan en detalle tipos ejemplares de tales regiones. En algunas realizaciones, los métodos de acuerdo con la divulgación comprenden determinar si las moléculas de ADNcf correspondientes al conjunto de regiones diana epigenéticas comprenden o indican modificaciones epigenéticas asociadas al cáncer (por ejemplo, hipermetilación en una o más regiones diana variables de hipermetilación; una o más perturbaciones de la unión a CTCF; y/o una o más perturbaciones de los sitios de inicio de la transcripción). El conjunto de regiones diana epigenéticas también puede comprender una o más regiones de control, por ejemplo, como se describe en la presente.

En algunas realizaciones, el conjunto de regiones diana epigenéticas tiene una huella de por lo menos 100 kb, por ejemplo, por lo menos 200 kb, por lo menos 300 kb o por lo menos 400 kb. En algunas realizaciones, el conjunto de regiones diana epigenéticas tiene una huella en el intervalo de 100-1000 kb, por ejemplo, 100-200 kb, 200-300 kb, 300-400 kb, 400-500 kb, 500-600 kb, 600-700 kb, 700-800 kb, 800-900 kb, y 900-1.000 kb.

i. Regiones diana variables de hipermetilación

En algunas realizaciones, el conjunto de regiones diana epigenéticas comprende una o más regiones diana variables de hipermetilación. En general, las regiones diana variables de hipermetilación se refieren a regiones en las que un aumento del nivel de metilación observado indica una mayor probabilidad de que una muestra (por ejemplo, de ADNcf) contenga ADN producido por células neoplásicas, como células tumorales o cancerosas. Por ejemplo, se ha observado repetidamente la hipermetilación de promotores de genes supresores de tumores. Véase, por ejemplo, Kang et al., Genome Biol. 18:53 (2017) y las referencias citadas en el mismo.

En Lam et al., Biochim Biophys Acta. 1866:106-20 (2016), se proporciona un análisis exhaustivo de las regiones diana variables de metilación en el cáncer colorrectal. 1866:106-20 (2016). Estas incluyen VIM, SEPT9, ITGA4, OSM4, GATA4 y NDRG4. En la Tabla 1 se proporciona un conjunto ejemplar de regiones diana variables de hipermetilación que comprenden los genes o partes de los mismos basados en los estudios de cáncer colorrectal (CRC). Es probable que muchos de estos genes tengan relevancia para cánceres más allá del cáncer colorrectal; por ejemplo, TP53 es ampliamente reconocido como un supresor tumoral de importancia crítica y la inactivación de este gen basada en la hipermetilación puede ser un mecanismo oncogénico común.

Tabla 1. Regiones diana de hipermetilación ejemplares (genes o partes de los mismos) basadas en estudios de CRC.

Nombre del Gen	Nombre del gen adicional	Cromosoma
VIM		chr10
SEPT9		chr17
CYCD2	CCND2	chr12
TFPI2		chr7
GATA4		chr8
RARB2	RARB	chr3
p16INK4a	CDKN2A	chr9
MGMT	MGMT	chr10
APC		chr5
NDRG4		chr16
HLTF		chr3
HPP1	TMEFF2	chr2
hMLH1	MLH1	chr3
RASSF1A	RASSF1	chr3
CDH13		chr16
IGFBP3		chr7
ITGA4		chr2

En algunas realizaciones, las regiones diana variables de hipermetilación comprenden una pluralidad de genes o partes de los mismos enumerados en la Tabla 1, por ejemplo, por lo menos el 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, o 100% de los genes o partes de los mismos enumerados en la Tabla 1. Por ejemplo, para cada locus incluido como región diana, puede haber una o más sondas con un sitio de hibridación que se une entre el sitio de inicio de la transcripción y el codón de parada (el último codón de parada para genes que están cortados y empalmados alternativamente) del gen. En algunas realizaciones, la una o más sondas se unen dentro de 300 pb en sentido ascendente y/o en sentido descendente de los genes o partes de los mismos enumerados en la Tabla 1, por ejemplo, dentro de 200 o 100 pb.

Las regiones diana variables de metilación en varios tipos de cáncer de pulmón se analizan en detalle, por ejemplo, en Ooki et al., Clin. Cancer Res. 23:7141-52 (2017); Belinsky, Annu. Rev. Physiol. 77:453- 74 (2015); Hulbert et al., Clin. Cancer Res. 23: 1998-2005 (2017); Shi et al., BMC Genomics 18:901 (2017); Schneider et al., BMC Cancer. 11:102 (2011); Lissa et al., Transl Lung Cancer Res 5(5):492-504 (2016); Skvortsova et al., Br. J. Cancer. 94(10):1492-1495 (2006); Kim et al., Cancer Res. 61:3419-3424 (2001); Furonaka et al., Pathology International 55:303-309 (2005); Gomes et al., Rev. Port. Pneumol. 20:20-30 (2014); Kim et al., Oncogene. 20:1765-70 (2001); Hopkins-Donaldson et al., Cell Death Differ. 10:356-64 (2003); Kikuchi et al., Clin. Cancer Res. 11:2954-61 (2005); Heller et al., Oncogene 25:959-968 (2006); Licchesi et al., Carcinogenesis. 29:895-904 (2008); Guo et al., Clin. Cancer Res. 10:7917-24 (2004); Palmisano et al., Cancer Res. 63:4620-4625 (2003); y Toyooka et al., Cancer Res. 61:4556-4560, (2001).

En la Tabla 2 se proporciona un conjunto ejemplar de regiones diana variables de hipermetilación que comprenden genes o partes de los mismos basados en los estudios de cáncer de pulmón. Es probable que muchos de estos genes tengan relevancia para cánceres además del cáncer de pulmón, por ejemplo, Casp8 (Caspasa 8) es una enzima clave en la muerte celular programada y la inactivación basada en la hipermetilación de este gen puede ser un mecanismo oncogénico común no limitado al cáncer de pulmón. Además, una serie de genes aparecen tanto en la Tabla 1 como en la Tabla 2, lo que indica generalidad.

Tabla 2. Regiones diana de hipermetilación ejemplares (genes o partes de los mismos) basadas en estudios de cáncer de pulmón.

Nombre del Gen	Cromosoma
MARCH11	chr5
TAC1	chr7
TCF21	chr6
SHOX2	chr3
p16	chr3
Casp8	chr2
CDH13	chr16
MGMT	chr10
MLH1	chr3
MSH2	chr2
TSLC1	chr11
APC	chr5
DKK1	chr10
DKK3	chr11
LKB1	chr11
WIF1	chr12
RUNX3	chr1
GATA4	chr8
GATA5	chr20
PAX5	chr9
E-Cadherina	chr16
H-Cadherina	chr16

Cualquiera de las realizaciones anteriores relativas a las regiones diana identificadas en la Tabla 2 puede combinarse con cualquiera de las realizaciones descritas anteriormente relativas a las regiones diana identificadas en la Tabla 1. En algunas realizaciones, las regiones diana variables de hipermetilación comprenden una pluralidad de genes o partes de los mismos enumerados en la Tabla 1 o Tabla 2, por ejemplo, por lo menos el 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, o 100% de los genes o partes de los mismos enumerados en la Tabla 1 o la Tabla 2.

Pueden obtenerse regiones diana de hipermetilación adicionales, por ejemplo, del Cancer Genome Atlas. Kang et al., Genome Biology 18:53 (2017), describen la construcción de un método probabilístico denominado Cancer Locator usando regiones diana de hipermetilación de mama, colon, riñón, hígado y pulmón. En algunas realizaciones, las regiones diana de hipermetilación pueden ser específicas de uno o más tipos de cáncer. Por consiguiente, en algunas realizaciones, las regiones diana de hipermetilación incluyen uno, dos, tres, cuatro o cinco subconjuntos de regiones diana de hipermetilación que muestran colectivamente hipermetilación en uno, dos, tres, cuatro o cinco de los cánceres de mama, colon, riñón, hígado y pulmón.

ii. Regiones diana variables de hipometilación

La hipometilación global es un fenómeno comúnmente observado en varios tipos de cáncer. Véase, por ejemplo, Hon et al., Genome Res. 22:246-258 (2012) (cáncer de mama); Ehrlich, Epigenomics 1:239-259 (2009)

(artículo de revisión que señala observaciones de hipometilación en cánceres de colon, ovario, próstata, leucemia, hepatocelular y cervical). Por ejemplo, regiones como elementos repetidos, por ejemplo, elementos LINE1, elementos Alu, repeticiones en tándem centroméricas, repeticiones en tándem pericentroméricas y ADN satélite, y regiones intergénicas que normalmente están metiladas en células sanas pueden mostrar una metilación reducida en células tumorales. Por consiguiente, en algunas realizaciones, el conjunto de regiones diana epigenéticas incluye regiones diana variables de hipometilación, en donde una disminución del nivel de metilación observado indica una mayor probabilidad de que una muestra (por ejemplo, de ADNcf) contenga ADN producido por células neoplásicas, como células tumorales o cancerosas.

En algunas realizaciones, las regiones diana variables de hipometilación incluyen elementos repetidos y/o regiones intergénicas. En algunas realizaciones, los elementos repetidos incluyen uno, dos, tres, cuatro o cinco elementos LINE1, elementos Alu, repeticiones en tándem centroméricas, repeticiones en tándem pericentroméricas y/o ADN satélite.

Las regiones genómicas específicas ejemplares que muestran hipometilación asociada al cáncer incluyen los nucleótidos 8403565-8953708 y 151104701-151106035 del cromosoma humano 1, por ejemplo, de acuerdo con el constructo del genoma humano hg19 o hg38. En algunas realizaciones, las regiones diana variables de hipometilación se solapan o comprenden una o ambas de estas regiones.

iii. Regiones de unión a CTCF

La CTCF es una proteína de unión al ADN que contribuye a la organización de la cromatina y a menudo está colocalizada con la cohesina. La perturbación de los sitios de unión de CTCF se ha informado en una variedad de diferentes cánceres. Véanse, por ejemplo, Katainen et al., *Nature Genetics*, doi: 10.1038/ng.3335, publicado en línea el 8 de junio de 2015; Guo et al., *Nat. Commun.* 9: 1520 (2018). La unión de CTCF da como resultado patrones reconocibles en ADNcf que pueden detectarse mediante secuenciación, por ejemplo, a través del análisis de longitud de fragmentos. Por ejemplo, se proporcionan detalles con respecto al análisis de longitud de fragmentos basado en secuenciación en Snyder et al., *Cell* 164:57-68 (2016); WO 2018/009723; y US20170211143A1.

Por tanto, las perturbaciones de la unión a CTCF dan como resultado variaciones en los patrones de fragmentación del ADNcf. Como tal, los sitios de unión de CTCF representan un tipo de regiones diana variables de fragmentación.

Se conocen muchos sitios de unión a CTCF. Véase, por ejemplo, la CTCFBSDB (Base de datos de sitios de unión a CTCF), disponible en Internet en insulatordb.uthsc.edu/; Cuddapah et al., *Genome Res.* 19:24-32 (2009); Martin et al., *Nat. Struct. Mol. Biol.* 18:708-14 (2011); Rhee et al., *Cell*. 147:1408-19 (2011). Los sitios de unión a CTCF ejemplares se encuentran en los nucleótidos 56014955-56016161 en el cromosoma 8 y en los nucleótidos 95359169-95360473 en el cromosoma 13, por ejemplo, de acuerdo con el constructo del genoma humano hg19 o hg38.

Por consiguiente, en algunas realizaciones, el conjunto de regiones diana epigenéticas incluye regiones de unión a CTCF. En algunas realizaciones, las regiones de unión a CTCF comprenden por lo menos 10, 20, 50, 100, 200 o 500 regiones de unión a CTCF, o 10-20, 20-50, 50-100, 100-200, 200-500 o 500-1000 regiones de unión a CTCF, por ejemplo, como las regiones de unión a CTCF descritas anteriormente o en uno o más de CTCFBSDB o los artículos de Cuddapah et al., Martin et al. o Rhee et al. citados anteriormente.

En algunas realizaciones, por lo menos algunos de los sitios CTCF pueden estar metilados o no metilados, en donde el estado de metilación está correlacionado con si la célula es o no una célula cancerosa. En algunas realizaciones, el conjunto de regiones diana epigenéticas comprende por lo menos 100 pb, por lo menos 200 pb, por lo menos 300 pb, por lo menos 400 pb, por lo menos 500 pb, por lo menos 750 pb, por lo menos 1000 pb de regiones en sentido ascendente y/o en sentido descendente de los sitios de unión a CTCF.

iv. Sitios de inicio de la transcripción

Los sitios de inicio de transcripción también pueden mostrar perturbaciones en células neoplásicas. Por ejemplo, la organización de nucleosomas en varios sitios de inicio de transcripción en células sanas del linaje hematopoyético, que contribuye sustancialmente al ADNcf en individuos sanos, puede diferir de la organización de nucleosomas en esos sitios de inicio de transcripción en células neoplásicas. Esto da como resultado a diferentes patrones de ADNcf que pueden detectarse mediante secuenciación, por ejemplo, como se analiza de manera general en Snyder et al., *Cell* 164:57-68 (2016); WO 2018/009723; y US20170211143A1.

Por tanto, las perturbaciones de los sitios de inicio de la transcripción también dan como resultado variaciones en los patrones de fragmentación del ADNcf. Como tales, los sitios de inicio de la transcripción también representan un tipo de regiones diana variables de fragmentación.

Los sitios de inicio transcripcional humano están disponibles en DBTSS (DataBase of Human Transcription

Start Sites), disponible en Internet en dbtss.hgc.jp y descritos en Yamashita et al., Nucleic Acids Res. 34(Database issue): D86-D89 (2006).

Por consiguiente, en algunas realizaciones, el conjunto de regiones diana epigenéticas incluye sitios de inicio transcripcional. En algunas realizaciones, los sitios de inicio transcripcional comprenden por lo menos 10, 20, 50, 100, 200, o 500 sitios de inicio transcripcional, o 10-20, 20-50, 50-100, 100-200, 200-500, o 500-1000 sitios de inicio transcripcional, por ejemplo, tales como sitios de inicio transcripcional listados en DBTSS. En algunas realizaciones, por lo menos algunos de los sitios de inicio de la transcripción pueden estar metilados o no metilados, en donde el estado de metilación está correlacionado con el hecho de que la célula sea o no cancerosa. En algunas realizaciones, el conjunto de regiones diana epigenéticas comprende por lo menos 100 pb, por lo menos 200 pb, por lo menos 300 pb, por lo menos 400 pb, por lo menos 500 pb, por lo menos 750 pb, por lo menos 1000 pb de regiones en sentido ascendente y/o en sentido descendente de los sitios de inicio de transcripción.

v. Variaciones del número de copias; amplificaciones focales

Aunque las variaciones en el número de copias, como las amplificaciones focales, son mutaciones somáticas, pueden detectarse mediante secuenciación basada en la frecuencia de lectura de manera análoga a los enfoques para detectar ciertos cambios epigenéticos, como los cambios en la metilación.

vi. Regiones de control de metilación

Puede ser útil incluir regiones de control para facilitar la validación de los datos. En algunas realizaciones, el conjunto de regiones diana epigenéticas incluye regiones de control que se espera que estén metiladas o no metiladas en esencialmente todas las muestras, independientemente de si el ADN se deriva de una célula cancerosa o de una célula normal. En algunas realizaciones, el conjunto de regiones diana epigenéticas incluye regiones hipometiladas de control que se espera que estén hipometiladas en esencialmente todas las muestras. En algunas realizaciones, el conjunto de regiones diana epigenéticas incluye regiones hipermetiladas de control que se espera que estén hipermetiladas en esencialmente todas las muestras.

b. Conjunto de regiones diana de secuencia variable

En algunas realizaciones, el conjunto de regiones diana de secuencia variable comprende una pluralidad de regiones que se sabe que experimentan mutaciones somáticas en el cáncer (denominadas en la presente mutaciones asociadas al cáncer). Por consiguiente, los métodos pueden comprender determinar si las moléculas de ADNcf correspondientes al conjunto de regiones diana de secuencia variable comprenden mutaciones asociadas al cáncer.

En algunas realizaciones, el conjunto de regiones diana de secuencia variable se dirige a una pluralidad de genes o regiones genómicas diferentes ("panel") seleccionados de tal manera que una proporción determinada de sujetos que tienen un cáncer presenta una variante genética o marcador tumoral en uno o más genes o regiones genómicas diferentes del panel. El panel puede seleccionarse para limitar una región de secuenciación a un número fijo de pares de bases. El panel puede seleccionarse para secuenciar una cantidad deseada de ADN, por ejemplo, ajustando la afinidad y/o la cantidad de las sondas como se describe en otra parte de la presente. El panel puede seleccionarse para lograr una profundidad de lectura de secuencia o una cobertura de lectura de secuencia deseadas para una cantidad de pares de bases secuenciados. El panel puede seleccionarse para alcanzar una sensibilidad teórica, una especificidad teórica y/o una precisión teórica para detectar una o más variantes genéticas en una muestra.

Las sondas para detectar el panel de regiones pueden incluir aquellas para detectar regiones genómicas de interés (regiones de puntos críticos) así como sondas conscientes de nucleosomas (por ejemplo, codones 12 y 13 de KRAS) y pueden diseñarse para optimizar la captura basándose en el análisis de la cobertura de ADNcf y la variación del tamaño del fragmento afectada por los patrones de unión de nucleosomas y la composición de la secuencia de GC. Las regiones usadas en la presente también pueden incluir regiones no críticas optimizadas sobre la base de las posiciones de los nucleosomas y los modelos de GC.

En la Tabla 3 y en la Tabla 4 pueden encontrarse ejemplos de listados de localizaciones genómicas de interés. En algunas realizaciones, un conjunto de regiones diana de secuencia variable usado en los métodos de la presente divulgación comprende por lo menos una parte de por lo menos 5, por lo menos 10, por lo menos 15, por lo menos 20, por lo menos 25, por lo menos 30, por lo menos 35, por lo menos 40, por lo menos 45, por lo menos 50, por lo menos 55, por lo menos 60, por lo menos 65 o 70 de los genes de la Tabla 3. En algunas realizaciones, un conjunto de regiones diana de secuencia variable usado en los métodos de la presente divulgación comprende por lo menos 5, por lo menos 10, por lo menos 15, por lo menos 20, por lo menos 25, por lo menos 30, por lo menos 35, por lo menos 40, por lo menos 45, por lo menos 50, por lo menos 55, por lo menos 60, por lo menos 65 o 70 de los SNV de la Tabla 3. En algunas realizaciones, un conjunto de regiones diana de secuencia variable usado en los métodos de la presente divulgación comprende por lo menos 1, por lo menos 2, por lo menos 3, por lo menos 4, por lo menos 5 o 6 de las fusiones de la Tabla 3. En algunas realizaciones, un conjunto de regiones diana de secuencia variable usado en los

métodos de la presente divulgación comprende por lo menos una parte de por lo menos 1, por lo menos 2 o 3 de los indeles de la Tabla 3. En algunas realizaciones, un conjunto de regiones diana de secuencia variable usado en los métodos de la presente divulgación comprende por lo menos una parte de por lo menos 5, por lo menos 10, por lo menos 15, por lo menos 20, por lo menos 25, por lo menos 30, por lo menos 35, por lo menos 40, por lo menos 45, por lo menos 50, por lo menos 55, por lo menos 60, por lo menos 65, por lo menos 70, o 73 de los genes de la Tabla 4. En algunas realizaciones, un conjunto de regiones diana de secuencia variable usado en los métodos de la presente divulgación comprende por lo menos 5, por lo menos 10, por lo menos 15, por lo menos 20, por lo menos 25, por lo menos 30, por lo menos 35, por lo menos 40, por lo menos 45, por lo menos 50, por lo menos 55, por lo menos 60, por lo menos 65, por lo menos 70 o 73 de los SNV de la Tabla 4. En algunas realizaciones, un conjunto de regiones diana de secuencia variable usado en los métodos de la presente divulgación comprende por lo menos 1, por lo menos 2, por lo menos 3, por lo menos 4, por lo menos 5 o 6 de las fusiones de la Tabla 4. En algunas realizaciones, un conjunto de regiones diana de secuencia variable usado en los métodos de la presente divulgación comprende por lo menos una parte de por lo menos 1, por lo menos 2, por lo menos 3, por lo menos 4, por lo menos 5, por lo menos 6, por lo menos 7, por lo menos 8, por lo menos 9, por lo menos 10, por lo menos 11, por lo menos 12, por lo menos 13, por lo menos 14, por lo menos 15, por lo menos 16, por lo menos 17, o 18 de los indeles de la Tabla 4. Cada una de estas localizaciones genómicas de interés puede identificarse como una región de estructura principal o región de punto crítico para un panel dado. En la Tabla 5 puede encontrarse un ejemplo de una lista de puntos críticos genómicos de interés. Las coordenadas de la Tabla 5 se basan en el ensamblaje hg19 del genoma humano, pero un experto en la técnica estará familiarizado con otros ensamblajes y podrá identificar conjuntos de coordenadas correspondientes a los exones, intrones, codones, etc. indicados en un ensamblaje de su elección. En algunas realizaciones, un conjunto de regiones diana de secuencia variable usado en los métodos de la presente divulgación comprende por lo menos una parte de por lo menos 1, por lo menos 2, por lo menos 3, por lo menos 4, por lo menos 5, por lo menos 6, por lo menos 7, por lo menos 8, por lo menos 9, por lo menos 10, por lo menos 11, por lo menos 12, por lo menos 13, por lo menos 14, por lo menos 15, por lo menos 16, por lo menos 17, por lo menos 18, por lo menos 19 o por lo menos 20 de los genes de la Tabla 5. Cada región genómica de punto crítico se enumera con varias características, incluido el gen asociado, el cromosoma en el que reside, la posición de inicio y parada del genoma que representa el locus del gen, la longitud del locus del gen en pares de bases, los exones cubiertos por el gen y la característica crítica (por ejemplo, el tipo de mutación) que puede tratar de captar una región genómica de interés dada.

Tabla 3

Mutaciones de punto (SNV) e indeles						Fusiones
AKT1	ALK	APC	AR	ARAF	ARID1A	ALK
ATM	BRAF	BRCA1	BRCA2	CCND1	CCND2	FGFR2
CCNE1	CDH1	CDK4	CDK6	CDKN2A	CDKN2B	FGFR3
CTNNB1	EGFR	ERBB2	ESR1	EZH2	FBXW7	NTRK1
FGFR1	FGFR2	FGFR3	GATA3	GNA11	GNAQ	RET
GNAS	HNF1A	HRAS	IDH1	IDH2	JAK2	ROS1
JAK3	KIT	KRAS	MAP2K1	MAP2K2	MET	
MLH1	MPL	MYC	NF1	NFE2L2	NOTCH1	
NPM1	NRAS	NTRK1	PDGFRA	PIK3CA	PTEN	
PTPN11	RAF1	RB1	RET	RHEB	RHOA	
RIT1	ROS1	SMAD4	SMO	SRC	STK11	
TERT	TP53	TSC1	VHL			

Tabla 4

Mutaciones de punto (SNV) e indeles						Fusiones
AKT1	ALK	APC	AR	ARAF	ARID1A	ALK
ATM	BRAF	BRCA1	BRCA2	CCND1	CCND2	FGFR2
CCNE1	CDH1	CDK4	CDK6	CDKN2A	DDR2	FGFR3
CTNNB1	EGFR	ERBB2	ESR1	EZH2	FBXW7	NTRK1
FGFR1	FGFR2	FGFR3	GATA3	GNA11	GNAQ	RET
GNAS	HNF1A	HRAS	IDH1	IDH2	JAK2	ROS1
JAK3	KIT	KRAS	MAP2K1	MAP2K2	MET	
MLH1	MPL	MYC	NF1	NFE2L2	NOTCH1	
NPM1	NRAS	NTRK1	PDGFRA	PIK3CA	PTEN	
PTPN11	RAF1	RB1	RET	RHEB	RHOA	
RIT1	ROS1	SMAD4	SMO	MAPK1	STK11	
TERT	TP53	TSC1	VHL	MAPK3	MTOR	
NTRK3						

Tabla 5

Gen	Cromosoma	Posición de inicio	Posición de parada	Longitud (bp)	Exones/intrones cubiertos	Característica
ALK	chr2	29446405	29446655	250	intrón 19	Fusión
ALK	chr2	29446062	29446197	135	intrón 20	Fusión
ALK	chr2	29446198	29446404	206	exón 20	Fusión
ALK	chr2	29447353	29447473	120	intrón 19	Fusión
ALK	chr2	29447614	29448316	702	intrón 19	Fusión
ALK	chr2	29448317	29448441	124	exón 19	Fusión
ALK	chr2	29449366	29449777	411	intrón 18	Fusión
ALK	chr2	29449778	29449950	172	exón 18	Fusión
BRAF	chr7	140453064	140453203	139	exón 15	BRAF V600
CTNNB1	chr3	41266007	41266254	247	exón 3	S37
EGFR	chr7	55240528	55240827	299	exones 18 y 19	G719 t delecciones
EGFR	chr7	55241603	55241746	143	exón 20	Inserciones/T790M
EGFR	chr7	55242404	55242523	119	exón 21	L858R
ERBB2	chr17	37880952	37881174	222	exón 20	Inserciones
ESR1	chr6	152419857	152420111	254	exón 10	V534, P535, L536, Y537, D538
FGFR2	chr10	123279482	123279693	211	exón 6	S252

ES 3 013 495 T3

(continuación)

Gen	Cromosoma	Posición de inicio	Posición de parada	Longitud (bp)	Exones/intrones cubiertos	Característica
GATA3	chr10	8111426	8111571	145	exón 5	SS/Indeles
GATA3	chr10	8115692	8116002	310	exón 6	SS/Indeles
GNAS	chr20	57484395	57484488	93	exón 8	R844
IDH1	chr2	209113083	209113394	311	exón 4	R132
IDH2	chr15	90631809	90631989	180	exón 4	R140, R172
KIT	chr4	55524171	55524258	87	exón 1	
KIT	chr4	55561667	55561957	290	exón 2	
KIT	chr4	55564439	55564741	302	exón 3	
KIT	chr4	55565785	55565942	157	exón 4	
KIT	chr4	55569879	55570068	189	exón 5	
KIT	chr4	55573253	55573463	210	exón 6	
KIT	chr4	55575579	55575719	140	exón 7	
KIT	chr4	55589739	55589874	135	exón 8	
KIT	chr4	55592012	55592226	214	exón 9	
KIT	chr4	55593373	55593718	345	exones 10 y 11	557, 559, 560, 576
KIT	chr4	55593978	55594297	319	exones 12 y 13	V654
KIT	chr4	55595490	55595661	171	exón 14	T670, S709
KIT	chr4	55597483	55597595	112	exón 15	D716
KIT	chr4	55598026	55598174	148	exón 16	L783
KIT	chr4	55599225	55599368	143	exón 17	C809, R815, D816, L818, D820, S821F, N822, Y823
KIT	chr4	55602653	55602785	132	exón 18	A829P
KIT	chr4	55602876	55602996	120	exón 19	
KIT	chr4	55603330	55603456	126	exón 20	
KIT	chr4	55604584	55604733	149	exón 21	
KRAS	chr12	25378537	25378717	180	exón 4	A146
KRAS	chr12	25380157	25380356	199	exón 3	Q61
KRAS	chr12	25398197	25398328	131	exón 2	G12/G13
MET	chr7	116411535	116412255	720	exón 13, exón 14, intrón 13, intrón 14	MET exón 14 SS

(continuación)

Gen	Cromosoma	Posición de inicio	Posición de parada	Longitud (bp)	Exones/intrones cubiertos	Característica
NRAS	chr1	115256410	115256609	199	exón 3	Q61
NRAS	chr1	115258660	115258791	131	exón 2	G12/G13
PIK3CA	chr3	178935987	178936132	145	exón 10	E545K
PIK3CA	chr3	178951871	178952162	291	exón 21	H1047R
PTEN	chr10	89692759	89693018	259	exón 5	R130
SMAD4	chr18	48604616	48604849	233	exón 12	D537
TERT	chr5	1294841	1295512	671	promotor	chr5:1295228
TP53	chr17	7573916	7574043	127	exón 11	Q331, R337, R342
TP53	chr17	7577008	7577165	157	exón 8	R273
TP53	chr17	7577488	7577618	130	exón 7	R248
TP53	chr17	7578127	7578299	172	exón 6	R213/Y220
TP53	chr17	7578360	7578564	204	exón 5	R175/Deleciones
TP53	chr17	7579301	7579600	299	exón 4	
				12574 (región diana total)		
				16330 (cobertura de sonda total)		

Además, o alternativamente, hay conjuntos de regiones diana adecuados disponibles en la bibliografía. Por ejemplo, Gale et al., PLoS One 13: e0194630 (2018), describen un panel de 35 dianas génicas relacionadas con el cáncer que pueden usarse como parte o la totalidad de un conjunto de regiones diana de secuencia variable. Estas 35 dianas son AKT1, ALK, BRAF, CCND1, CDK2A, CTNNB1, EGFR, ERBB2, ESR1, FGFR1, FGFR2, FGFR3, FOXL2, GATA3, GNA11, GNAQ, GNAS, HRAS, IDH1, IDH2, KIT, KRAS, MED12, MET, MYC, NFE2L2, NRAS, PDGFRA, PIK3CA, PPP2R1A, PTEN, RET, STK11, TP53, y U2AF1.

En algunas realizaciones, el conjunto de regiones diana de secuencia variable comprende regiones diana de por lo menos 10, 20, 30 o 35 genes relacionados con el cáncer, como los genes relacionados con el cáncer enumerados anteriormente.

3. División; Análisis de las características epigenéticas

Una población de diferentes formas de ácidos nucleicos (por ejemplo, ADN hipermetilado e hipometilado en una muestra, como un conjunto capturado de ADNcf como se describe en la presente) se divide físicamente sobre la base del nivel de metilación antes del análisis, por ejemplo, secuenciación, o etiquetado y secuenciación. Este enfoque puede usarse para determinar, por ejemplo, si las regiones diana epigenéticas variables de hipermetilación muestran hipermetilación característica de las células tumorales o las regiones diana epigenéticas variables de hipometilación muestran hipometilación característica de las células tumorales. Además, al dividir una población heterogénea de ácidos nucleicos, pueden aumentarse las señales raras, por ejemplo, enriqueciendo moléculas raras de ácidos nucleicos que son más prevalentes en una fracción (o división) de la población. Por ejemplo, una variación genética presente en el ADN hipermetilado pero menos (o nada) en el ADN hipometilado puede detectarse más fácilmente dividiendo una muestra en moléculas de ácido nucleico hipermetiladas e hipometiladas. Al analizar múltiples fracciones de una muestra, puede realizarse un análisis multidimensional de un único locus de un genoma o especie de ácido

nucleico y, por lo tanto, puede lograrse una mayor sensibilidad.

En algunos casos, una muestra de ácido nucleico heterogénea se divide en dos o más divisiones (por ejemplo, por lo menos 3, 4, 5, 6 o 7 divisiones). En algunas realizaciones, cada división se etiqueta diferencialmente. Las divisiones etiquetadas pueden agruparse para la preparación y/o secuenciación colectiva de la muestra. Los pasos de división-etiquetado-agrupamiento pueden producirse más de una vez, con cada ronda de división produciéndose sobre la base de características diferentes (ejemplos proporcionados en la presente) y etiquetada usando etiquetas diferenciales que se distinguen de otras divisiones y medios de división.

Los ejemplos de características que pueden usarse para la división incluyen la longitud de la secuencia, el nivel de metilación, la unión a nucleosomas, el mal apareamiento de secuencias, la inmunoprecipitación y/o las proteínas que se unen al ADN. Las divisiones resultantes pueden incluir una o más de las siguientes formas de ácido nucleico: ADN de cadena sencilla (ADNmc), ADN de cadena doble (ADNdc), fragmentos de ADN más cortos y fragmentos de ADN más largos. Una población heterogénea de ácidos nucleicos se divide en dos o más fracciones sobre la base del nivel de metilación. Otros ejemplos de modificaciones epigenéticas incluyen la presencia o ausencia de metilación; el tipo de metilación (por ejemplo, 5-metilcitosina frente a otros tipos de metilación, como la metilación de adenina y/o la hidroximetilación de citosina); y la asociación y el nivel de asociación con una o más proteínas, como las histonas. Además, una población heterogénea de ácidos nucleicos puede dividirse en moléculas de ácido nucleico asociadas con nucleosomas y moléculas de ácido nucleico desprovistas de nucleosomas. Además, una población heterogénea de ácidos nucleicos puede dividirse en ADN de cadena sencilla (ADNmc) y ADN de cadena doble (ADNdc). Además, una población heterogénea de ácidos nucleicos puede dividirse sobre la base de la longitud del ácido nucleico (por ejemplo, moléculas de hasta 160 pb y moléculas con una longitud de más de 160 pb).

En algunos casos, cada división (representativa de una forma diferente de ácido nucleico) se marca diferencialmente, y las divisiones se agrupan entre sí antes de la secuenciación. En otros casos, las diferentes formas se secuencian por separado.

La FIG. 1 ilustra un método de división. Una población de diferentes ácidos nucleicos (101) se divide (102) en dos o más divisiones diferentes (103 a, b). Cada división (103 a, b) es representativa de una forma de ácido nucleico diferente. Cada división está etiquetada de manera distinta (104). Los ácidos nucleicos marcados se agrupan juntos (107) antes de la secuenciación (108). Las lecturas se analizan in silico. Las etiquetas se usan para clasificar las lecturas de las distintas divisiones. El análisis para detectar variantes genéticas puede realizarse a nivel de división por división, así como a nivel de toda la población de ácidos nucleicos. Por ejemplo, el análisis puede incluir análisis in silico para determinar variantes genéticas, como CNV, SNV, indel, fusión en ácidos nucleicos en cada división. En algunos casos, el análisis in silico puede incluir la determinación de la estructura de la cromatina. Por ejemplo, la cobertura de las lecturas de secuencia puede usarse para determinar la colocación de los nucleosomas en la cromatina. Una mayor cobertura puede correlacionarse con una mayor ocupación de nucleosomas en la región genómica, mientras que una menor cobertura puede correlacionarse con una menor ocupación de nucleosomas o una región con nucleosomas agotada (NDR).

Las muestras pueden incluir ácidos nucleicos que varían en modificaciones incluyendo modificaciones post-replicación a nucleótidos y unión, habitualmente no covalente, a una o más proteínas.

En una realización, la población de ácidos nucleicos es una obtenida de una muestra de suero, plasma o sangre de un sujeto que se sospecha que tiene neoplasia, un tumor o cáncer o diagnosticado previamente con neoplasia, un tumor o cáncer. La población de ácidos nucleicos incluye ácidos nucleicos con varios niveles de metilación. La metilación puede producirse a partir de una o más modificaciones post-replicación o transcripcionales. Las modificaciones post-replicación incluyen modificaciones del nucleótido citosina, particularmente en la posición 5 de la nucleobase, por ejemplo, 5-metilcitosina, 5-hidroximetilcitosina, 5-formilcitosina y 5-carboxilcitosina.

En algunas realizaciones, los ácidos nucleicos de la población original pueden ser de cadena sencilla y/o de cadena doble. La división basada en la cadena sencilla frente a la doble de los ácidos nucleicos puede realizarse, por ejemplo, usando sondas de captura marcadas para dividir el ADNmc y usando adaptadores de cadena doble para dividir el ADNdc.

Los agentes de afinidad pueden ser anticuerpos con la especificidad deseada, compañeros de unión naturales o variantes de los mismos (Bock et al., Nat Biotech 28: 1106-1114 (2010); Song et al., Nat Biotech 29: 68-72 (2011)), o péptidos artificiales seleccionados, por ejemplo, mediante visualización en fagos para que tengan especificidad para una diana determinada.

Ejemplos de fracciones de captura contempladas en la presente incluyen dominios de unión a metilo (MBD) y proteínas de unión a metilo (MBP) como se describe en la presente.

De igual manera, la división de diferentes formas de ácidos nucleicos puede realizarse usando proteínas de unión a histonas que pueden separar ácidos nucleicos unidos a histonas de ácidos nucleicos libres o no unidos.

Ejemplos de proteínas de unión a histonas que pueden usarse en los métodos divulgados en la presente incluyen RBBP4 (RbAp48) y péptidos de dominio SANT.

Aunque para algunos agentes de afinidad y modificaciones, la unión al agente puede producirse de manera esencialmente total o nula dependiendo de si un ácido nucleico lleva una modificación, la separación puede ser una de grado. En tales casos, los ácidos nucleicos sobrerrepresentados en una modificación se unen al agente en mayor medida que los ácidos nucleicos infrarrepresentados en la modificación. Alternativamente, los ácidos nucleicos que tienen modificaciones pueden unirse de manera total o parcial. Pero entonces, pueden eluirse secuencialmente del agente de unión varios niveles de modificaciones.

Por ejemplo, en algunas realizaciones, la división puede ser binaria o basarse en el grado/nivel de modificaciones. Por ejemplo, todos los fragmentos metilados pueden dividirse de los fragmentos no metilados usando proteínas de dominio de unión a metilo (por ejemplo, el Kit de enriquecimiento de ADN metilado MethylMiner (Thermo Fisher Scientific)). Posteriormente, la división adicional puede implicar la elución de fragmentos que tienen diferentes niveles de metilación ajustando la concentración de sal en una solución con el dominio de unión a metilo y los fragmentos unidos. A medida que aumenta la concentración de sal, se eluyen los fragmentos que tengan mayores niveles de metilación.

En algunos casos, las divisiones finales son representativas de ácidos nucleicos que tienen diferentes extensiones de modificaciones (sobrerrepresentación o infrarrepresentación de modificaciones). La sobrerrepresentación y la infrarrepresentación pueden definirse por el número de modificaciones que presenta un ácido nucleico con respecto a la mediana del número de modificaciones por cadena en una población. Por ejemplo, si la mediana del número de residuos de 5-metilcitosina en el ácido nucleico de una muestra es 2, un ácido nucleico que incluya más de dos residuos de 5-metilcitosina estará sobrerrepresentado en esta modificación y un ácido nucleico con 1 o cero residuos de 5-metilcitosina estará infrarrepresentado. El efecto de la separación por afinidad es enriquecer los ácidos nucleicos sobrerrepresentados en una modificación en una fase ligada y los ácidos nucleicos infrarrepresentados en una modificación en una fase no ligada (es decir, en solución). Los ácidos nucleicos de la fase ligada pueden eluirse antes del procesamiento posterior.

Cuando se usa el kit de enriquecimiento de ADN metilado MethylMiner (Thermo Fisher Scientific) pueden separarse varios niveles de metilación usando eluciones secuenciales. Por ejemplo, una división hipometilada (por ejemplo, sin metilación) puede separarse de una división metilada poniendo en contacto la población de ácido nucleico con el MBD del kit, que está unido a perlas magnéticas. Las perlas se usan para separar los ácidos nucleicos metilados de los no metilados. Posteriormente, se realizan uno o más pasos de elución secuencialmente para eluir los ácidos nucleicos que tienen diferentes niveles de metilación. Por ejemplo, un primer conjunto de ácidos nucleicos metilados puede eluirse a una concentración salina de 160 mM o mayor, por ejemplo, por lo menos 200 mM, 300 mM, 400 mM, 500 mM, 600 mM, 700 mM, 800 mM, 900 mM, 1000 mM o 2000 mM. Después de que hayan eluido dichos ácidos nucleicos metilados, se usa de nuevo separación magnética para separar los ácidos nucleicos con mayor nivel de metilación de los ácidos nucleicos con menor nivel de metilación. Los pasos de elución y separación magnética pueden repetirse para crear varias divisiones, como una división hipometilada (por ejemplo, representativa de la ausencia de metilación), una división metilada (representativa de un bajo nivel de metilación) y una división hipermetilada (representativa de un alto nivel de metilación).

En algunos métodos, los ácidos nucleicos unidos a un agente usado para la separación por afinidad se someten a un paso de lavado. El paso de lavado elimina los ácidos nucleicos débilmente unidos al agente de afinidad. Tales ácidos nucleicos pueden enriquecerse en ácidos nucleicos que tienen la modificación en una medida cercana a la media o mediana (es decir, intermedia entre los ácidos nucleicos que permanecen unidos a la fase sólida y los ácidos nucleicos que no se unen a la fase sólida en el contacto inicial de la muestra con el agente).

La separación por afinidad da como resultado por lo menos dos, y a veces tres o más divisiones de ácidos nucleicos con diferentes extensiones de una modificación. Mientras que las divisiones están todavía separadas, los ácidos nucleicos de por lo menos una división, y habitualmente dos o tres (o más) divisiones están enlazados a etiquetas de ácido nucleico, habitualmente proporcionadas como componentes de adaptadores, con los ácidos nucleicos en diferentes divisiones recibiendo diferentes etiquetas que distinguen a los miembros de una división de otra. Las etiquetas enlazadas a las moléculas de ácido nucleico de la misma división pueden ser iguales o diferentes entre sí. Pero si son diferentes entre sí, las etiquetas pueden tener parte de su código en común para identificar las moléculas a las que están unidas como pertenecientes a una división particular.

Para detalles adicionales referentes a la división de muestras de ácido nucleico basada en características como la metilación, véase el documento WO2018/119452.

Las moléculas de ácido nucleico pueden fraccionarse en diferentes divisiones sobre la base de las moléculas de ácido nucleico que están unidas a una proteína específica (un dominio de unión a metilo) o a un fragmento de la misma y las que no están unidas a esa proteína específica o fragmento de la misma.

Las moléculas de ácido nucleico pueden fraccionarse basándose en la unión ADN-proteína. Los complejos ADN-proteína pueden fraccionarse basándose en una propiedad específica de una proteína. Ejemplos de tales propiedades incluyen varios epítomos, modificaciones (por ejemplo, metilación o acetilación de histonas) o actividad enzimática. Ejemplos de proteínas que pueden unirse al ADN y servir de base para el fraccionamiento pueden incluir, entre otras, la proteína A y la proteína G. Puede usarse cualquier método adecuado para fraccionar las moléculas de ácido nucleico basándose en las regiones unidas a proteínas. Ejemplos de métodos usados para fraccionar moléculas de ácido nucleico sobre la base de regiones unidas a proteínas incluyen, pero no se limitan a, SDS-PAGE, cromatina-inmuno-precipitación (ChIP), cromatografía de heparina y fraccionamiento de flujo de campo asimétrico (AF4).

La división de los ácidos nucleicos se realiza poniendo en contacto los ácidos nucleicos con un dominio de unión a metilación ("MBD") de una proteína de unión a metilación ("MBP"). La MBD se une a la 5-metilcitosina (5mC). La MBD se acopla a perlas paramagnéticas, como Dynabeads® M-280 Streptavidin, mediante un conector de biotina. La separación en fracciones con diferentes grados de metilación puede realizarse eluyendo las fracciones mediante el aumento de la concentración de NaCl.

Ejemplos de MBP contemplados en la presente incluyen, pero no se limitan a:

(a) MeCP2 es una proteína que se une preferentemente a la 5-metil-citosina sobre la citosina no modificada.

(b) RPL26, PRP8 y la proteína reparadora de mal apareamientos del ADN MHS6 se unen preferentemente a la 5-hidroximetil-citosina sobre la citosina no modificada.

(c) FOXK1, FOXK2, FOXP1, FOXP4 y FOXI3 se unen preferentemente a la 5-formil citosina sobre la citosina no modificada (Iurlaro et al., Genome Biol. 14: R119 (2013)).

(d) Anticuerpos específicos de una o más bases nucleotídicas metiladas.

En general, la elución es una función del número de sitios metilados por molécula, con moléculas que tienen más metilación eluyendo bajo concentraciones de sal aumentadas. Para eluir el ADN en poblaciones distintas basadas en el grado de metilación, puede usarse una serie de tampones de elución de concentración creciente de NaCl. La concentración de sal puede variar de aproximadamente 100 mM a aproximadamente 2500 mM de NaCl. En una realización, el proceso da como resultado tres (3) divisiones. Las moléculas se ponen en contacto con una solución a una primera concentración de sal y que comprende una molécula con un dominio de unión a metilo, cuya molécula puede estar unida a una fracción de captura, tal como estreptavidina. A la primera concentración de sal, una población de moléculas se unirá al MBD y una población permanecerá no unida. La población no unida puede separarse como población "hipometilada". Por ejemplo, una primera división representativa de la forma hipometilada del ADN es la que permanece sin unirse a una concentración baja de sal, por ejemplo, 100 mM o 160 mM. Una segunda división representativa del ADN metilado intermedio se eluye usando una concentración intermedia de sal, por ejemplo, entre 100 mM y 2000 mM de concentración. También se separa de la muestra. Una tercera división representativa de la forma hipermetilada del ADN se eluye usando una concentración elevada de sal, por ejemplo, por lo menos de aproximadamente 2000 mM.

a. Etiquetado de divisiones

En algunas realizaciones, dos o más divisiones, por ejemplo, cada división, están marcadas diferencialmente. Las etiquetas pueden ser moléculas, como ácidos nucleicos, que contienen información que indica una característica de la molécula con la que está asociada la etiqueta. Por ejemplo, las moléculas pueden llevar una etiqueta de muestra (que distingue las moléculas de una muestra de las de otra muestra diferente), una etiqueta de división (que distingue las moléculas de una división de las de otra división diferente) o una etiqueta molecular (que distingue moléculas diferentes entre sí, tanto en escenarios de etiquetado único como no único). En ciertas realizaciones, una etiqueta puede comprender uno o una combinación de códigos de barras. Como se usa en la presente, el término "código de barras" se refiere a una molécula de ácido nucleico que tiene una secuencia particular de nucleótidos, o a la propia secuencia de nucleótidos, dependiendo del contexto. Un código de barras puede tener, por ejemplo, entre 10 y 100 nucleótidos. Una colección de códigos de barras puede tener secuencias degeneradas o puede tener secuencias que tengan una cierta distancia de Hamming, según se desee para el propósito específico. Así, por ejemplo, un índice de muestra, un índice de división o un índice molecular pueden estar compuestos por un código de barras o por una combinación de dos códigos de barras, cada uno unido a extremos diferentes de una molécula.

Las etiquetas pueden usarse para etiquetar las divisiones individuales de la población de polinucleótidos para correlacionar la etiqueta (o etiquetas) con una división específica. En algunas realizaciones, puede usarse una única etiqueta para etiquetar una división específica. En algunas realizaciones, pueden usarse múltiples etiquetas diferentes para etiquetar una división específica. En las realizaciones que emplean múltiples etiquetas diferentes para etiquetar una división específica, el conjunto de etiquetas usadas para etiquetar una división puede diferenciarse fácilmente del conjunto de etiquetas usadas para etiquetar otras divisiones. En algunas realizaciones, las etiquetas pueden tener funciones adicionales, por ejemplo las etiquetas pueden usarse para indexar fuentes de muestras o usarse como

identificadores moleculares únicos (que pueden usarse para mejorar la calidad de los datos de secuenciación diferenciando errores de secuenciación de mutaciones, por ejemplo como en Kinde et al., Proc Natl Acad Sci USA 108: 9530-9535 (2011), Kou et al., PLoS ONE, 11:e0146638 (2016)) o usarse como identificadores moleculares no únicos, por ejemplo como se describe en la Patente de Estados Unidos Nº 9.598.731. De manera similar, en algunas realizaciones, las etiquetas pueden tener funciones adicionales, por ejemplo, las etiquetas pueden usarse para indexar fuentes de muestras o usarse como identificadores moleculares no únicos (que pueden usarse para mejorar la calidad de los datos de secuenciación diferenciando los errores de secuenciación de las mutaciones).

En una realización, el etiquetado de divisiones comprende el etiquetado de moléculas en cada división con una etiqueta de división. Después de recombinar las divisiones y secuenciar las moléculas, las etiquetas de división identifican la división de origen. En otra realización, las diferentes divisiones se etiquetan con diferentes conjuntos de etiquetas moleculares, por ejemplo, compuestas por un par de códigos de barras. De esta manera, cada código de barras molecular indica la división de origen, además de ser útil para distinguir moléculas dentro de una división. Por ejemplo, puede usarse un primer conjunto de 35 códigos de barras para etiquetar moléculas en una primera división, mientras que puede usarse un segundo conjunto de 35 códigos de barras para etiquetar moléculas en una segunda división.

En algunas realizaciones, después de la división y etiquetado con etiquetas de división, las moléculas pueden agruparse para secuenciación en una única serie. En algunas realizaciones, a las moléculas se les añade una etiqueta de muestra, por ejemplo, en un paso posterior a la adición de etiquetas de división y agrupación. Las etiquetas de muestra pueden facilitar la agrupación de material generado a partir de múltiples muestras para su secuenciación en un único ciclo de secuenciación.

Alternativamente, en algunas realizaciones, las etiquetas de división pueden estar correlacionadas con la muestra así como con la división. Como ejemplo simple, una primera etiqueta puede indicar una primera división de una primera muestra; una segunda etiqueta puede indicar una segunda división de la primera muestra; una tercera etiqueta puede indicar una primera división de una segunda muestra; y una cuarta etiqueta puede indicar una segunda división de la segunda muestra.

Aunque las etiquetas pueden unirse a moléculas ya divididas basándose en una o más características, las moléculas etiquetadas finales de la biblioteca pueden no poseer ya esa característica. Por ejemplo, mientras que las moléculas de ADN de cadena sencilla pueden dividirse y etiquetarse, es probable que las moléculas etiquetadas finales de la biblioteca sean de cadena doble. De manera similar, mientras que el ADN puede someterse a división sobre la base de diferentes niveles de metilación, en la biblioteca final, las moléculas marcadas derivadas de estas moléculas probablemente no estén metiladas. Por consiguiente, la etiqueta unida a la molécula de la biblioteca indica típicamente la característica de la "molécula original" de la que se deriva la molécula marcada final, y no necesariamente la característica de la propia molécula marcada.

Por ejemplo, los códigos de barras 1, 2, 3, 4, etc. se usan para etiquetar y marcar moléculas en la primera división; los códigos de barras A, B, C, D, etc. se usan para etiquetar y marcar moléculas en la segunda división; y los códigos de barras a, b, c, d, etc. se usan para etiquetar y marcar moléculas en la tercera división. Las divisiones marcadas de manera diferencial pueden agruparse antes de la secuenciación. Las divisiones marcadas diferencialmente pueden secuenciarse por separado o secuenciarse juntas concurrentemente, por ejemplo, en la misma celda de flujo de un secuenciador Illumina.

Después de la secuenciación, el análisis de las lecturas para detectar variantes genéticas puede realizarse a nivel de división por división, así como a nivel de toda la población de ácido nucleico. Se usan etiquetas para clasificar las lecturas de diferentes divisiones. El análisis puede incluir análisis in silico para determinar la variación genética y epigenética (una o más de metilación, estructura de la cromatina, etc.) usando información de secuencia, longitud de coordenadas genómicas, cobertura y/o número de copias. En algunas realizaciones, una mayor cobertura puede correlacionarse con una mayor ocupación de nucleosomas en la región genómica, mientras que una menor cobertura puede correlacionarse con una menor ocupación de nucleosomas o una región agotada de nucleosomas (NDR).

b. Determinación del patrón de 5-metilcitosina de los ácidos nucleicos; secuenciación por bisulfito

La secuenciación basada en bisulfito y las variantes de la misma proporciona otro medio para determinar el patrón de metilación de un ácido nucleico que no depende de la división basada en el nivel de metilación antes de la secuenciación. En algunas realizaciones, la determinación del patrón de metilación comprende distinguir la 5-metilcitosina (5mC) de la citosina no metilada. En algunas realizaciones, la determinación del patrón de metilación comprende distinguir la N-metiladenina de la adenina no metilada. En algunas realizaciones, la determinación del patrón de metilación comprende distinguir la 5-hidroximetilcitosina (5hmC), la 5-formilcitosina (5fC) y la 5-carboxilcitosina (5caC) de la citosina no metilada. Ejemplos de secuenciación por bisulfito incluyen pero no se limitan a la secuenciación por bisulfito oxidativa (OX-BS-seq), secuenciación por bisulfito asistida por Tet (TAB-seq), y secuenciación por bisulfito reducida (redBS-seq). En algunas realizaciones, la determinación del patrón de metilación comprende la secuenciación por bisulfito de genoma completo, por ejemplo, como en MethylC-seq (Ulrich et al., Nature

Protocols 10:475-483 (2015)). En algunas realizaciones, la determinación del patrón de metilación comprende la determinación del patrón de metilación basado en matrices, por ejemplo, como en Methylation EPIC Beadchip o el uso de matrices Illumina Infinium (por ejemplo, matrices HumanMethylation450) (véase The Cancer Genome Atlas Research Network, Nature 507:315-322 (2014)). En algunas realizaciones, la determinación del patrón de metilación comprende PCR de bisulfito. En algunas realizaciones, la determinación del patrón de metilación comprende EM-Seq (US 2013/0244237 A1). En algunas realizaciones, la determinación del patrón de metilación comprende TAPS (WO 2019/136413 A1).

La secuenciación por bisulfito oxidativa (OX-BS-seq) se usa para distinguir entre 5mC y 5hmC, convirtiendo primero la 5hmC en 5fC, y procediendo después a la secuenciación por bisulfito. Para distinguir entre 5mC y 5hmC también puede usarse la secuenciación por bisulfito asistida por Tet (TAB-seq). En la TAB-seq, la 5hmC está protegida por glucosilación. A continuación, se usa una enzima Tet para convertir la 5mC en 5caC antes de proceder a la secuenciación por bisulfito. La secuenciación por bisulfito reducida se usa para distinguir 5fC de citosinas modificadas.

En general, en la secuenciación con bisulfito, una muestra de ácido nucleico se divide en dos alícuotas y una alícuota se trata con bisulfito. El bisulfito convierte la citosina nativa y ciertos nucleótidos de citosina modificados (por ejemplo, 5-formilcitosina o 5-carboxilcitosina) en uracilo, mientras que otras citosinas modificadas (por ejemplo, 5-metilcitosina, 5-hidroximetilcitosina) no se convierten. La comparación de las secuencias de ácido nucleico de las moléculas de las dos alícuotas indica qué citosinas se convirtieron en uracilos y cuáles no. En consecuencia, puede determinarse las citosinas que se modificaron y las que no. La división inicial de la muestra en dos alícuotas es desventajosa para las muestras que sólo contienen pequeñas cantidades de ácidos nucleicos, y/o compuestas de orígenes heterogéneos de células/tejidos, como los fluidos corporales que contienen ADN libre de células.

Por consiguiente, en algunas realizaciones, la secuenciación por bisulfito se lleva a cabo sin dividir inicialmente una muestra en dos alícuotas, por ejemplo, de la siguiente manera. En algunas realizaciones, los ácidos nucleicos de una población se enlazan a una fracción de captura como cualquiera de las descritas en la presente, es decir, un marcador que puede capturarse o inmovilizarse. Después del enlace de las fracciones de captura a los ácidos nucleicos de la muestra, los ácidos nucleicos de la muestra sirven como plantillas para la amplificación. Después de la amplificación, las plantillas originales permanecen enlazadas a las fracciones de captura, pero los amplicones no están enlazados a las fracciones de captura.

La fracción de captura puede enlazarse a los ácidos nucleicos de muestra como componente de un adaptador, que también puede proporcionar sitios de unión de cebadores de amplificación y/o secuenciación. En algunos métodos, los ácidos nucleicos de muestra se enlazan a adaptadores en ambos extremos, y ambos adaptadores llevan una fracción de captura. Preferiblemente, los residuos de citosina de los adaptadores se modifican, por ejemplo con 5metilcitosina, para protegerlos frente a la acción del bisulfito. En algunos casos, las fracciones de captura están enlazadas a las plantillas originales mediante un enlace escindible (por ejemplo, destiobitina-TEG fotoescindible o residuos de uracilo escindibles con la enzima USER™, Chem. Commun. (Camb). 51: 3266-3269 (2015)), en cuyo caso pueden eliminarse, si se desea, las fracciones de captura.

Los amplicones se desnaturalizan y se ponen en contacto con un reactivo de afinidad para la etiqueta de captura. Las plantillas originales se unen al reactivo de afinidad mientras que las moléculas de ácido nucleico resultantes de la amplificación no lo hacen. Por tanto, las plantillas originales pueden separarse de las moléculas de ácido nucleico resultantes de la amplificación.

Después de la separación de las plantillas originales de las moléculas de ácido nucleico resultantes de la amplificación, las plantillas originales pueden someterse a tratamiento con bisulfito. Alternativamente, los productos de amplificación pueden someterse a tratamiento con bisulfito y la población de plantillas originales no. Después de dicho tratamiento, las poblaciones respectivas pueden amplificarse (lo que en el caso de la población molde original convierte los uracilos en timinas). Las poblaciones también pueden someterse a hibridación con sonda de biotina para su captura. A continuación, se analizan las poblaciones respectivas y se comparan las secuencias para determinar qué citosinas estaban 5-metiladas (o 5-hidroximetiladas) en la muestra original. La detección de un nucleótido T en la población plantilla (correspondiente a una citosina no metilada convertida en uracilo) y de un nucleótido C en la posición correspondiente de la población amplificada indica una C no modificada. La presencia de C en las posiciones correspondientes de la plantilla original y de las poblaciones amplificadas indica una C modificada en la muestra original.

En algunas realizaciones, un método usa la preparación secuencial de bibliotecas NGS de ADN-seq y bisulfito-seq (BIS-seq) de bibliotecas de ADN molecular etiquetado (véase el documento WO 2018/119452, por ejemplo, en la FIG. 4). Este proceso se lleva a cabo mediante marcado de adaptadores (por ejemplo, biotina), amplificación de ADN-seq de toda la biblioteca, recuperación de molécula original (por ejemplo, extracción de perlas de estreptavidina), conversión de bisulfito y BIS-seq. En algunas realizaciones, el método identifica la 5-metilcitosina con resolución de bases individuales, a través de la amplificación secuencial NGS-preparativa de moléculas de biblioteca originales con y sin tratamiento de bisulfito. Esto puede lograrse modificando los adaptadores de NGS 5-metilados (adaptadores direccionales; con forma de Y/forcados con sustitución de 5-metilcitosina) usados en BIS-seq

con un marcador (por ejemplo, biotina) en una de las dos cadenas del adaptador. Las moléculas de ADN de muestra se ligan al adaptador y se amplifican (por ejemplo, mediante PCR). Como sólo las moléculas originales tendrán un extremo de adaptador marcado, pueden recuperarse selectivamente de su progenie amplificada mediante métodos de captura específicos (por ejemplo, perlas magnéticas de estreptavidina). Como las moléculas originales retienen las marcas de 5-metilación, la conversión por bisulfito en la biblioteca capturada producirá un estado de 5-metilación de resolución de una sola base en la BIS-seq, conservando información molecular para la ADN-seq correspondiente. En algunas realizaciones, la biblioteca tratada con bisulfito puede combinarse con una biblioteca no tratada antes de la captura/NGS mediante la adición de una secuencia de ADN de etiqueta de muestra en un flujo de trabajo de NGS multiplexado estándar. Al igual que con los flujos de trabajo BIS-seq, puede llevarse a cabo un análisis bioinformático para la alineación genómica y la identificación de bases 5-metiladas. En resumen, este método proporciona la capacidad de recuperar selectivamente las moléculas originales ligadas, portadoras de marcas de 5-metilcitosina, después de la amplificación de la biblioteca, permitiendo de este modo el procesamiento paralelo del ADN convertido en bisulfito. Esto supera la naturaleza destructiva del tratamiento por bisulfito sobre la calidad/sensibilidad de la información de ADN-seq extraída del flujo de trabajo. Con este método, las moléculas de ADN original ligadas y recuperadas (a través de adaptadores marcados) permiten la amplificación de la biblioteca de ADN completa y la aplicación paralela de tratamientos que provocan modificaciones epigenéticas del ADN. La presente divulgación analiza el uso de métodos BIS-seq para identificar la citosina-5-metilación (5-metilcitosina), pero en muchas realizaciones el uso de métodos BIS-seq no es necesario. Se han desarrollado variantes de BIS-seq para identificar citosinas hidroximetiladas (5hmC; OX- BS-seq, TAB-seq), formilcitosina (5fC; redBS-seq) y carboxilcitosinas. Estas metodologías pueden implementarse con la preparación secuencial/paralela de bibliotecas descrita en la presente.

c. Métodos alternativos de análisis de ácidos nucleicos modificados

En algunos de estos métodos, una población de ácidos nucleicos que llevan la modificación en diferentes grados (por ejemplo, 0, 1, 2, 3, 4, 5 o más grupos metilo por molécula de ácido nucleico) se pone en contacto con adaptadores antes del fraccionamiento de la población dependiendo del grado de la modificación. Los adaptadores se unen a uno o a ambos extremos de las moléculas de ácido nucleico de la población. Preferiblemente, los adaptadores incluyen diferentes etiquetas en número suficiente para que el número de combinaciones de etiquetas dé como resultado una probabilidad baja, por ejemplo, del 95, 99 o 99,9% de que dos ácidos nucleicos con los mismos puntos de inicio y parada reciban la misma combinación de etiquetas. Después de la unión de los adaptadores, los ácidos nucleicos se amplifican a partir de la unión de cebadores a los sitios de unión de cebadores dentro de los adaptadores. Los adaptadores, ya lleven las mismas etiquetas o unas diferentes, pueden incluir los mismos sitios de unión de cebadores o unos diferentes, pero preferiblemente los adaptadores incluyen el mismo sitio de unión de cebadores. Después de la amplificación, los ácidos nucleicos se ponen en contacto con un agente que preferiblemente se une a los ácidos nucleicos que llevan la modificación (como los agentes descritos anteriormente). Los ácidos nucleicos se separan en por lo menos dos divisiones que difieren en el grado en que los ácidos nucleicos llevan la modificación de la unión a los agentes. Por ejemplo, si el agente tiene afinidad por los ácidos nucleicos portadores de la modificación, los ácidos nucleicos sobrerrepresentados en la modificación (en comparación con la mediana de la representación en la población) se unen preferentemente al agente, mientras que los ácidos nucleicos infrarrepresentados para la modificación no se unen o se eluyen más fácilmente del agente. Después de la separación, las diferentes divisiones pueden someterse a pasos de procesamiento adicionales, típicamente incluyen la amplificación y el análisis de secuencias, en paralelo pero por separado. A continuación, pueden compararse los datos de secuencia de las distintas divisiones.

Tal esquema de separación puede realizarse usando el siguiente procedimiento ejemplar. Los ácidos nucleicos se enlazan en ambos extremos a adaptadores con forma de Y que incluyen sitios de unión de cebadores y etiquetas. Las moléculas se amplifican. A continuación, las moléculas amplificadas se fraccionan por contacto con un anticuerpo que se une preferentemente a la 5-metilcitosina para producir dos divisiones. Una división incluye moléculas originales que carecen de metilación y copias de amplificación que han perdido la metilación. La otra división incluye moléculas de ADN originales con metilación. A continuación, las dos divisiones se procesan y secuencian por separado con amplificación adicional de la división metilada. A continuación, pueden compararse los datos de secuencia de las dos divisiones. En este ejemplo, las etiquetas no se usan para distinguir entre ADN metilado y no metilado, sino más bien para distinguir entre diferentes moléculas dentro de estas divisiones, de tal manera que pueda determinarse si las lecturas con los mismos puntos de inicio y parada se basan en la misma molécula o en moléculas diferentes.

La divulgación proporciona métodos adicionales para analizar una población de ácidos nucleicos en la que por lo menos algunos de los ácidos nucleicos incluyen uno o más residuos de citosina modificados, como 5-metilcitosina y cualquiera de las otras modificaciones descritas anteriormente. En estos métodos, la población de ácidos nucleicos se pone en contacto con adaptadores que incluyen uno o más residuos de citosina modificados en la posición 5C, como la 5-metilcitosina. Preferiblemente, en tales adaptadores también se modifican todos los residuos de citosina, o se modifican todas las citosinas en una región de unión al cebador de los adaptadores. Los adaptadores se unen a ambos extremos de las moléculas de ácido nucleico de la población. Preferiblemente, los adaptadores incluyen diferentes etiquetas en número suficiente de tal manera que el número de combinaciones de etiquetas dé como resultado una baja probabilidad, por ejemplo, del 95, 99 o 99,9% de que dos ácidos nucleicos con los mismos puntos de inicio y parada reciban la misma combinación de etiquetas. Los sitios de unión del cebador en dichos

adaptadores pueden ser iguales o diferentes, pero preferiblemente son iguales. Después de la unión de los adaptadores, los ácidos nucleicos se amplifican a partir de cebadores que se unen a los sitios de unión de cebadores de los adaptadores. Los ácidos nucleicos amplificados se dividen en una primera y una segunda alícuotas. La primera alícuota se analiza para obtener datos de secuencia con o sin procesamiento adicional. De este modo, se determinan los datos de secuencia de las moléculas de la primera alícuota independientemente del estado de metilación inicial de las moléculas de ácido nucleico. Las moléculas de ácido nucleico de la segunda alícuota se tratan con bisulfito. Este tratamiento convierte las citosinas no modificadas en uracilos. A continuación, los ácidos nucleicos tratados con bisulfito se someten a una amplificación cebada por cebadores para los sitios de unión originales de los adaptadores enlazados al ácido nucleico. Sólo las moléculas de ácido nucleico originalmente enlazadas a los adaptadores (a diferencia de los productos de amplificación de los mismos) son ahora amplificables porque estos ácidos nucleicos conservan citosinas en los sitios de unión del cebador de los adaptadores, mientras que los productos de amplificación han perdido la metilación de estos residuos de citosina, que han experimentado conversión a uracilos en el tratamiento con bisulfito. Así, sólo se someten a amplificación las moléculas originales de las poblaciones, por lo menos algunas de las cuales están metiladas. Después la amplificación, estos ácidos nucleicos se someten a un análisis de secuencias. La comparación de las secuencias determinadas a partir de la primera y la segunda alícuotas puede indicar, entre otras cosas, qué citosinas de la población de ácidos nucleicos fueron objeto de metilación.

Dicho análisis puede realizarse usando el siguiente procedimiento ejemplar. El ADN metilado se enlaza a adaptadores con forma de Y en ambos extremos que incluyen sitios de unión de cebadores y etiquetas. Las citosinas en los adaptadores son 5-metiladas. La metilación de los cebadores sirve para proteger los sitios de unión de los cebadores en un paso posterior de bisulfito. Después de la unión de los adaptadores, se amplifican las moléculas de ADN. El producto de la amplificación se divide en dos alícuotas para la secuenciación con y sin tratamiento con bisulfito. La alícuota no sometida a secuenciación con bisulfito puede someterse a un análisis de secuencia con o sin tratamiento posterior. La otra alícuota se trata con bisulfito, que convierte las citosinas no metiladas en uracilos. Sólo los sitios de unión del cebador protegidos por la metilación de las citosinas pueden soportar la amplificación cuando se ponen en contacto con cebadores específicos para los sitios de unión del cebador original. Por tanto, sólo se someten a una amplificación posterior las moléculas originales y no las copias de la primera amplificación. A continuación, las moléculas amplificadas adicionalmente se someten a un análisis de secuencia. A continuación, pueden compararse las secuencias de las dos alícuotas. Al igual que en el esquema de separación analizado anteriormente, las etiquetas de ácido nucleico en los adaptadores no se usan para distinguir entre ADN metilado y no metilado, sino para distinguir moléculas de ácido nucleico dentro de la misma división.

d. PCR sensible a la metilación

En algunas realizaciones, se usa amplificación sensible a la metilación para evaluar la metilación en regiones diana variables de hipermetilación y/o variables de hipometilación. Varios pasos pueden hacerse sensibles a la metilación adaptando enfoques conocidos a los métodos descritos en la presente.

Por ejemplo, una muestra puede dividirse en alícuotas, por ejemplo, antes o después de un paso de captura como se describe en la presente, y una alícuota puede digerirse con una enzima de restricción sensible a la metilación, por ejemplo, como se describe en Moore et al., *Methods Mol Biol.* 325:239-49 (2006). Las secuencias no metiladas se digieren en esta alícuota. A continuación, las alícuotas digeridas y no digeridas pueden someterse a los pasos adecuados descritos en la presente (amplificación, opcionalmente etiquetado, secuenciación y similares) y las secuencias se analizan para determinar el grado de digestión en la muestra tratada, que refleja la presencia de citosinas no metiladas. Alternativamente, puede evitarse la división en alícuotas amplificando una muestra, separando el material amplificado de las plantillas originales y digiriendo a continuación el material original con una enzima de restricción sensible a la metilación antes de realizar una amplificación adicional, por ejemplo, como se ha analizado anteriormente con respecto a la secuenciación por bisulfito.

En otro ejemplo, una muestra puede dividirse en alícuotas y una alícuota puede tratarse para convertir las citosinas no metiladas en uracilo, por ejemplo, como se describe en el documento US 2003/0082600, antes de la captura. La conversión de citosinas no metiladas en uracilo reducirá la eficacia de la captura de regiones diana con baja metilación al alterar la secuencia de las regiones. Las alícuotas tratadas y no tratadas pueden luego someterse a los pasos apropiados descritos en la presente (captura, amplificación, opcionalmente etiquetado, secuenciación y similares) y analizarse las secuencias para determinar el grado de agotamiento de las regiones diana en la muestra tratada, que refleja la presencia de citosinas no metiladas.

4. Sujetos

En algunas realizaciones, el ADN (por ejemplo, ADNcf) se obtiene de un sujeto que tiene un cáncer. En algunas realizaciones, el ADN (por ejemplo, ADNcf) se obtiene de un sujeto que se sospecha que tiene cáncer. En algunas realizaciones, el ADN (por ejemplo, ADNcf) se obtiene de un sujeto que tiene un tumor. En algunas realizaciones, el ADN (por ejemplo, ADNcf) se obtiene de un sujeto que se sospecha que tiene un tumor. En algunas realizaciones, el ADN (por ejemplo, ADNcf) se obtiene de un sujeto con neoplasia. En algunas realizaciones, el ADN (por ejemplo, ADNcf) se obtiene de un sujeto que se sospecha que tiene una neoplasia. En algunas realizaciones, el

ADN (por ejemplo, ADNcf) se obtiene de un sujeto en remisión de un tumor, cáncer o neoplasia (por ejemplo, después de quimioterapia, resección quirúrgica, radiación o una combinación de las mismas). En cualquiera de las realizaciones anteriores, el cáncer, tumor o neoplasia, o la sospecha de cáncer, tumor o neoplasia, puede ser de pulmón, colon, recto, riñón, mama, próstata o hígado. En algunas realizaciones, el cáncer, tumor o neoplasia o sospecha de cáncer, tumor o neoplasia es de pulmón. En algunas realizaciones, el cáncer, tumor o neoplasia o sospecha de cáncer, tumor o neoplasia es de colon o recto. En algunas realizaciones, el cáncer, tumor o neoplasia o sospecha de cáncer, tumor o neoplasia es de mama. En algunas realizaciones, el cáncer, tumor o neoplasia o sospecha de cáncer, tumor o neoplasia es de próstata. En cualquiera de las realizaciones anteriores, el sujeto puede ser un humano.

En algunas realizaciones, el sujeto ha sido diagnosticado previamente con un cáncer, por ejemplo, cualquiera de los cánceres indicados anteriormente o en otra parte de la presente. Dicho sujeto puede haber recibido previamente uno o más tratamientos contra el cáncer, por ejemplo, cirugía, quimioterapia, radiación y/o inmunoterapia. En algunas realizaciones, se obtiene una muestra (por ejemplo, ADNcf) de un sujeto que ha sido previamente diagnosticado y tratado en uno o más puntos temporales preseleccionados después de uno o más tratamientos previos contra el cáncer.

La muestra (por ejemplo, ADNc) obtenida del sujeto puede secuenciarse para proporcionar un conjunto de información de secuencia, que puede incluir la secuenciación de moléculas de ADN capturadas del conjunto de regiones diana de secuencia variable con una mayor profundidad de secuenciación que las moléculas de ADN capturadas del conjunto de regiones diana epigenéticas, como se describe en detalle en otra parte de la presente.

5. Método ejemplar para la identificación de etiquetas moleculares de bibliotecas divididas con perlas MBD

Un método ejemplar para la identificación de etiquetas moleculares de bibliotecas divididas con perlas MBD mediante NGS es el siguiente:

i) División física de una muestra de ADN extraída (por ejemplo, ADN extraído de plasma sanguíneo de una muestra humana, que opcionalmente se ha sometido a captura de dianas como se describe en la presente) usando un kit de purificación de perlas de proteína de dominio de unión a metilo, guardando todas las eluciones del proceso para su procesamiento posterior.

ii) Aplicación paralela de etiquetas moleculares diferenciales y secuencias adaptadoras habilitantes para NGS a cada división. Por ejemplo, las divisiones hipermetilada, de metilación residual ('lavado') e hipometilada se ligan con adaptadores de NGS con etiquetas moleculares.

iii) Recombinación de todas las divisiones etiquetadas moleculares, y posterior amplificación usando secuencias de cebadores de ADN específicas del adaptador.

iv) Captura/hibridación de la biblioteca total recombinada y amplificada, dirigiéndose a regiones genómicas de interés (por ejemplo, variantes genéticas específicas del cáncer y regiones metiladas diferencialmente).

v) Reamplificación de la biblioteca de ADN capturada, añadiendo una etiqueta de muestra. Se agrupan diferentes muestras y se analizan en multiplexación en un instrumento NGS.

vi) Análisis bioinformático de los datos NGS, usando las etiquetas moleculares para identificar moléculas únicas, así como la deconvolución de la muestra en moléculas que se dividieron por MBD diferencialmente. Este análisis puede proporcionar información sobre la 5-metilcitosina relativa de las regiones genómicas, concurrente con la secuenciación genética estándar y la detección de variantes.

El método ejemplar expuesto anteriormente puede comprender además cualquier característica compatible de los métodos de acuerdo con esta divulgación expuestos en otra parte de la presente.

6. Flujos de trabajo ejemplares

En la presente se proporcionan flujos de trabajo ejemplares para la división y la preparación de bibliotecas. En algunas realizaciones, pueden usarse en combinación algunas o todas las características de los flujos de trabajo de división y preparación de bibliotecas. Los flujos de trabajo ejemplares expuestos anteriormente pueden comprender además cualquier característica compatible de los métodos de acuerdo con esta divulgación expuestos en otra parte de la presente.

a. División

En algunas realizaciones, el ADN de muestra (por ejemplo, entre 1 y 300 ng) se mezcla con una cantidad apropiada de tampón de dominio de unión a metilo (MBD) (la cantidad de tampón MBD depende de la cantidad de ADN usada) y perlas magnéticas conjugadas con proteínas de MBD y se incuban durante la noche. Durante esta

incubación el ADN metilado (ADN hipermetilado) se une a la proteína de MBD en las perlas magnéticas. El ADN no metilado (ADN hipometilado) o menos metilado (metilado intermedio) se lava de las perlas con tampones que contienen concentraciones crecientes de sal. Por ejemplo, a partir de tales lavados pueden obtenerse una, dos o más fracciones que contengan ADN no metilado, hipometilado y/o metilado de manera intermedia. Por último, para eluir el ADN fuertemente metilado (ADN hipermetilado) de la proteína MBD se usa un tampón con alto contenido en sal. En algunas realizaciones, estos lavados dan como resultado tres divisiones (división hipometilada, fracción metilada de manera intermedia y división hipermetilada) de ADN que tiene niveles crecientes de metilación.

En algunas realizaciones, las tres divisiones de ADN se desalan y concentran en preparación para los pasos enzimáticos de la preparación de bibliotecas.

b. Preparación de bibliotecas

En algunas realizaciones (por ejemplo, después de concentrar el ADN en las divisiones), el ADN dividido se hace ligable, por ejemplo, extendiendo los salientes de los extremos de las moléculas de ADN, y añadiendo residuos de adenosina a los extremos 3' de los fragmentos y fosforilando el extremo 5' de cada fragmento de ADN. Para ligar cada molécula de ADN dividida con un adaptador en cada extremo se añaden ADN ligasa y adaptadores. Estos adaptadores contienen etiquetas de división (por ejemplo, códigos de barras no aleatorios y no únicos) que son distinguibles de las etiquetas de división en los adaptadores usados en las otras divisiones. Después de la ligación, las tres divisiones se agrupan juntas y se amplifican (por ejemplo, mediante PCR, por ejemplo con cebadores específicos para los adaptadores).

Después de la PCR, el ADN amplificado puede limpiarse y concentrarse antes de la captura. El ADN amplificado se pone en contacto con una colección de sondas descritas en la presente (que pueden ser, por ejemplo, sondas de ARN biotiniladas) que se dirigen a regiones específicas de interés. La mezcla se incuba, por ejemplo, durante la noche, en un tampón salino. Las sondas se capturan (por ejemplo, usando perlas magnéticas de estreptavidina) y se separan del ADN amplificado que no se capturó, por ejemplo, mediante una serie de lavados con sal, proporcionando de este modo un conjunto capturado de ADN. Después de la captura, el ADN del conjunto capturado se amplifica mediante PCR. En algunas realizaciones, los cebadores de PCR contienen una etiqueta de muestra, incorporando de este modo la etiqueta de muestra en las moléculas de ADN. En algunas realizaciones, el ADN de diferentes muestras se agrupa junto y luego se secuencian en multiplex, por ejemplo, usando un secuenciador Illumina NovaSeq.

III. CARACTERÍSTICAS GENERALES DE LOS MÉTODOS

1. Muestras

Una muestra puede ser cualquier muestra biológica aislada de un sujeto. Una muestra puede ser una muestra corporal. Las muestras pueden incluir tejidos corporales, como tumores sólidos conocidos o sospechosos, sangre completa, plaquetas, suero, plasma, heces, glóbulos rojos, glóbulos blancos o leucocitos, células endoteliales, biopsias de tejidos, líquido cefalorraquídeo, líquido sinovial, líquido linfático, líquido de ascitis, líquido intersticial o extracelular, el líquido de los espacios entre células, incluyendo el líquido crevicular gingival, médula ósea, derrames pleurales, líquido cefalorraquídeo, saliva, mucosa, esputo, semen, orina. Las muestras son preferiblemente fluidos corporales, en particular sangre y fracciones de la misma, y orina. Una muestra puede estar en la forma originalmente aislada de un sujeto o puede haberse sometido a un procesamiento posterior para eliminar o añadir componentes, como células, o enriquecerla para un componente con respecto a otro. Por tanto, un fluido corporal preferido para el análisis es el plasma o el suero que contiene ácidos nucleicos libres de células. Una muestra puede aislarse u obtenerse de un sujeto y transportarse a un sitio de análisis de la muestra. La muestra puede conservarse y transportarse a una temperatura deseable, por ejemplo, temperatura ambiente, 4°C, -20°C, y/o -80°C. Una muestra puede aislarse u obtenerse de un sujeto en el lugar del análisis de la muestra. El sujeto puede ser un humano, un mamífero, un animal, un animal de compañía, un animal de servicio o una mascota. El sujeto puede tener un cáncer. El sujeto puede no tener cáncer o un síntoma detectable de cáncer. El sujeto puede haber sido tratado con una o más terapias contra el cáncer, por ejemplo, una o más quimioterapias, anticuerpos, vacunas o biología. El sujeto puede estar en remisión. Al sujeto se le puede haber diagnosticado o no susceptibilidad al cáncer o a mutaciones/trastornos genéticos asociados al cáncer.

El volumen de plasma puede depender de la profundidad de lectura deseada para las regiones secuenciadas. Los volúmenes ejemplares son 0,4-40 ml, 5-20 ml, 10-20 ml. Por ejemplo, el volumen puede ser de 0,5 ml, 1 ml, 5 ml, 10 ml, 20 ml, 30 ml o 40 ml. Un volumen de plasma muestreado puede ser de 5 a 20 ml.

Una muestra puede comprender varias cantidades de ácido nucleico que contiene equivalentes de genoma. Por ejemplo, una muestra de aproximadamente 30 ng de ADN puede contener aproximadamente 10.000 (10⁴) equivalentes de genoma humano haploide y, en el caso de ADNcf, aproximadamente 200.000 millones (2x10¹¹) de moléculas de polinucleótidos individuales. De manera similar, una muestra de aproximadamente 100 ng de ADN puede contener aproximadamente 30.000 equivalentes de genoma humano haploide y, en el caso del ADNcf,

aproximadamente 600.000 millones de moléculas individuales.

Una muestra puede comprender ácidos nucleicos de diferentes fuentes, por ejemplo, de células y libres de células del mismo sujeto, de células y libres de células de diferentes sujetos. Una muestra puede comprender ácidos nucleicos portadores de mutaciones. Por ejemplo, una muestra puede comprender ADN portador de mutaciones de la línea germinal y/o mutaciones somáticas. Las mutaciones de la línea germinal se refieren a mutaciones existentes en el ADN de la línea germinal de un sujeto. Las mutaciones somáticas se refieren a mutaciones que se originan en células somáticas de un sujeto, por ejemplo, células cancerosas. Una muestra puede comprender ADN portador de mutaciones asociadas al cáncer (por ejemplo, mutaciones somáticas asociadas al cáncer). Una muestra puede comprender una variante epigenética (es decir, una modificación química o proteica), en donde la variante epigenética se asocia con la presencia de una variante genética como una mutación asociada al cáncer. En algunas realizaciones, la muestra comprende una variante epigenética asociada con la presencia de una variante genética, en donde la muestra no comprende la variante genética.

Las cantidades ejemplares de ácidos nucleicos libres de células en una muestra antes de la amplificación varían de aproximadamente 1 fg a aproximadamente 1 µg, por ejemplo, de 1 pg a 200 ng, de 1 ng a 100 ng, de 10 ng a 1000 ng. Por ejemplo, la cantidad puede ser de hasta aproximadamente 600 ng, hasta aproximadamente 500 ng, hasta aproximadamente 400 ng, hasta aproximadamente 300 ng, hasta aproximadamente 200 ng, hasta aproximadamente 100 ng, hasta aproximadamente 50 ng, o hasta aproximadamente 20 ng de moléculas de ácido nucleico libres de células. La cantidad puede ser de por lo menos 1 fg, por lo menos 10 fg, por lo menos 100 fg, por lo menos 1 pg, por lo menos 10 pg, por lo menos 100 pg, por lo menos 1 ng, por lo menos 10 ng, por lo menos 100 ng, por lo menos 150 ng, o por lo menos 200 ng de moléculas de ácido nucleico libres de células. La cantidad puede ser de hasta 1 femtogramo (fg), 10 fg, 100 fg, 1 picogramo (pg), 10 pg, 100 pg, 1 ng, 10 ng, 100 ng, 150 ng, 200 ng, 250 ng o 300 ng de moléculas de ácido nucleico libres de células. El método puede comprender obtener de 1 femtogramo (fg) a 200 ng.

Los ácidos nucleicos libres de células son ácidos nucleicos no contenidos dentro de una célula o unidos de otro modo a una célula o, en otras palabras, ácidos nucleicos que permanecen en una muestra después de retirar las células intactas. Los ácidos nucleicos libres de células incluyen ADN, ARN e híbridos de los mismos, incluyendo ADN genómico, ADN mitocondrial, ARNip, ARNmi, ARN circulante (ARNc), ARNt, ARNr, ARN nucleolar pequeño (ARNsno), ARN que interactúa con Piwi (ARNpi), ARN no codificante largo (ARNnc largo) o fragmentos de cualquiera de ellos. Los ácidos nucleicos libres de células pueden ser de cadena sencilla, de cadena doble o híbridos de los mismos. Un ácido nucleico libre de células puede liberarse al fluido corporal a través de procesos de secreción o muerte celular, por ejemplo, necrosis celular y apoptosis. Algunos ácidos nucleicos libres de células se liberan en el fluido corporal a partir de células cancerosas, como el ADN tumoral circulante (ADNtc). Otros son liberados por células sanas. En algunas realizaciones, el ADNcf es ADN fetal libre de células (ADNcfl). En algunas realizaciones, los ácidos nucleicos libres de células son producidos por células tumorales. En algunas realizaciones, los ácidos nucleicos libres de células son producidos por una mezcla de células tumorales y no tumorales.

Los ácidos nucleicos libres de células tienen una distribución de tamaño ejemplar de aproximadamente 100-500 nucleótidos, con moléculas de 110 a aproximadamente 230 nucleótidos que representan aproximadamente el 90% de las moléculas, con una moda de aproximadamente 168 nucleótidos y un segundo pico menor en un intervalo entre 240 y 440 nucleótidos.

Los ácidos nucleicos libres de células pueden aislarse de fluidos corporales mediante un paso de fraccionamiento o división en el que los ácidos nucleicos libres de células, tal como se encuentran en solución, se separan de las células intactas y otros componentes no solubles del fluido corporal. La división puede incluir técnicas como la centrifugación o la filtración. Alternativamente, pueden lisarse las células de los fluidos corporales y procesarse conjuntamente los ácidos nucleicos celulares y libres de células. En general, después de la adición de tampones y pasos de lavado, los ácidos nucleicos pueden precipitarse con un alcohol. Para eliminar contaminantes o sales pueden usarse pasos adicionales de limpieza, como columnas a base de sílice. A lo largo de la reacción pueden añadirse ácidos nucleicos portadores a granel no específicos, como ADN C 1, ADN o proteínas para secuenciación con bisulfito, hibridación y/o ligación, para optimizar ciertos aspectos del procedimiento, como el rendimiento.

Después de dicho procesamiento, las muestras pueden incluir varias formas de ácido nucleico, incluyendo ADN de cadena doble, ADN de cadena sencilla y ARN de cadena sencilla. En algunas realizaciones, el ADN y el ARN de cadena sencilla pueden convertirse en formas de cadena doble para que se incluyan en los pasos posteriores de procesamiento y análisis.

Las moléculas de ADN de cadena doble de una muestra y las moléculas de ácido nucleico de cadena sencilla convertidas en moléculas de ADN de cadena doble pueden enlazarse a adaptadores en cualquier extremo o en ambos extremos. Típicamente, las moléculas de cadena doble se terminan de manera roma mediante tratamiento con una polimerasa con una polimerasa 5'-3' y una exonucleasa 3'-5' (o función de lectura de pruebas), en presencia de los cuatro nucleótidos estándar. El fragmento grande de Klenow y la polimerasa T4 son ejemplos de polimerasas adecuadas. Las moléculas de ADN de extremo roma pueden ligarse con un adaptador por lo menos parcialmente de

cadena doble (por ejemplo, un adaptador con forma de Y o de campana). Alternativamente, a los extremos romos de los ácidos nucleicos de muestra y a los adaptadores se les pueden añadir nucleótidos complementarios para facilitar la ligación. En la presente se contemplan tanto la ligadura de extremo romo como la ligadura de extremo pegajoso. En la ligadura de extremos romos, tanto las moléculas de ácido nucleico como las etiquetas de adaptadores tienen extremos romos. En la ligadura de extremo pegajoso, típicamente, las moléculas de ácido nucleico llevan un saliente "A" y los adaptadores llevan un saliente "T".

2. Etiquetas

Pueden incorporarse o unirse de otro modo a adaptadores etiquetas que comprenden códigos de barras. Las etiquetas pueden incorporarse por ligadura, PCR de extensión solapada, entre otros métodos.

a. Estrategias de marcado molecular

El etiquetado molecular se refiere a una práctica de etiquetado que permite diferenciar las moléculas a partir de las cuales se originaron las lecturas de secuencias. Las estrategias de etiquetado pueden dividirse en estrategias de etiquetado único y estrategias de etiquetado no único. En el etiquetado único, todas o casi todas las moléculas de una muestra llevan una etiqueta diferente, de tal manera que las lecturas pueden asignarse a las moléculas originales basándose únicamente en la información de la etiqueta. A veces, las etiquetas usadas en estos métodos se denominan "etiquetas únicas". En el etiquetado no único, diferentes moléculas de la misma muestra pueden llevar la misma etiqueta, de tal manera que se usa otra información además de la información de la etiqueta para asignar una lectura de secuencia a una molécula original. Dicha información puede incluir la coordenada de inicio y parada, la coordenada para la que mapea la molécula, la coordenada de inicio o parada sola, etc. Las etiquetas usadas en tales métodos se denominan a veces "etiquetas no únicas". Por consiguiente, no es necesario etiquetar de manera única cada molécula de una muestra. Basta con etiquetar de manera única las moléculas que pertenecen a una clase identificable dentro de una muestra. Por tanto, moléculas de diferentes familias identificables pueden llevar la misma etiqueta sin pérdida de información sobre la identidad de la molécula etiquetada.

En ciertas realizaciones de etiquetado no único, el número de etiquetas diferentes usadas puede ser suficiente para que haya una probabilidad muy alta (por ejemplo, por lo menos del 99%, por lo menos del 99,9%, por lo menos del 99,99% o por lo menos del 99,999%) de que todas las moléculas de un grupo particular lleven una etiqueta diferente. Cabe señalar que cuando se usan códigos de barras como etiquetas, y cuando los códigos de barras se unen, por ejemplo, aleatoriamente, a ambos extremos de una molécula, la combinación de códigos de barras, en conjunto, puede constituir una etiqueta. Este número, en término, es una función del número de moléculas que entran en las llamadas. Por ejemplo, la clase puede ser todas las moléculas que corresponden a la misma posición de inicio-parada en un genoma de referencia. La clase puede ser todas las moléculas que mapean a través de un locus genético particular, por ejemplo, una base particular o una región particular (por ejemplo, hasta 100 bases o un gen o un exón de un gen). En ciertas realizaciones, el número de etiquetas diferentes usadas para identificar de manera única un número de moléculas, z , en una clase puede estar entre cualquiera de 2^*z , 3^*z , 4^*z , 5^*z , 6^*z , 7^*z , 8^*z , 9^*z , 10^*z , 11^*z , 12^*z , 13^*z , 14^*z , 15^*z , 16^*z , 17^*z , 18^*z , 19^*z , 20^*z o 100^*z (por ejemplo, límite inferior) y cualquiera de 100.000^*z , 10.000^*z , 1.000^*z o 100^*z (por ejemplo, límite superior).

Por ejemplo, en una muestra de aproximadamente 3 ng a 30 ng de ADN libre de células humanas, se espera que se mapeen alrededor de 10^3 - 10^4 moléculas a una coordenada de nucleótido particular, y entre aproximadamente 3 y 10 moléculas que tengan cualquier coordenada de inicio compartan la misma coordenada de parada. Por consiguiente, entre aproximadamente 50 y aproximadamente 50.000 etiquetas diferentes (por ejemplo, entre aproximadamente 6 y aproximadamente 220 combinaciones de códigos de barras) pueden ser suficientes para etiquetar de manera única todas estas moléculas. Para etiquetar de manera única las 10^3 - 10^4 moléculas que mapean a través de una coordenada nucleotídica, se necesitarían de aproximadamente 1 millón a aproximadamente 20 millones de etiquetas diferentes.

Generalmente, la asignación de códigos de barras de etiquetas únicas o no únicas en reacciones sigue los métodos y sistemas descritos en las Solicitudes de Patente de Estados Unidos 20010053519, 20030152490, 20110160078, y la Patente de Estados Unidos N° 6.582.908 y la Patente de Estados Unidos N° 7.537.898 y la Patente de Estados Unidos N° 9.598.731. Las etiquetas pueden enlazarse a los ácidos nucleicos de la muestra aleatoria o no aleatoriamente.

En algunas realizaciones, los ácidos nucleicos marcados se secuencian después de cargarlos en una placa de micropocillos. La placa de micropocillos puede tener 96, 384 o 1536 micropocillos. En algunos casos, se introducen en una proporción esperada de etiquetas únicas por micropocillos. Por ejemplo, las etiquetas únicas pueden cargarse de modo que se carguen más de aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000, 5000, 10000, 50.000, 100.000, 500.000, 1.000.000, 10.000.000, 50.000.000 o 1.000.000.000 de etiquetas únicas por muestra de genoma. En algunos casos, las etiquetas únicas pueden cargarse de tal manera que se carguen menos de aproximadamente 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000, 5000, 10000, 50.000, 100.000, 500.000, 1.000.000, 10.000.000, 50.000.000 o 1.000.000.000 de etiquetas únicas por muestra de genoma. En algunos casos, el número

medio de etiquetas únicas cargadas por muestra genómica es menor de, o mayor de, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 500, 1000, 5000, 10000, 50.000, 100.000, 500.000, 1.000.000, 10.000.000, 50.000.000 o 1.000.000.000 de etiquetas únicas por muestra genómica.

Un formato preferido usa 20-50 etiquetas diferentes (por ejemplo, códigos de barras) ligadas a ambos extremos de los ácidos nucleicos diana. Por ejemplo, 35 etiquetas diferentes (por ejemplo, códigos de barras) ligadas a ambos extremos de moléculas diana creando 35 x 35 permutaciones, lo que equivale a 1225 combinaciones de etiquetas para 35 etiquetas. Este número de etiquetas es suficiente para que diferentes moléculas que tengan los mismos puntos de inicio y parada tengan una alta probabilidad (por ejemplo, por lo menos un 94%, 99,5%, 99,99%, 99,999%) de recibir diferentes combinaciones de etiquetas. Otras combinaciones de códigos de barras incluyen cualquier número entre 10 y 500, por ejemplo, aproximadamente 15x15, aproximadamente 35x35, aproximadamente 75x75, aproximadamente 100x100, aproximadamente 250x250, aproximadamente 500x500.

En algunos casos, las etiquetas únicas pueden ser oligonucleótidos de secuencia predeterminada o aleatoria o semialeatoria. En otros casos, puede usarse una pluralidad de códigos de barras de tal manera que los códigos de barras no sean necesariamente únicos entre sí en la pluralidad. En este ejemplo, los códigos de barras pueden ligarse a moléculas individuales de tal manera que la combinación del código de barras y la secuencia a la que puede ligarse cree una secuencia única que pueda rastrearse individualmente. Como se describe en la presente, la detección de códigos de barras no únicos en combinación con los datos de secuencia de las partes de comienzo (inicio) y final (parada) de las lecturas de secuencia puede permitir la asignación de una identidad única a una molécula particular. También puede usarse la longitud o el número de pares de bases de una lectura de secuencia individual para asignar una identidad única a dicha molécula. Como se describe en la presente, los fragmentos de una única cadena de ácido nucleico a los que se ha asignado una identidad única pueden permitir por lo tanto la identificación posterior de fragmentos de la cadena original.

3. Amplificación

Los ácidos nucleicos de muestra flanqueados por adaptadores pueden amplificarse mediante PCR y otros métodos de amplificación. La amplificación se realiza típicamente mediante cebadores que se unen a sitios de unión de cebadores en adaptadores que flanquean una molécula de ADN a amplificar. Los métodos de amplificación pueden implicar ciclos de desnaturalización, apareamiento y extensión, resultantes del termociclado, o pueden ser isotérmicos, como en la amplificación mediada por transcripción. Otros métodos de amplificación incluyen la reacción en cadena de la ligasa, la amplificación por desplazamiento de cadena, la amplificación basada en secuencias de ácidos nucleicos y la replicación auto sostenida basada en secuencias.

Preferiblemente, los presentes métodos realizan "ligaciones TV A" de ADNdc con adaptadores de cola T y cola C, que dan como resultado la amplificación de por lo menos el 50, 60, 70 u 80% de los ácidos nucleicos de cadena doble antes de ligarse a los adaptadores. Preferiblemente, los presentes métodos aumentan la cantidad o el número de moléculas amplificadas con respecto a los métodos de control realizados con adaptadores de cola T solos en por lo menos un 10, 15 o 20%.

4. Conjuntos de cebos; Fracciones de captura; Enriquecimiento

Como se ha analizado anteriormente, los ácidos nucleicos de una muestra pueden someterse a un paso de captura, en el que se capturan las moléculas que tienen secuencias diana para su análisis posterior. La captura de la diana puede implicar el uso de un conjunto de cebos que comprenda cebos de oligonucleótidos marcados con una fracción de captura, como biotina u otros ejemplos indicados a continuación. Las sondas pueden tener secuencias seleccionadas para formar segmento a través de un panel de regiones, como genes. En algunas realizaciones, un conjunto de cebos puede tener rendimientos de captura más altos y más bajos para conjuntos de regiones diana que los del conjunto de regiones diana de secuencia variable y el conjunto de regiones diana epigenéticas, respectivamente, como se analiza en otra parte de la presente. Tales conjuntos de cebos se combinan con una muestra en condiciones que permiten la hibridación de las moléculas diana con los cebos. A continuación, las moléculas capturadas se aíslan usando la fracción de captura. Por ejemplo, una fracción de captura de biotina mediante estreptavidina basada en perlas. Tales métodos se describen con más detalle en, por ejemplo, la Patente de Estados Unidos 9.850.523, concedida el 26 de diciembre de 2017.

Las fracciones de captura incluyen, sin limitación, biotina, avidina, estreptavidina, un ácido nucleico que comprende una secuencia de nucleótidos particular, un hapteno reconocido por un anticuerpo y partículas atraíbles magnéticamente. La fracción de extracción puede ser un miembro de un par de unión, como biotina/estreptavidina o hapteno/anticuerpo. En algunas realizaciones, una fracción de captura que está unida a un analito es capturada por su par de unión que está unido a una fracción aislable, como una partícula atraíble magnéticamente o una partícula grande que puede sedimentarse mediante centrifugación. La fracción de captura puede ser cualquier tipo de molécula que permita la separación por afinidad de los ácidos nucleicos que llevan la fracción de captura de los ácidos nucleicos que carecen de la fracción de captura. Las fracciones de captura ejemplares son la biotina, que permite la separación por afinidad mediante la unión a estreptavidina enlazada o enlazable a una fase sólida, o un oligonucleótido, que

permite la separación por afinidad mediante la unión a un oligonucleótido complementario enlazado o enlazable a una fase sólida.

5. Secuenciación

Los ácidos nucleicos de muestra, opcionalmente flanqueados por adaptadores, con o sin amplificación previa se someten generalmente a secuenciación. Los métodos de secuenciación o formatos disponibles comercialmente que se utilizan opcionalmente incluyen, por ejemplo, secuenciación de Sanger, secuenciación de alto rendimiento, pirosecuenciación, secuenciación por síntesis, secuenciación de molécula única, secuenciación basada en nanoporos, secuenciación por semiconductores, secuenciación por ligación, secuenciación por hibridación, RNA-Seq (Illumina), Digital Gene Expression (Helicos), secuenciación de próxima generación (NGS), secuenciación de molécula única por síntesis (SMSS) (Helicos), secuenciación masivamente paralela, Clonal Single Molecule Array (Solexa), secuenciación por escopeta, Ion Torrent, Oxford Nanopore, Roche Genia, secuenciación Maxim-Gilbert, caminata de cebadores, secuenciación mediante plataformas PacBio, SOLiD, Ion Torrent o Nanopore. Las reacciones de secuenciación pueden realizarse en una variedad de unidades de procesamiento de muestras, que pueden incluir múltiples carriles, múltiples canales, múltiples pocillos u otros medios para procesar múltiples conjuntos de muestras de manera sustancialmente simultánea. Las unidades de procesamiento de muestras también pueden incluir múltiples cámaras de muestras para permitir el procesamiento de múltiples series simultáneamente.

Las reacciones de secuenciación pueden realizarse en uno o más tipos de fragmentos de ácido nucleico o regiones que contengan marcadores de cáncer o de otras enfermedades. Las reacciones de secuenciación también pueden realizarse en cualquier fragmento de ácido nucleico presente en la muestra. Las reacciones de secuenciación pueden realizarse en por lo menos aproximadamente un 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 99%, 99,9% o 100% del genoma. En otros casos, las reacciones de secuenciación pueden realizarse en menos de aproximadamente el 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 99%, 99,9% o 100% del genoma.

Pueden realizarse reacciones de secuenciación simultáneas usando técnicas de secuenciación multiplex. En algunas realizaciones, los polinucleótidos libres de células se secuencian con por lo menos aproximadamente 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, o 100.000 reacciones de secuenciación. En otras realizaciones, los polinucleótidos libres de células se secuencian con menos de aproximadamente 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000 o 100.000 reacciones de secuenciación. Las reacciones de secuenciación se realizan típicamente secuencial o simultáneamente. El análisis de datos posterior se realiza generalmente en todas o en parte de las reacciones de secuenciación. En algunas realizaciones, el análisis de datos se realiza en por lo menos aproximadamente 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000 o 100.000 reacciones de secuenciación. En otras realizaciones, el análisis de datos puede realizarse en menos de aproximadamente 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000 o 100.000 reacciones de secuenciación. Un ejemplo de profundidad de lectura es de aproximadamente 1000 a aproximadamente 50000 lecturas por locus (por ejemplo, posición de base).

a. Profundidad diferencial de secuenciación

En algunas realizaciones, los ácidos nucleicos correspondientes al conjunto de regiones diana de secuencia variable se secuencian a una mayor profundidad de secuenciación que los ácidos nucleicos correspondientes al conjunto de regiones diana epigenéticas. Por ejemplo, la profundidad de secuenciación para ácidos nucleicos correspondientes al conjunto de regiones diana de variantes de secuencia puede ser de por lo menos 1,25-, 1,5-, 1,75-, 2-, 2,25-, 2,5-, 2,75-, 3-, 3,5-, 4-, 4,5-, 5-, 6-, 7-, 8-, 9-, 10-, 11-, 12-, 13-, 14-, o 15 veces mayor, o de 1,25- a 1,5-, de 1,5- a 1,75-, de 1,75- a 2-, de 2- a 2,25-, de 2,25- a 2,5-, de 2,5- a 2,75-, de 2,75- a 3-, de 3- a 3,5-, de 3,5- a 4-, de 4- a 4,5-, de 4,5- a 5-, de 5- a 5,5-, de 5,5- a 6-, de 6- a 7-, de 7- a 8-, de 8- a 9-, de 9- a 10-, de 10- a 11-, de 11- a 12-, de 12- a 13-, de 13- a 14-, de 14- a 15- veces o de 15- a 100- veces mayor que la profundidad de secuenciación de los ácidos nucleicos correspondientes al conjunto de regiones diana epigenéticas. En algunas realizaciones, dicha profundidad de secuenciación es por lo menos 2 veces mayor. En algunas realizaciones, dicha profundidad de secuenciación es por lo menos 5 veces mayor. En algunas realizaciones, dicha profundidad de secuenciación es por lo menos 10 veces mayor. En algunas realizaciones, dicha profundidad de secuenciación es de 4 a 10 veces mayor. En algunas realizaciones, dicha profundidad de secuenciación es de 4 a 100 veces mayor. Cada una de estas realizaciones se refiere a la medida en que los ácidos nucleicos correspondientes al conjunto de regiones diana de secuencia variable se secuencian a una mayor profundidad de secuenciación que los ácidos nucleicos correspondientes al conjunto de regiones diana epigenéticas.

En algunas realizaciones, el ADNcf capturado correspondiente al conjunto de regiones diana de secuencia variable y el ADNcf capturado correspondiente al conjunto de regiones diana epigenéticas se secuencian concurrentemente, por ejemplo, en la misma celda de secuenciación (como la celda de flujo de un secuenciador Illumina) y/o en la misma composición, que puede ser una composición agrupada resultante de recombinar conjuntos capturados por separado o una composición obtenida capturando el ADNcf correspondiente al conjunto de regiones diana de secuencia variable y el ADNcf capturado correspondiente al conjunto de regiones diana epigenéticas en el mismo recipiente.

6. Análisis

La secuenciación puede generar una pluralidad de lecturas de secuencia o lecturas. Las lecturas de secuencia o lecturas pueden incluir secuencias de datos de nucleótidos de menos de aproximadamente 150 bases de longitud, o menos de aproximadamente 90 bases de longitud. En algunas realizaciones, las lecturas tienen entre aproximadamente 80 bases y aproximadamente 90 bases, por ejemplo, aproximadamente 85 bases de longitud. En algunas realizaciones, los métodos de la presente divulgación se aplican a lecturas muy cortas, por ejemplo, menos de aproximadamente 50 bases o aproximadamente 30 bases de longitud. Los datos de lectura de secuencias pueden incluir los datos de secuencia así como la metainformación. Los datos de lectura de secuencias pueden almacenarse en cualquier formato de archivo adecuado incluyendo, por ejemplo, archivos VCF, archivos FASTA o archivos FASTQ.

FASTA puede referirse a un programa informático para buscar bases de datos de secuencias, y el nombre FASTA también puede referirse a un formato de archivo estándar. FASTA se describe, por ejemplo, en Pearson & Lipman, 1988, Improved tools for biological sequence comparison, PNAS 85:2444- 2448. Una secuencia en formato FASTA comienza con una descripción de única línea, seguida de líneas de datos de secuencia. La línea de descripción se distingue de los datos de la secuencia por un símbolo mayor que (">") en la primera columna. La palabra que sigue al símbolo ">" es el identificador de la secuencia, y el resto de la línea es la descripción (ambas son opcionales). Puede no haber ningún espacio entre el ">" y la primera letra del identificador. Se recomienda que todas las líneas de texto tengan menos de 80 caracteres. La secuencia termina si aparece otra línea que empiece por ">"; esto indica el inicio de otra secuencia.

El formato FASTQ es un formato basado en texto para almacenar tanto una secuencia biológica (habitualmente la secuencia de nucleótidos) como sus correspondientes puntuaciones de calidad. Es similar al formato FASTA pero con las puntuaciones de calidad a continuación de los datos de la secuencia. Tanto la letra de la secuencia como la puntuación de calidad se codifican con un único carácter ASCII por brevedad. El formato FASTQ es un estándar de facto para almacenar la salida de instrumentos de secuenciación de alto rendimiento como el Illumina Genome Analyzer, como se describe, por ejemplo, por Cock et al. ("The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants," Nucleic Acids Res 38(6): 1767-1771, 2009).

Para archivos FASTA y FASTQ, la metainformación incluye la línea de descripción y no las líneas de datos de secuencia. En algunas realizaciones, para los archivos FASTQ, la metainformación incluye las puntuaciones de calidad. Para los archivos FASTA y FASTQ, los datos de secuencia comienzan después de la línea de descripción y se presentan típicamente usando algún subconjunto de códigos de ambigüedad IUPAC, opcionalmente con "-". En una realización, los datos de secuencia pueden usar los caracteres A, T, C, G y N, incluyendo opcionalmente "-" o U según sea necesario (por ejemplo, para representar huecos o uracilo).

En algunas realizaciones, el por lo menos un archivo de lectura de secuencia maestra y el archivo de salida se almacenan como archivos de texto sin formato (por ejemplo, usando una codificación como ASCII; ISO/IEC 646; EBCDIC; UTF-8; o UTF-16). Un sistema informático proporcionado por la presente divulgación puede incluir un programa editor de texto capaz de abrir los archivos de texto sin formato. Un programa editor de texto puede referirse a un programa informático capaz de presentar el contenido de un archivo de texto (como un archivo de texto sin formato) en una pantalla de ordenador, permitiendo a un humano editar el texto (por ejemplo, usando un monitor, teclado y ratón). Ejemplos de editores de texto incluyen, sin limitación, Microsoft Word, emacs, pico, vi, BBEdit, y TextWrangler. El programa editor de texto puede ser capaz de mostrar los archivos de texto sin formato en una pantalla de ordenador, mostrando la metainformación y la secuencia leída en un formato legible por humanos (por ejemplo, no codificado en binario, sino usando caracteres alfanuméricos como pueden ser usados en la impresión o escritura humanas).

Aunque los métodos se han analizado con referencia a archivos FASTA o FASTQ, los métodos y sistemas de la presente divulgación pueden usarse para comprimir cualquier formato de archivo de secuencias adecuado incluyendo, por ejemplo, archivos en el formato Variant Call Format (VCF). Un archivo VCF típico puede incluir una sección de encabezamiento y una sección de datos. El encabezamiento contiene un número arbitrario de líneas de metainformación, cada una de las cuales comienza con el carácter "##", y una línea de definición de campos delimitada por TAB que comienza con un único carácter "#". La línea de definición de campos nombra ocho columnas obligatorias y la sección de cuerpo contiene líneas de datos que rellenan las columnas definidas por la línea de definición de campos. El formato VCF se describe, por ejemplo, en Danecek et al. ("The variant call format and VCF tools," Bioinformatics 27(15):2156-2158, 2011). La sección de encabezado puede tratarse como la metainformación que se escribirá en los archivos comprimidos y la sección de datos puede tratarse como las líneas, cada una de las cuales puede almacenarse en un archivo maestro solo si es única.

Algunas realizaciones permiten el ensamblaje de lecturas de secuencias. En el ensamblaje por alineación, por ejemplo, las lecturas de secuencia se alinean entre sí o se alinean con una secuencia de referencia. Al alinear cada lectura, a su vez con un genoma de referencia, todas las lecturas se posicionan en relación con las demás para crear el ensamblaje. Además, la alineación o el mapeo de la lectura de secuencia con una secuencia de referencia

también puede usarse para identificar secuencias variantes dentro de la lectura de secuencia. La identificación de secuencias variantes puede usarse en combinación con los métodos y sistemas descritos en la presente para ayudar en el diagnóstico o pronóstico de una enfermedad o afección, o para guiar las decisiones de tratamiento.

En algunas realizaciones, cualquiera o todos los pasos están automatizados. Alternativamente, los métodos de la presente divulgación pueden incorporarse total o parcialmente en uno o más programas dedicados, por ejemplo, cada uno opcionalmente escrito en un lenguaje compilado como C++, luego compilado y distribuido como un binario. Los métodos de la presente divulgación pueden implementarse total o parcialmente como módulos dentro de, o invocando funcionalidad dentro de, plataformas de análisis de secuencias existentes. En algunas realizaciones, los métodos de la presente divulgación incluyen una serie de pasos que se aplican todos automáticamente en respuesta a una única cola de inicio (por ejemplo, uno o una combinación de eventos desencadenantes procedentes de la actividad humana, otro programa informático o una máquina). Por tanto, la presente divulgación proporciona métodos en los que pueden producirse automáticamente cualquiera de los pasos o cualquier combinación de los pasos en respuesta a una cola. "Automáticamente" generalmente significa sin intervención, influencia o interacción humana (por ejemplo, en respuesta sólo a la actividad humana original o previa a la cola).

Los métodos de la presente divulgación también pueden abarcar varias formas de salida, que incluye una interpretación precisa y sensible de la muestra de ácido nucleico de un sujeto. La salida de la recuperación puede proporcionarse en el formato de un archivo informático. En algunas realizaciones, la salida es un archivo FASTA, un archivo FASTQ o un archivo VCF. La salida puede procesarse para producir un archivo de texto o un archivo XML que contenga datos de secuencia, como una secuencia del ácido nucleico alineada con una secuencia del genoma de referencia. En otras realizaciones, el procesamiento proporciona un resultado que contiene coordenadas o una cadena que describe una o más mutaciones en el ácido nucleico en cuestión con respecto al genoma de referencia. Las cadenas de alineación pueden incluir Simple UnGapped Alignment Report (SUGAR), Verbose Useful Labeled Gapped Alignment Report (VULGAR), y Compact Idiosyncratic Gapped Alignment Report (CIGAR) (como se describe, por ejemplo, en Ning et al., *Genome Research* 11 (10): 1725-9, 2001). Estas cadenas pueden implementarse, por ejemplo, en el software de alineación de secuencias Exonerate del Instituto Europeo de Bioinformática (Hinxton, Reino Unido).

En algunas realizaciones, se produce un alineamiento de secuencias -como, por ejemplo, un mapa de alineamiento de secuencias (SAM) o un mapa de alineamiento binario (BAM)- que comprende una cadena CIGAR (el formato SAM se describe, por ejemplo, en Li et al., "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, 25(16):2078-9, 2009). En algunas realizaciones, CIGAR muestra o incluye alineaciones con espacios una por línea. CIGAR es un formato de alineación por pares comprimido que se notifica como una cadena CIGAR. Una cadena CIGAR puede ser útil para representar alineaciones por pares largas (por ejemplo, genómicas). Una cadena CIGAR puede usarse en formato SAM para representar alineaciones de lecturas con una secuencia genómica de referencia.

Una cadena CIGAR puede seguir un motivo establecido. Cada carácter va precedido de un número, dando los recuentos de bases del evento. Los caracteres usados pueden incluir M, I, D, N, y S (M=coincidencia; I=inserción; D=delección; N=hueco; S=sustitución). La cadena CIGAR define la secuencia de coincidencias y/o mal apareamientos y deleciones (o huecos). Por ejemplo, la cadena CIGAR 2MD3M2D2M puede indicar que el alineamiento contiene 2 coincidencias, 1 delección (se omite el número 1 para ahorrar espacio), 3 coincidencias, 2 deleciones y 2 coincidencias.

En algunas realizaciones, se prepara una población de ácidos nucleicos para secuenciación formando enzimáticamente extremos romos en ácidos nucleicos de cadena doble con salientes de cadena sencilla en uno o ambos extremos. En estas realizaciones, la población se trata típicamente con una enzima que tiene una actividad 5'-3' de ADN polimerasa y una actividad 3'-5' de exonucleasa en presencia de los nucleótidos (por ejemplo, A, C, G y T o U). Ejemplos de enzimas o fragmentos catalíticos de las mismas que pueden usarse opcionalmente incluyen el fragmento grande de Klenow y la polimerasa T4. En los salientes de 5', la enzima típicamente extiende el extremo rebajado 3' en la cadena opuesta hasta que se alinea con el extremo 5' para producir un extremo romo. En los salientes 3', la enzima generalmente digiere desde el extremo 3' hasta y a veces más allá del extremo 5' de la cadena opuesta. Si esta digestión sobrepasa el extremo 5' de la cadena opuesta, el hueco puede rellenarse con una enzima que tenga la misma actividad polimerasa que la usada para los salientes 5'. La formación de extremos romos en ácidos nucleicos de cadena doble facilita, por ejemplo, la unión de adaptadores y la posterior amplificación.

En algunas realizaciones, las poblaciones de ácidos nucleicos se someten a procesamiento adicional, como la conversión de ácidos nucleicos de cadena sencilla en ácidos nucleicos de cadena doble y/o la conversión de ARN en ADN (por ejemplo, ADN complementario o ADNc). Estas formas de ácido nucleico también se enlazan opcionalmente a adaptadores y se amplifican.

Con o sin amplificación previa, los ácidos nucleicos sometidos al proceso de formación de extremos romos descrito anteriormente, y opcionalmente otros ácidos nucleicos en una muestra, pueden secuenciarse para producir ácidos nucleicos secuenciados. Un ácido nucleico secuenciado puede referirse a la secuencia de un ácido nucleico (por ejemplo, información de secuencia) o a un ácido nucleico cuya secuencia se ha determinado. La secuenciación puede realizarse para proporcionar datos de secuencia de moléculas de ácido nucleico individuales en una muestra, ya sea directa o indirectamente a partir de una secuencia de consenso de productos de amplificación de una molécula

de ácido nucleico individual en la muestra.

En algunas realizaciones, los ácidos nucleicos de cadena doble con salientes de cadena sencilla en una muestra después de la formación del extremo romo se enlazan en ambos extremos a adaptadores que incluyen códigos de barras, y la secuenciación determina las secuencias de ácidos nucleicos así como los códigos de barras en línea introducidos por los adaptadores. Las moléculas de ADN con extremo romo se ligan opcionalmente a un extremo romo de un adaptador por lo menos parcialmente de cadena doble (por ejemplo, un adaptador con forma de Y o de campana). Alternativamente, a los extremos romos de los ácidos nucleicos de la muestra y los adaptadores se les pueden añadir una cola con nucleótidos complementarios para facilitar la ligación (por ejemplo, ligación de extremos pegajosos).

La muestra de ácido nucleico se pone en contacto típicamente con un número suficiente de adaptadores para que haya una probabilidad baja (por ejemplo, menos de aproximadamente el 1 o el 0,1 %) de que dos copias cualesquiera del mismo ácido nucleico reciban la misma combinación de códigos de barras de adaptador de los adaptadores enlazados en ambos extremos. El uso de adaptadores de esta manera puede permitir la identificación de familias de secuencias de ácidos nucleicos con los mismos puntos de inicio y parada en un ácido nucleico de referencia y enlazadas a la misma combinación de códigos de barras. Dicha familia puede representar secuencias de productos de amplificación de un ácido nucleico en la muestra antes de la amplificación. Las secuencias de los miembros de la familia pueden compilarse para derivar nucleótido o nucleótidos consenso o una secuencia consenso completa para una molécula de ácido nucleico en la muestra original, modificada por la formación del extremo romo y la unión del adaptador. En otras palabras, el nucleótido que ocupa una posición específica de un ácido nucleico en la muestra puede determinarse como el consenso de los nucleótidos que ocupan esa posición correspondiente en las secuencias de los miembros de la familia. Las familias pueden incluir secuencias de una o ambas cadenas de un ácido nucleico de cadena doble. Si los miembros de una familia incluyen secuencias de ambas cadenas de un ácido nucleico de cadena doble, las secuencias de una cadena pueden convertirse en sus complementos a efectos de compilar secuencias para derivar nucleótido o nucleótidos o secuencias de consenso. Algunas familias incluyen una única secuencia miembro. En este caso, esta secuencia puede tomarse como la secuencia de un ácido nucleico en la muestra antes de la amplificación. Alternativamente, pueden eliminarse del análisis posterior las familias con una única secuencia miembro.

Las variaciones de nucleótidos (por ejemplo, SNV o indeles) en ácidos nucleicos secuenciados pueden determinarse comparando los ácidos nucleicos secuenciados con una secuencia de referencia. La secuencia de referencia es a menudo una secuencia conocida, por ejemplo, una secuencia genómica completa o parcial conocida de un sujeto (por ejemplo, una secuencia genómica completa de un sujeto humano). La secuencia de referencia puede ser, por ejemplo, hG19 o hG38. Los ácidos nucleicos secuenciados pueden representar secuencias determinadas directamente para un ácido nucleico de una muestra, o un consenso de secuencias de productos de amplificación de dicho ácido nucleico, como se ha descrito anteriormente. Puede realizarse una comparación en una o más posiciones designadas de una secuencia de referencia. Puede identificarse un subconjunto de ácidos nucleicos secuenciados que incluya una posición correspondiente con una posición designada de la secuencia de referencia cuando están alineadas al máximo las secuencias respectivas. Dentro de dicho subconjunto puede determinarse qué ácidos nucleicos secuenciados, si los hay, incluyen una variación de nucleótidos en la posición designada y, opcionalmente cuáles, si los hay, incluyen un nucleótido de referencia (por ejemplo, el mismo que en la secuencia de referencia). Si el número de ácidos nucleicos secuenciados en el subconjunto que incluye una variante nucleotídica excede un umbral seleccionado, entonces puede llamarse a un nucleótido variante en la posición designada. El umbral puede ser un número simple, tal como por lo menos 1, 2, 3, 4, 5, 6, 7, 8, 9, o 10 ácidos nucleicos secuenciados dentro del subconjunto que incluya la variante nucleotídica o puede ser una proporción, como por lo menos 0,5, 1, 2, 3, 4, 5, 10, 15 o 20, de ácidos nucleicos secuenciados dentro del subconjunto que incluyen la variante nucleotídica, entre otras posibilidades. La comparación puede repetirse para cualquier posición de interés designada en la secuencia de referencia. A veces puede realizarse una comparación para posiciones designadas que ocupan por lo menos aproximadamente 20, 100, 200 o 300 posiciones contiguas en una secuencia de referencia, por ejemplo, aproximadamente 20-500, o aproximadamente 50-300 posiciones contiguas.

Detalles adicionales referentes a la secuenciación de ácidos nucleicos, incluyendo los formatos y las aplicaciones descritos en la presente también se proporcionan, por ejemplo, en Levy et al., Annual Review of Genomics and Human Genetics, 17: 95-115 (2016), Liu et al., J. of Biomedicine and Biotechnology, Volume 2012, Article ID 251364: 1-11 (2012), Voelkerding et al., Clinical Chem, 55: 641-658 (2009), MacLean et al., Nature Rev. Microbial., 7: 287-296 (2009), Astier et al., J Am Chem Soc., 128(5):1705-10 (2006), Patente de Estados Unidos Nº 6.210.891, Patente de Estados Unidos Nº 6.258.568, Patente de Estados Unidos Nº 6.833.246, Patente de Estados Unidos Nº 7.115.400, Patente de Estados Unidos Nº 6.969.488, Patente de Estados Unidos Nº 5.912.148, Patente de Estados Unidos Nº 6.130.073, Patente de Estados Unidos Nº 7.169.560, Patente de Estados Unidos Nº 7.282.337, Patente de Estados Unidos Nº 7.482.120, Patente de Estados Unidos Nº 7.501.245, Patente de Estados Unidos Nº 6.818.395, Patente de Estados Unidos Nº 6.911.345, Patente de Estados Unidos Nº 7.501.245, Patente de Estados Unidos Nº 7.329.492, Patente de Estados Unidos Nº 7.170.050, Patente de Estados Unidos Nº 7.302.146, Patente de Estados Unidos Nº 7.313.308, y Patente de Estados Unidos Nº 7.476.503.

IV. COLECCIONES DE SONDAS ESPECÍFICAS DE LA DIANA; COMPOSICIONES

1. Colecciones de sondas específicas

Una colección útil de sondas específicas de la diana comprende sondas de unión a diana específicas para un conjunto de regiones diana de secuencia variable y sondas de unión a diana específicas para un conjunto de regiones diana epigenéticas. En algunas realizaciones, el rendimiento de captura de las sondas de unión a diana específicas para el conjunto de regiones diana de secuencia variable es mayor (por ejemplo, por lo menos 2 veces mayor) que el rendimiento de captura de las sondas de unión a diana específicas para el conjunto de regiones diana epigenéticas. En algunas realizaciones, la colección de sondas específicas de la diana está configurada para que tenga un rendimiento de captura específico para el conjunto de regiones diana de secuencia variable mayor (por ejemplo, por lo menos 2 veces mayor) que su rendimiento de captura específico para el conjunto de regiones diana epigenéticas.

En algunas realizaciones, el rendimiento de captura de las sondas de unión a diana específicas para el conjunto de regiones diana de secuencia variable es de por lo menos 1,25-, 1,5-, 1,75-, 2-, 2,25-, 2,5-, 2,75-, 3-, 3,5-, 4-, 4,5-, 5-, 6-, 7-, 8-, 9-, 10-, 11-, 12-, 13-, 14- o 15 veces mayor que el rendimiento de captura de las sondas de unión a diana específicas para el conjunto de regiones diana epigenéticas. En algunas realizaciones, el rendimiento de captura de las sondas de unión a diana específicas para el conjunto de regiones diana de secuencia variable es de 1,25- a 1,5-, de 1,5- a 1,75-, de 1,75- a 2-, de 2- a 2,25-, de 2,25- a 2,5-, de 2,5- a 2,75-, de 2,75- a 3-, de 3- a 3,5-, de 3,5- a 4-, de 4- a 4,5-, de 4,5- a 5-, de 5- a 5,5-, de 5,5- a 6-, de 6- a 7-, de 7- a 8-, de 8- a 9-, de 9- a 10-, de 10- a 11-, de 11- a 12-, de 12- a 13-, de 13- a 14-, o de 14- a 15 veces mayor que el rendimiento de captura de las sondas de unión a diana específicas para el conjunto de regiones diana epigenéticas. En algunas realizaciones, el rendimiento de captura de las sondas de unión a diana específicas para el conjunto de regiones diana de secuencia variable es por lo menos 10 veces mayor que el rendimiento de captura de las sondas de unión a diana específicas para el conjunto de regiones diana epigenéticas, por ejemplo, de 10 a 20 veces mayor que el rendimiento de captura de las sondas de unión a diana específicas para el conjunto de regiones diana epigenéticas.

En algunas realizaciones, la colección de sondas específicas de la diana está configurada para que tenga un rendimiento de captura específico para el conjunto de regiones diana de secuencia variable por lo menos 1,25-, 1,5-, 1,75-, 2-, 2,25-, 2,5-, 2,75-, 3-, 3,5-, 4-, 4,5-, 5-, 6-, 7-, 8-, 9-, 10-, 11-, 12-, 13-, 14-, o 15 veces mayor que su rendimiento de captura para el conjunto de regiones diana epigenéticas. En algunas realizaciones, la colección de sondas específicas de la diana está configurada para que tenga un rendimiento de captura específico para el conjunto de regiones diana de secuencia variable de 1,25- a 1,5-, de 1,5- a 1,75-, de 1,75- a 2-, de 2- a 2,25-, de 2,25- a 2,5-, de 2,5- a 2,75-, de 2,75- a 3-, de 3- a 3,5-, de 3,5- a 4-, de 4- a 4,5-, de 4,5- a 5-, de 5- a 5,5-, de 5,5- a 6-, de 6- a 7-, de 7- a 8-, de 8- a 9-, de 9- a 10-, de 10- a 11-, de 11- a 12-, de 12- a 13-, de 13- a 14-, o de 14- a 15 veces mayor que su rendimiento de captura específico para el conjunto de regiones diana epigenéticas. En algunas realizaciones, la colección de sondas específicas de la diana está configurada para que tenga un rendimiento de captura específico para el conjunto de regiones diana de secuencia variable por lo menos 10 veces mayor que su rendimiento de captura para el conjunto de regiones diana epigenéticas, por ejemplo, de 10 a 20 veces mayor que su rendimiento de captura para el conjunto de regiones diana epigenéticas.

La colección de sondas puede configurarse para proporcionar mayores rendimientos de captura para la región diana de secuencia variable establecida de varias maneras, incluyendo concentración, diferentes longitudes y/o químicas (por ejemplo, que afecten a la afinidad), y combinaciones de las mismas. La afinidad puede modularse ajustando la longitud de la sonda y/o incluyendo modificaciones de nucleótidos, como se analiza a continuación.

En algunas realizaciones, las sondas específicas de la diana para el conjunto de regiones diana de secuencia variable están presentes a una concentración mayor que las sondas específicas de la diana para el conjunto de regiones diana epigenéticas. En algunas realizaciones, la concentración de las sondas de unión a diana específicas para el conjunto de regiones diana de secuencia variable es por lo menos 1,25-, 1,5-, 1,75-, 2-, 2,25-, 2,5-, 2,75-, 3-, 3,5-, 4-, 4,5-, 5-, 6-, 7-, 8-, 9-, 10-, 11-, 12-, 13-, 14-, o 15 veces mayor que la concentración de las sondas de unión a diana específicas para el conjunto de regiones diana epigenéticas. En algunas realizaciones, la concentración de las sondas de unión a diana específicas para el conjunto de regiones diana de secuencia variable es de 1,25- a 1,5-, de 1,5- a 1,75-, de 1,75- a 2-, de 2- a 2,25-, de 2,25- a 2,5-, de 2,5- a 2,75-, de 2,75- a 3-, de 3- a 3,5-, de 3,5- a 4-, de 4- a 4,5-, de 4,5- a 5-, de 5- a 5,5-, de 5,5- a 6-, de 6- a 7-, de 7- a 8-, de 8- a 9-, de 9- a 10-, de 10- a 11-, de 11- a 12-, de 12- a 13-, de 13- a 14-, o de 14- a 15 veces mayor que la concentración de las sondas de unión a diana específicas para el conjunto de regiones diana epigenéticas. En algunas realizaciones, la concentración de las sondas de unión a diana específicas para el conjunto de regiones diana de secuencia variable es por lo menos 2 veces mayor que la concentración de las sondas de unión a diana específicas para el conjunto de regiones diana epigenéticas. En algunas realizaciones, la concentración de las sondas de unión a diana específicas para el conjunto de regiones diana de secuencia variable es por lo menos 10 veces mayor que la concentración de las sondas de unión a diana específicas para el conjunto de regiones diana epigenéticas, por ejemplo, de 10 a 20 veces mayor que la concentración de las sondas de unión a diana específicas para el conjunto de regiones diana epigenéticas. En tales realizaciones, la concentración puede referirse a la concentración media de masa por volumen de sondas individuales en cada conjunto.

En algunas realizaciones, las sondas específicas de la diana para el conjunto de regiones diana de secuencia variable tienen una mayor afinidad por sus dianas que las sondas específicas de la diana para el conjunto de regiones diana epigenéticas. La afinidad puede modularse de cualquier forma conocida por los expertos en la técnica, incluyendo el uso de diferentes químicas de sonda. Por ejemplo, ciertas modificaciones de nucleótidos, como la 5-metilación de citosina (en ciertos contextos de secuencia), modificaciones que proporcionan un heteroátomo en la posición de azúcar 2', y nucleótidos LNA, pueden aumentar la estabilidad de los ácidos nucleicos de cadena doble, lo que indica que los oligonucleótidos con tales modificaciones tienen una afinidad relativamente mayor por sus secuencias complementarias. Véanse, por ejemplo, Severin et al., *Nucleic Acids Res.* 39: 8740- 8751 (2011); Freier et al., *Nucleic Acids Res.* 25: 4429-4443 (1997); Patente de Estados Unidos N° 9.738.894. Además, las longitudes de secuencia más largas proporcionarán generalmente una mayor afinidad. Otras modificaciones de nucleótidos, como la sustitución de la nucleobase hipoxantina por guanina, reducen la afinidad al reducir la cantidad de enlace de hidrógeno entre el oligonucleótido y su secuencia complementaria. En algunas realizaciones, las sondas específicas de la diana para el conjunto de regiones diana de secuencia variable tienen modificaciones que aumentan su afinidad por sus dianas. En algunas realizaciones, alternativa o adicionalmente, las sondas específicas de la diana para el conjunto de regiones diana epigenéticas tienen modificaciones que disminuyen su afinidad por sus dianas. En algunas realizaciones, las sondas específicas de la diana para el conjunto de regiones diana de secuencia variable tienen longitudes medias más largas y/o temperaturas medias de fusión más altas que las sondas específicas de la diana para el conjunto de regiones diana epigenéticas. Estas realizaciones pueden combinarse entre sí y/o con diferencias de concentración, como se ha analizado anteriormente, para conseguir una diferencia de veces deseada en el rendimiento de captura, como cualquier diferencia de veces o intervalo de los mismos descrito anteriormente.

En algunas realizaciones, las sondas específicas de la diana comprenden una fracción de captura. La fracción de captura puede ser cualquiera de las fracciones de captura descritas en la presente, por ejemplo, biotina. En algunas realizaciones, las sondas específicas de la diana están enlazadas a un soporte sólido, por ejemplo, de manera covalente o no covalente, por ejemplo, a través de la interacción de un par de unión de fracciones de captura. En algunas realizaciones, el soporte sólido es una perla, como una perla magnética.

En algunas realizaciones, las sondas específicas de la diana para el conjunto de regiones diana de secuencia variable y/o las sondas específicas de la diana para el conjunto de regiones diana epigenéticas son un conjunto de cebo como se ha analizado con anterioridad, por ejemplo, sondas que comprenden elementos de captura y secuencias seleccionadas para formar segmentos a través de un panel de regiones, como genes.

En algunas realizaciones, las sondas específicas de la diana se proporcionan en una única composición. La única composición puede ser una solución (líquida o congelada). Alternativamente, puede ser un liofilizado.

Alternativamente, las sondas específicas de la diana pueden proporcionarse como una pluralidad de composiciones, por ejemplo, comprendiendo una primera composición que comprende sondas específicas para el conjunto de regiones diana epigenéticas y una segunda composición que comprende sondas específicas para el conjunto de regiones diana de secuencia variable. Estas sondas pueden mezclarse en proporciones adecuadas para proporcionar una composición de sonda combinada con cualquiera de las diferencias de veces anteriores en concentración y/o rendimiento de captura. Alternativamente, pueden usarse en procedimientos de captura separados (por ejemplo, con alícuotas de una muestra o secuencialmente con la misma muestra) para proporcionar una primera y una segunda composiciones que comprendan regiones diana epigenéticas capturadas y regiones diana de secuencia variable, respectivamente.

a. Sondas específicas para regiones diana epigenéticas

Las sondas para el conjunto de regiones diana epigenéticas pueden comprender sondas específicas para uno o más tipos de regiones diana que diferencian ADN de células neoplásicas (por ejemplo, tumorales o cancerosas) de células sanas, por ejemplo, células circulantes no neoplásicas. Los tipos ejemplares de tales regiones se analizan en detalle en la presente, por ejemplo, en las secciones anteriores referentes a los conjuntos capturados. Las sondas para el conjunto de regiones diana epigenéticas también pueden comprender sondas para una o más regiones de control, por ejemplo, como se describe en la presente.

En algunas realizaciones, las sondas para el conjunto de sondas de regiones diana epigenéticas tienen una huella de por lo menos 100 kb, por ejemplo, por lo menos 200 kb, por lo menos 300 kb o por lo menos 400 kb. En algunas realizaciones, las sondas para el conjunto de regiones diana epigenéticas tienen una huella en el intervalo de 100-1000 kb, por ejemplo, 100-200 kb, 200-300 kb, 300-400 kb, 400-500 kb, 500-600 kb, 600-700 kb, 700-800 kb, 800-900 kb, y 900-1.000 kb.

i. Regiones diana variables de hipermetilación

En algunas realizaciones, las sondas para el conjunto de regiones diana epigenéticas comprenden sondas específicas para una o más regiones diana variables de hipermetilación. Las regiones diana variables de

hipermetilación pueden ser cualquiera de las expuestas anteriormente. Por ejemplo, en algunas realizaciones, las sondas específicas para regiones diana variables de hipermetilación comprenden sondas específicas para una pluralidad de loci enumerados en la Tabla 1, por ejemplo, por lo menos un 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, o 100% de los loci enumerados en la Tabla 1. En algunas realizaciones, las sondas específicas para regiones diana variables de hipermetilación comprenden sondas específicas para una pluralidad de loci enumerados en la Tabla 2, por ejemplo, por lo menos un 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, o 100% de los loci enumerados en la Tabla 2. En algunas realizaciones, las sondas específicas para regiones diana variables de hipermetilación comprenden sondas específicas para una pluralidad de loci enumerados en la Tabla 1 o Tabla 2, por ejemplo, por lo menos un 10%, 20%, 30%, 40%, 50%,

60%, 70%, 80%, 90% o 100% de los loci enumerados en la Tabla 1 o la Tabla 2. En algunas realizaciones, para cada locus incluido como región diana, puede haber una o más sondas con un sitio de hibridación que se une entre el sitio de inicio de la transcripción y el codón de parada (el último codón de parada para genes que están cortados y empalmados alternativamente) del gen. En algunas realizaciones, la una o más sondas se unen dentro de 300 pb de la posición indicada, por ejemplo, dentro de 200 o 100 pb. En algunas realizaciones, una sonda tiene un sitio de hibridación que se solapa con la posición enumerada anteriormente. En algunas realizaciones, las sondas específicas para las regiones diana de hipermetilación incluyen sondas específicas para uno, dos, tres, cuatro o cinco subconjuntos de regiones diana de hipermetilación que muestran colectivamente hipermetilación en uno, dos, tres, cuatro o cinco de los cánceres de mama, colon, riñón, hígado y pulmón.

ii. Regiones diana variables de hipometilación

En algunas realizaciones, las sondas para el conjunto de regiones diana epigenéticas comprenden sondas específicas para una o más regiones diana variables de hipometilación. Las regiones diana variables de hipometilación pueden ser cualquiera de las expuestas anteriormente. Por ejemplo, las sondas específicas para una o más regiones diana variables de hipometilación pueden incluir sondas para regiones como elementos repetidos, por ejemplo, elementos LINE1, elementos Alu, repeticiones en tándem centroméricas, repeticiones en tándem pericentroméricas, y ADN satélite, y regiones intergénicas que normalmente están metiladas en células sanas pueden mostrar metilación reducida en células tumorales.

En algunas realizaciones, las sondas específicas para regiones diana variables de hipometilación incluyen sondas específicas para elementos repetidos y/o regiones intergénicas. En algunas realizaciones, las sondas específicas para elementos repetidos incluyen sondas específicas para uno, dos, tres, cuatro o cinco elementos LINE1, elementos Alu, repeticiones en tándem centroméricas, repeticiones en tándem pericentroméricas y/o ADN satélite.

Sondas ejemplares específicas para regiones genómicas que muestran hipometilación asociada a cáncer incluyen sondas específicas para los nucleótidos 8403565-8953708 y/o 151104701- 151106035 del cromosoma humano 1. En algunas realizaciones, las sondas específicas para regiones diana variables de hipometilación incluyen sondas específicas para regiones que se solapan o comprenden los nucleótidos 8403565-8953708 y/o 151104701- 151106035 del cromosoma 1 humano.

iii. Regiones de unión a CTCF

En algunas realizaciones, las sondas para el conjunto de regiones diana epigenéticas incluyen sondas específicas para regiones de unión a CTCF. En algunas realizaciones, las sondas específicas para regiones de unión a CTCF comprenden sondas específicas para por lo menos 10, 20, 50, 100, 200 o 500 regiones de unión a CTCF, o 10-20, 20-50, 50-100, 100-200, 200-500, o 500-1000 regiones de unión a CTCF, por ejemplo, como las regiones de unión a CTCF descritas anteriormente o en uno o más de CTCFBSDB o los artículos de Cuddapah et al., Martin et al. o Rhee et al. citados anteriormente. En algunas realizaciones, las sondas para el conjunto de regiones diana epigenéticas comprenden por lo menos 100 pb, por lo menos 200 pb, por lo menos 300 pb, por lo menos 400 pb, por lo menos 500 pb, por lo menos 750 pb, o por lo menos 1000 pb de regiones en sentido ascendente y en sentido descendente de los sitios de unión a CTCF.

iv. Sitios de inicio de la transcripción

En algunas realizaciones, las sondas para el conjunto de regiones diana epigenéticas incluyen sondas específicas para sitios de inicio transcripcional. En algunas realizaciones, las sondas específicas para sitios de inicio transcripcional comprenden sondas específicas para por lo menos 10, 20, 50, 100, 200 o 500 sitios de inicio transcripcional, o 10-20, 20-50, 50-100, 100-200, 200-500, o 500-1000 sitios de inicio transcripcional, por ejemplo, tales como sitios de inicio transcripcional enumerados en DBTSS. En algunas realizaciones, las sondas para el conjunto de regiones diana epigenéticas comprenden sondas para secuencias de por lo menos 100 pb, por lo menos 200 pb, por lo menos 300 pb, por lo menos 400 pb, por lo menos 500 pb, por lo menos 750 pb o por lo menos 1000 pb en sentido ascendente y en sentido descendente de los sitios de inicio transcripcional.

v. Amplificaciones focales

Como se ha indicado anteriormente, aunque las amplificaciones focales son mutaciones somáticas, pueden detectarse mediante secuenciación basada en la frecuencia de lectura de manera análoga a los enfoques para detectar ciertos cambios epigenéticos como los cambios en la metilación.

vi. Regiones de control

Puede ser útil incluir regiones de control para facilitar la validación de los datos. En algunas realizaciones, las sondas específicas para el conjunto de regiones diana epigenéticas incluyen sondas específicas para regiones metiladas de control que se espera que estén metiladas en esencialmente todas las muestras. En algunas realizaciones, las sondas específicas para el conjunto de regiones diana epigenéticas incluyen sondas específicas para regiones hipometiladas de control que se espera que estén hipometiladas en esencialmente todas las muestras.

b. Sondas específicas para regiones diana de secuencia variable

Las sondas para el conjunto de regiones diana de secuencia variable pueden comprender sondas específicas para una pluralidad de regiones que se sabe que sufren mutaciones somáticas en el cáncer. Las sondas pueden ser específicas para cualquier conjunto de regiones diana de secuencia variable descrito en la presente. Los conjuntos de regiones diana de secuencia variable ejemplares se tratan en detalle en la presente, por ejemplo, en las secciones anteriores referentes a los conjuntos capturados.

En algunas realizaciones, el conjunto de sondas de regiones diana de secuencia variable tiene una huella de por lo menos 10 kb, por ejemplo, por lo menos 20 kb, por lo menos 30 kb o por lo menos 40 kb. En algunas realizaciones, el conjunto de sondas de regiones diana epigenéticas tiene una huella en el intervalo de 10-100 kb, por ejemplo, 10-20 kb, 20-30 kb, 30-40 kb, 40-50 kb, 50-60 kb, 60-70 kb, 70-80 kb, 80-90 kb, y 90-100 kb.

En algunas realizaciones, las sondas específicas para el conjunto de regiones diana de secuencia variable comprenden sondas específicas para por lo menos una parte de por lo menos 5, por lo menos 10, por lo menos 15, por lo menos 20, por lo menos 25, por lo menos 30, por lo menos 35, por lo menos 40, por lo menos 45, por lo menos 50, por lo menos 55, por lo menos 60, por lo menos 65, o 70 de los genes de la Tabla 3. En algunas realizaciones, las sondas específicas para el conjunto de regiones diana de secuencia variable comprenden sondas específicas para por lo menos 5, por lo menos 10, por lo menos 15, por lo menos 20, por lo menos 25, por lo menos 30, por lo menos 35, por lo menos 40, por lo menos 45, por lo menos 50, por lo menos 55, por lo menos 60, por lo menos 65, o 70 de los SNV de la Tabla 3. En algunas realizaciones, las sondas específicas para el conjunto de regiones diana de secuencia variable comprenden sondas específicas para por lo menos 1, por lo menos 2, por lo menos 3, por lo menos 4, por lo menos 5, o 6 de las fusiones de la Tabla 3. En algunas realizaciones, las sondas específicas para el conjunto de regiones diana de secuencia variable comprenden sondas específicas para por lo menos una parte de por lo menos 1, por lo menos 2, o 3 de los indeles de la Tabla 3. En algunas realizaciones, las sondas específicas para el conjunto de regiones diana de secuencia variable comprenden sondas específicas para por lo menos una parte de por lo menos 5, por lo menos 10, por lo menos 15, por lo menos 20, por lo menos 25, por lo menos 30, por lo menos 35, por lo menos 40, por lo menos 45, por lo menos 50, por lo menos 55, por lo menos 60, por lo menos 65, por lo menos 70, o 73 de los genes de la Tabla 4. En algunas realizaciones, las sondas específicas para el conjunto de regiones diana de secuencia variable comprenden sondas específicas para por lo menos 5, por lo menos 10, por lo menos 15, por lo menos 20, por lo menos 25, por lo menos 30, por lo menos 35, por lo menos 40, por lo menos 45, por lo menos 50, por lo menos 55, por lo menos 60, por lo menos 65, por lo menos 70, o 73 de las SNV de la Tabla 4. En algunas realizaciones, las sondas específicas para el conjunto de regiones diana de secuencia variable comprenden sondas específicas para por lo menos 1, por lo menos 2, por lo menos 3, por lo menos 4, por lo menos 5, o 6 de las fusiones de la Tabla 4. En algunas realizaciones, las sondas específicas para el conjunto de regiones diana de secuencia variable comprenden sondas específicas para por lo menos una parte de por lo menos 1, por lo menos 2, por lo menos 3, por lo menos 4, por lo menos 5, por lo menos 6, por lo menos 7, por lo menos 8, por lo menos 9, por lo menos 10, por lo menos 11, por lo menos 12, por lo menos 13, por lo menos 14, por lo menos 15, por lo menos 16, por lo menos 17, o 18 de los indeles de la Tabla 4. En algunas realizaciones, las sondas específicas para el conjunto de regiones diana de secuencia variable comprenden sondas específicas para por lo menos una parte de por lo menos 1, por lo menos 2, por lo menos 3, por lo menos 4, por lo menos 5, por lo menos 6, por lo menos 7, por lo menos 8, por lo menos 9, por lo menos 10, por lo menos 11, por lo menos 12, por lo menos 13, por lo menos 14, por lo menos 15, por lo menos 16, por lo menos 17, por lo menos 18, por lo menos 19, o por lo menos 20 de los genes de la Tabla 5.

En algunas realizaciones, las sondas específicas para el conjunto de regiones diana de secuencia variable comprenden sondas específicas para regiones diana de por lo menos 10, 20, 30 o 35 genes relacionados con el cáncer, como AKT1, ALK, BRAF, CCND1, CDK2A, CTNNB1, EGFR, ERBB2, ESR1, FGFR1, FGFR2, FGFR3, FOXL2, GATA3, GNA11, GNAQ, GNAS, HRAS, IDH1, IDH2, KIT, KRAS, MED12, MET, MYC, NFE2L2, NRAS, PDGFRA, PIK3CA, PPP2R1A, PTEN, RET, STK11, TP53 y U2AF1.

c. Composición de las sondas

En algunas realizaciones, se proporciona una única composición que comprende sondas para el conjunto de regiones diana de secuencia variable y sondas para el conjunto de regiones diana epigenéticas. Las sondas pueden proporcionarse en dicha composición en cualquier proporción de concentración descrita en la presente.

En algunas realizaciones, se proporciona una primera composición que comprende sondas para el conjunto de regiones diana epigenéticas y una segunda composición que comprende sondas para el conjunto de regiones diana de secuencia variable. La proporción entre la concentración de las sondas en la primera composición y la concentración de las sondas en la segunda composición puede ser cualquiera de las proporciones descritas en la presente.

2. Composiciones que comprenden ADNc capturado

El ADNc capturado puede tener cualquiera de las características descritas en la presente referentes a los conjuntos capturados incluyendo, por ejemplo, una mayor concentración del ADN correspondiente al conjunto de regiones diana de secuencia variable (normalizado para el tamaño de la huella como se ha analizado anteriormente) que del ADN correspondiente al conjunto de regiones diana epigenéticas. En algunas realizaciones, el ADNc del conjunto capturado comprende etiquetas de secuencia, que pueden añadirse al ADNc como se describe en la presente. En general, la inclusión de etiquetas de secuencia hace que las moléculas de ADNc difieran de su forma de origen natural no etiquetada.

Tales composiciones pueden comprender además un conjunto de sondas descrito en la presente o cebadores de secuenciación, cada uno de los cuales puede diferir de moléculas de ácido nucleico de origen natural. Por ejemplo, un conjunto de sondas descrito en la presente puede comprender una fracción de captura, y los cebadores de secuenciación pueden comprender una etiqueta de origen no natural.

V. SISTEMAS INFORMÁTICOS

Los métodos de la presente divulgación pueden implementarse usando, o con la ayuda de, sistemas informáticos. Por ejemplo, tales métodos, que pueden comprender: recoger ADNc de un sujeto de prueba; capturar una pluralidad de conjuntos de regiones diana del ADNc, en donde la pluralidad de conjuntos de regiones diana comprende un conjunto de regiones diana de secuencia variable y un conjunto de regiones diana epigenéticas, por lo que se produce un conjunto capturado de moléculas de ADNc; secuenciar las moléculas de ADNc capturadas, en donde las moléculas de ADNc capturadas del conjunto de regiones diana de secuencia variable se secuencian a una mayor profundidad de secuenciación que las moléculas de ADNc capturadas del conjunto de regiones diana epigenéticas; obtener una pluralidad de lecturas de secuencias generadas por un secuenciador de ácidos nucleicos a partir de la secuenciación de las moléculas de ADNc capturadas; mapear la pluralidad de lecturas de secuencias con una o más secuencias de referencia para generar lecturas de secuencias mapeadas; y procesar las lecturas de secuencias mapeadas correspondientes al conjunto de regiones diana de secuencia variable y al conjunto de regiones diana epigenéticas para determinar la probabilidad de que el sujeto tenga cáncer.

La FIG. 2 muestra un sistema informático 201 que está programado o configurado de otro modo para implementar los métodos de la presente divulgación. El sistema informático 201 puede regular varios aspectos preparación, secuenciación, y/o análisis de muestras. En algunos ejemplos, el sistema informático 201 se configura para realizar la preparación de la muestra y el análisis de la muestra, incluyendo la secuenciación del ácido nucleico.

El sistema informático 201 incluye una unidad central de procesamiento (CPU, también "procesador" y "procesador informático" en la presente) 205, que puede ser un procesador de un solo núcleo o de múltiples núcleos, o una pluralidad de procesadores para el procesamiento en paralelo. El sistema informático 201 también incluye memoria o ubicación de memoria 210 (por ejemplo, memoria de acceso aleatorio, memoria de sólo lectura, memoria flash), unidad de almacenamiento electrónico 215 (por ejemplo, disco duro), interfaz de comunicación 220 (por ejemplo, adaptador de red) para comunicarse con uno o más sistemas distintos, y dispositivos periféricos 225, como caché, otra memoria, almacenamiento de datos y/o adaptadores de pantalla electrónica. La memoria 210, la unidad de almacenamiento 215, la interfaz 220 y los dispositivos periféricos 225 están en comunicación con la CPU 205 a través de una red o bus de comunicación (líneas continuas), como una placa base. La unidad de almacenamiento 215 puede ser una unidad de almacenamiento de datos (o repositorio de datos) para almacenar datos. El sistema informático 201 puede acoplarse operativamente a una red informática 230 con la ayuda de la interfaz de comunicación 220. La red informática 230 puede ser Internet, una Internet y/o extranet, o una intranet y/o extranet que está en comunicación con Internet. En algunos casos, la red informática 230 es una red de telecomunicaciones y/o de datos. La red informática 230 puede incluir uno o más servidores informáticos, que pueden permitir la computación distribuida, como la computación en nube. La red informática 230, en algunos casos con la ayuda del sistema informático 201, puede implementar una red peer-to-peer, que puede permitir que los dispositivos acoplados al sistema informático 201 se comporten como un cliente o un servidor.

La CPU 205 puede ejecutar una secuencia de instrucciones legibles por máquina, que pueden estar incorporadas en un programa o software. Las instrucciones pueden almacenarse en una ubicación de memoria, como

la memoria 210. Ejemplos de operaciones realizadas por la CPU 205 pueden incluir búsqueda, decodificación, ejecución y escritura.

La unidad de almacenamiento 215 puede almacenar archivos, como controladores, bibliotecas y programas guardados. La unidad de almacenamiento 215 puede almacenar programas generados por usuarios y sesiones grabadas, así como la salida o salidas asociadas a los programas. La unidad de almacenamiento 215 puede almacenar datos de usuario, por ejemplo, preferencias de usuario y programas de usuario. El sistema informático 201 en algunos casos puede incluir una o más unidades de almacenamiento de datos adicionales que son externas al sistema informático 201, como ubicadas en un servidor remoto que está en comunicación con el sistema informático 201 a través de una intranet o Internet. Los datos pueden transferirse de una ubicación a otra usando, por ejemplo, una red de comunicación o una transferencia física de datos (por ejemplo, usando un disco duro, una memoria USB u otro mecanismo de almacenamiento de datos).

El sistema informático 201 puede comunicarse con uno o más sistemas informáticos remotos a través de la red 230. Para la realización, el sistema informático 201 puede comunicarse con un sistema informático remoto de un usuario (por ejemplo, un operador). Ejemplos de sistemas informáticos remotos incluyen ordenadores personales (por ejemplo, PC portátiles), pizarras o tabletas (por ejemplo, Apple® iPad, Samsung® Galaxy Tab), teléfonos, teléfonos inteligentes (por ejemplo, Apple® iPhone, dispositivos con Android, BlackBerry®) o asistentes digitales personales. El usuario puede acceder al sistema informático 201 a través de la red 230.

Los métodos descritos en la presente pueden implementarse mediante código ejecutable por máquina (por ejemplo, procesador informático) almacenado en una ubicación de almacenamiento electrónico del sistema informático 201, como, por ejemplo, en la memoria 210 o en la unidad de almacenamiento electrónico 215. El código ejecutable por máquina o legible por máquina puede proporcionarse en forma de software. Durante su uso, el código puede ser ejecutado por el procesador 205. En algunos casos, el código puede recuperarse de la unidad de almacenamiento 215 y almacenarse en la memoria 210 para que el procesador 205 pueda acceder a él. En algunas situaciones, puede excluirse la unidad de almacenamiento electrónico 215, y las instrucciones ejecutables por máquina se almacenan en la memoria 210.

En un aspecto, la presente divulgación proporciona un medio legible por ordenador no transitorio que comprende instrucciones ejecutables por ordenador que, cuando se ejecutan por al menos un procesador electrónico, realizan por lo menos una parte de un método que comprende: recoger ADNcf de un sujeto de prueba; capturar una pluralidad de conjuntos de regiones diana a partir del ADNcf, en donde la pluralidad de conjuntos de regiones diana comprende un conjunto de regiones diana de secuencia variable y un conjunto de regiones diana epigenéticas, por medio de lo cual se produce un conjunto capturado de moléculas de ADNcf; secuenciar las moléculas de ADNcf capturadas, en donde las moléculas de ADNcf capturadas del conjunto de regiones diana de secuencia variable se secuencian a una mayor profundidad de secuenciación que las moléculas de ADNcf capturadas del conjunto de regiones diana epigenéticas; obtener una pluralidad de lecturas de secuencias generadas por un secuenciador de ácidos nucleicos a partir de la secuenciación de las moléculas de ADNcf capturadas; mapear la pluralidad de lecturas de secuencias con una o más secuencias de referencia para generar lecturas de secuencias mapeadas; y procesar las lecturas de secuencias mapeadas correspondientes al conjunto de regiones diana de secuencia variable y al conjunto de regiones diana epigenéticas para determinar la probabilidad de que el sujeto tenga cáncer.

El código puede estar precompilado y configurado para su uso con una máquina que tenga un procesador adaptado para ejecutar el código o puede compilarse durante el tiempo de ejecución. El código puede suministrarse en un lenguaje de programación que puede seleccionarse para permitir que el código se ejecute de manera precompilada o compilada.

Aspectos de los sistemas y métodos proporcionados en la presente, como el sistema informático 201, pueden incorporarse en programación. Varios aspectos de la tecnología pueden considerarse como "productos" o "artículos de fabricación" típicamente en forma de código ejecutable por máquina (o procesador) y/o datos asociados que se transportan o incorporan en un tipo de medio legible por máquina. El código ejecutable por máquina puede almacenarse en una unidad de almacenamiento electrónico, como una memoria (por ejemplo, memoria de sólo lectura, memoria de acceso aleatorio, memoria flash) o un disco duro. Los medios de tipo "almacenamiento" pueden incluir cualquiera o todas las memorias tangibles de los ordenadores, procesadores o similares, o módulos asociados a los mismos, como varias memorias semiconductoras, unidades de cinta, unidades de disco y similares, que pueden proporcionar almacenamiento no transitorio en cualquier momento para la programación del software.

Todo o parte del software puede comunicarse a veces a través de Internet o de varias otras redes de telecomunicaciones. Tales comunicaciones, por ejemplo, pueden permitir la carga del software desde un ordenador o procesador a otro, por ejemplo, desde un servidor de gestión u ordenador anfitrión en la plataforma informática de un servidor de aplicaciones. Así, otro tipo de medios que pueden portar los elementos de software incluyen las ondas ópticas, eléctricas y electromagnéticas, como las que se usan a través de interfaces físicas entre dispositivos locales, a través de redes fijas cableadas y ópticas, y a través de varios enlaces aéreos. Los elementos físicos que transportan dichas ondas, como los enlaces por cable o inalámbricos, los enlaces ópticos o similares, también pueden

considerarse medios portadores del software. Como se usan en la presente, a menos que se limiten a medios de "almacenamiento" tangibles no transitorios, términos como "medio legible" por ordenador o máquina se refieren a cualquier medio que participe en el suministro de instrucciones a un procesador para su ejecución.

Por lo tanto, un medio legible por máquina, como un código ejecutable por ordenador, puede adoptar muchas formas, incluyendo pero no limitadas a, un medio de almacenamiento tangible, un medio de onda portadora o un medio de transmisión física. Los medios de almacenamiento no volátiles incluyen, por ejemplo, discos ópticos o magnéticos, como cualquiera de los dispositivos de almacenamiento de cualquier ordenador u ordenadores o similares, como los que pueden usarse para implementar las bases de datos, etc. que se muestran en los dibujos. Los medios de almacenamiento volátiles incluyen la memoria dinámica, como la memoria principal de dicha plataforma informática. Los medios de transmisión tangibles incluyen cables coaxiales, cables de cobre y fibra óptica, incluyendo los cables que componen un bus dentro de un sistema informático. Los medios de transmisión de ondas portadoras pueden adoptar la forma de señales eléctricas o electromagnéticas, u ondas acústicas o luminosas como las generadas durante las comunicaciones de datos por radiofrecuencia (RF) e infrarrojos (IR). Por lo tanto, las formas comunes de medios legibles por ordenador incluyen, por ejemplo: un disquete, un disco flexible, un disco duro, una cinta magnética, cualquier otro medio magnético, un CD-ROM, DVD o DVD-ROM, cualquier otro medio óptico, tarjetas perforadas, cinta de papel, cualquier otro medio de almacenamiento físico con patrones de agujeros, una RAM, una ROM, una PROM y EPROM, una FLASH EPROM, cualquier otro chip o cartucho de memoria, una onda portadora que transporte datos o instrucciones, cables o enlaces que transporten dicha onda portadora, o cualquier otro medio a partir del cual un ordenador pueda leer código de programación y/o datos. Muchas de estas formas de medios legibles por ordenador pueden participar en el transporte de una o más secuencias de una o más instrucciones a un procesador para su ejecución.

El sistema informático 201 puede incluir o estar en comunicación con una pantalla electrónica que comprende una interfaz de usuario (UI) para proporcionar, por ejemplo, uno o más resultados del análisis de la muestra. Ejemplos de UI incluyen, sin limitación, una interfaz gráfica de usuario (GUI) y una interfaz de usuario basada en web.

Detalles adicionales relativos a sistemas y redes informáticos, bases de datos y productos de programas informáticos también se proporcionan en, por ejemplo, Peterson, Computer Networks: A Systems Approach, Morgan Kaufmann, 5ª ed. (2011), Kurose, Computer Networking: A Top Down Approach, Pearson, 7ª ed. (2016), Elmasri, Fundamentals of Database Systems, Addison Wesley, 6ª ed. (2010), Coronel, Database Systems: Design, Implementation, & Management, Cengage Learning, 11ª ed. (2014), Tucker, Programming Languages, McGraw-Hill Science/Engineering/Math, 2ª ed. (2006), y Rhoton, Cloud Computing Architected: Solution Design Handbook, Recursive Press (2011).

VI. APLICACIONES

1. Cáncer y otras enfermedades

Los presentes métodos pueden usarse para diagnosticar la presencia de afecciones, particularmente cáncer, en un tema, para caracterizar afecciones (por ejemplo, estadificando el cáncer o determinando heterogeneidad de un cáncer), monitorizar la respuesta al tratamiento de una afección, efectuar el riesgo del pronóstico de desarrollar una afección o un curso subsecuente de una afección. La presente divulgación también puede ser útil para determinar la eficacia de una opción de tratamiento particular. Las opciones de tratamiento exitosas pueden aumentar la cantidad de variación en el número de copias o mutaciones raras detectadas en la sangre del sujeto si el tratamiento tiene éxito, ya que más cánceres pueden morir y despojarse del ADN. En otros ejemplos, esto puede no suceder. En otro ejemplo, quizás ciertas opciones de tratamiento puedan correlacionarse con los perfiles genéticos de los cánceres a lo largo del tiempo. Esta correlación puede ser útil para seleccionar una terapia.

Además, si se observa que un cáncer está en remisión después del tratamiento, los presentes métodos pueden usarse para monitorizar la enfermedad residual o la recurrencia de la enfermedad.

En algunas realizaciones, los métodos y sistemas divulgados en la presente pueden usarse para identificar terapias personalizadas o dirigidas para tratar una enfermedad o afección dada en pacientes basándose en la clasificación de una variante de ácido nucleico como de origen somático o de línea germinal. Típicamente, la enfermedad en consideración es un tipo de cáncer. Ejemplos no limitativos de tales cánceres incluyen cáncer del tracto biliar, cáncer de vejiga, carcinoma de células transicionales, carcinoma urotelial, cáncer cerebral, gliomas, astrocitomas, carcinoma de mama, carcinoma metaplásico, cáncer cervical, carcinoma cervical de células escamosas, cáncer rectal, carcinoma colorrectal, cáncer de colon, cáncer colorrectal hereditario no polipósico, adenocarcinomas colorrectales, tumores del estroma gastrointestinal (GIST), carcinoma de endometrio, sarcomas del estroma endometrial, cáncer de esófago, carcinoma esofágico de células escamosas, adenocarcinoma esofágico, melanoma ocular, melanoma uveal, carcinomas de vesícula biliar, adenocarcinoma de vesícula biliar, carcinoma de células renales, carcinoma de células renales claras, carcinoma de células transicionales, carcinomas uroteliales, tumor de Wilms, leucemia, leucemia linfocítica aguda (ALL), leucemia mieloide aguda (AML), leucemia linfocítica crónica (CLL), leucemia mieloide crónica (CML), leucemia mielomonocítica crónica (CMML), cáncer de hígado, carcinoma hepático,

hepatoma, carcinoma hepatocelular, colangiocarcinoma, hepatoblastoma, cáncer de pulmón, cáncer de pulmón de células no pequeñas (NSCLC), mesotelioma, linfomas de células B, linfoma no Hodgkin, linfoma difuso de células B grandes, linfoma de células del manto, linfomas de células T, linfoma no Hodgkin, linfoma/leucemia linfoblástica T precursora, linfomas de células T periféricas, mieloma múltiple, carcinoma nasofaríngeo (NPC), neuroblastoma, cáncer de orofaringe, carcinomas de células escamosas de la cavidad oral, osteosarcoma, carcinoma de ovario, cáncer de páncreas, adenocarcinoma ductal pancreático, neoplasias pseudopapilares, carcinomas de células acinares, cáncer de próstata, adenocarcinoma de próstata, cáncer de piel, melanoma, melanoma maligno, melanoma cutáneo, carcinomas del intestino delgado, cáncer de estómago, carcinoma gástrico, tumor del estroma gastrointestinal (GIST), cáncer de útero o sarcoma uterino. El tipo y/o estadio del cáncer puede detectarse a partir de variaciones genéticas que incluyen mutaciones, mutaciones raras, indeles, variaciones del número de copias, transversiones, translocaciones, inversiones, deleciones, aneuploidías, aneuploidías parciales, poliploidías, inestabilidad cromosómica, alteraciones de la estructura cromosómica, fusiones génicas, fusiones cromosómicas, truncamientos génicos, amplificación génica, duplicaciones génicas, lesiones cromosómicas, lesiones del ADN, cambios anormales en las modificaciones químicas de los ácidos nucleicos, cambios anormales en los patrones epigenéticos y cambios anormales en la 5-metilcitosina de los ácidos nucleicos.

Los datos genéticos también pueden usarse para caracterizar una forma específica de cáncer. Los cánceres a menudo son heterogéneos tanto en su composición como en su estadificación. Los datos del perfil genético pueden permitir la caracterización de subtipos específicos de cáncer que pueden ser importantes en el diagnóstico o tratamiento de ese subtipo específico. Esta información también puede proporcionar a un sujeto o profesional pistas sobre el pronóstico de un tipo específico de cáncer y permitir a un sujeto o profesional adaptar las opciones de tratamiento de acuerdo con el progreso de la enfermedad. Algunos cánceres pueden progresar y volverse más agresivos y genéticamente inestables. Otros cánceres pueden permanecer benignos, inactivos o latentes. El sistema y los métodos de esta divulgación pueden ser útiles para determinar la progresión de la enfermedad.

Además, los métodos de la divulgación pueden usarse para caracterizar la heterogeneidad de una condición anormal en un sujeto. Tales métodos pueden incluir, por ejemplo, la generación de un perfil genético de polinucleótidos extracelulares derivados del sujeto, en donde el perfil genético comprende una pluralidad de datos resultantes de la variación del número de copias y análisis de mutaciones raras. En algunas realizaciones, una condición anormal es cáncer. En algunas realizaciones, la condición anormal puede ser una que dé como resultado una población genómica heterogénea. En el ejemplo del cáncer, se sabe que algunos tumores comprenden células tumorales en diferentes estadios del cáncer. En otros ejemplos, la heterogeneidad puede comprender múltiples focos de enfermedad. De nuevo, en el ejemplo del cáncer, puede haber múltiples focos tumorales, quizás donde uno o más focos son el resultado de metástasis que se han extendido desde un sitio primario.

Los presentes métodos pueden usarse para generar un perfil, huella o conjunto de datos que es una suma de información genética derivada de diferentes células en una enfermedad heterogénea. Este conjunto de datos puede comprender variación del número de copias, variación epigenética y análisis de mutaciones solos o en combinación.

Los presentes métodos pueden usarse para diagnosticar, pronosticar, monitorizar u observar cánceres u otras enfermedades. En algunas realizaciones, los métodos de la presente no implican el diagnóstico, pronóstico o monitorización de un feto y, como tales, no están dirigidos a pruebas prenatales no invasivas. En otras realizaciones, estas metodologías pueden emplearse en un sujeto embarazado para diagnosticar, pronosticar, monitorizar u observar cánceres u otras enfermedades en un sujeto no nacido cuyo ADN y otros polinucleótidos pueden cocircular con moléculas maternas.

Ejemplos no limitativos de otras enfermedades, trastornos o afecciones de base genética que se evalúan opcionalmente usando los métodos y sistemas divulgados en la presente incluyen acondroplasia, deficiencia de alfa-1 antitripsina, síndrome antifosfolípido, autismo, poliquistosis renal autosómica dominante, Charcot-Marie-Tooth (CMT), cri du chat, enfermedad de Crohn, fibrosis quística, enfermedad de Dercum, síndrome de Down, síndrome de Duane, distrofia muscular de Duchenne, trombofilia Factor V Leiden, hipercolesterolemia familiar, fiebre mediterránea familiar, síndrome X frágil, enfermedad de Gaucher, hemocromatosis, hemofilia, holoprosencefalia, enfermedad de Huntington, síndrome de Klinefelter, síndrome de Marfan, distrofia miotónica, neurofibromatosis, síndrome de Noonan, osteogénesis imperfecta, enfermedad de Parkinson, fenilcetonuria, anomalía de Poland, porfiria, progeria, retinitis pigmentosa, inmunodeficiencia combinada grave (SCID), enfermedad de células falciformes, atrofia muscular espinal, Tay-Sachs, talasemia, trimetilaminuria, síndrome de Turner, síndrome velocardiofacial, síndrome WAGR, enfermedad de Wilson, o similares.

En algunas realizaciones, un método descrito en la presente comprende detectar una presencia o ausencia de ADN originado o derivado de una célula tumoral en un punto de tiempo preseleccionado después de un tratamiento previo contra el cáncer de un sujeto previamente diagnosticado con cáncer usando un conjunto de información de secuencia obtenida como se describe en la presente. El método puede comprender además determinar una puntuación de recurrencia del cáncer que es indicativa de la presencia o ausencia del ADN originado o derivado de la célula tumoral para el sujeto de prueba.

Quando se ha determinado una puntuación de recurrencia de cáncer, puede usarse además para determinar un estado de recurrencia de cáncer. El estado de recurrencia del cáncer puede estar en riesgo de recurrencia del cáncer, por ejemplo, cuando la puntuación de recurrencia del cáncer está por encima de un umbral predeterminado. El estado de recurrencia del cáncer puede ser de riesgo bajo o más bajo para la recurrencia del cáncer, por ejemplo, cuando la puntuación de recurrencia del cáncer está por encima de un umbral predeterminado. En realizaciones particulares, una puntuación de recurrencia del cáncer igual al umbral predeterminado puede dar como resultado un estado de recurrencia del cáncer de en riesgo de recurrencia del cáncer o en riesgo bajo o más bajo de recurrencia del cáncer.

En algunas realizaciones, una puntuación de recurrencia de cáncer se compara con un umbral de recurrencia de cáncer predeterminado, y el sujeto de prueba se clasifica como candidato para un tratamiento de cáncer posterior cuando la puntuación de recurrencia de cáncer está por encima del umbral de recurrencia de cáncer o como no candidato para terapia cuando la puntuación de recurrencia de cáncer está por debajo del umbral de recurrencia de cáncer. En realizaciones particulares, una puntuación de recurrencia del cáncer igual al umbral de recurrencia del cáncer puede dar como resultado la clasificación o como candidato para un tratamiento posterior del cáncer o como no candidato para la terapia.

Los métodos analizados anteriormente pueden comprender además cualquier característica o conjunto de características compatibles expuestas en otras partes de la presente, incluyendo en la sección referente a los métodos para determinar un riesgo de recurrencia del cáncer en un sujeto de prueba y/o clasificar a un sujeto de prueba como candidato para un tratamiento posterior del cáncer.

2. Métodos para determinar el riesgo de recurrencia del cáncer en un sujeto de prueba y/o clasificar a un sujeto de prueba como candidato a un tratamiento posterior contra el cáncer.

En algunas realizaciones, un método proporcionado en la presente es un método para determinar un riesgo de recurrencia del cáncer en un sujeto de prueba. En algunas realizaciones, un método proporcionado en la presente es un método para clasificar a un sujeto de prueba como candidato para un tratamiento posterior contra el cáncer.

Cualquiera de dichos métodos puede comprender la recogida de ADN (por ejemplo, originado o derivado de una célula tumoral) del sujeto de prueba diagnosticado con el cáncer en uno o más puntos de tiempo preseleccionados después de uno o más tratamientos previos contra el cáncer al sujeto de prueba. El sujeto puede ser cualquiera de los sujetos descritos en la presente. El ADN puede ser ADNcf. El ADN puede obtenerse de una muestra de tejido.

Cualquiera de tales métodos puede comprender la captura de una pluralidad de conjuntos de regiones diana del ADN del sujeto, en donde la pluralidad de conjuntos de regiones diana comprende un conjunto de regiones diana de secuencia variable y un conjunto de regiones diana epigenéticas, por lo que se produce un conjunto capturado de moléculas de ADN. El paso de captura puede realizarse de acuerdo con cualquiera de las realizaciones descritas en otras partes de la presente.

En cualquiera de tales métodos, el tratamiento previo contra el cáncer puede comprender cirugía, administración de una composición terapéutica y/o quimioterapia.

Cualquiera de tales métodos puede comprender la secuenciación de las moléculas de ADN capturadas, por lo que se produce un conjunto de información de secuencia. Las moléculas de ADN capturadas del conjunto de regiones diana de secuencia variable pueden secuenciarse a una mayor profundidad de secuenciación que las moléculas de ADN capturadas del conjunto de regiones diana epigenéticas.

Cualquiera de tales métodos puede comprender la detección de la presencia o ausencia de ADN originado o derivado de una célula tumoral en un punto temporal preseleccionado usando el conjunto de información de secuencia. La detección de la presencia o ausencia de ADN originado o derivado de una célula tumoral puede realizarse de acuerdo con cualquiera de las realizaciones de lo mismo descritas en la presente.

Los métodos para determinar un riesgo de recurrencia de cáncer en un sujeto de prueba pueden comprender la determinación de una puntuación de recurrencia de cáncer que es indicativa de la presencia o ausencia, o cantidad, del ADN originado o derivado de la célula tumoral para el sujeto de prueba. La puntuación de recurrencia del cáncer puede usarse además para determinar el estado de recurrencia del cáncer. El estado de recurrencia del cáncer puede estar en riesgo de recurrencia del cáncer, por ejemplo, cuando la puntuación de recurrencia del cáncer está por encima de un umbral predeterminado. El estado de recurrencia del cáncer puede ser de riesgo bajo o más bajo de recurrencia del cáncer, por ejemplo, cuando la puntuación de recurrencia del cáncer está por encima de un umbral predeterminado. En realizaciones particulares, una puntuación de recurrencia del cáncer igual al umbral predeterminado puede dar como resultado un estado de recurrencia del cáncer de en riesgo de recurrencia del cáncer o en riesgo bajo o más bajo de recurrencia del cáncer.

Los métodos para clasificar a un sujeto de prueba como candidato para un tratamiento posterior contra el

cáncer pueden comprender la comparación de la puntuación de recurrencia del cáncer del sujeto de prueba con un umbral predeterminado de la recurrencia del cáncer, clasificando de este modo al sujeto de prueba como candidato para el tratamiento posterior contra el cáncer cuando la puntuación de recurrencia del cáncer está por encima del umbral de recurrencia del cáncer o no un candidato para la terapia cuando la puntuación de recurrencia del cáncer está por debajo del umbral de la repetición del cáncer. En realizaciones particulares, una puntuación de recurrencia del cáncer igual al umbral de recurrencia del cáncer puede dar como resultado la clasificación como candidato para un tratamiento posterior contra el cáncer o como no candidato para la terapia. En algunas realizaciones, el tratamiento posterior contra el cáncer comprende quimioterapia o la administración de una composición terapéutica.

Cualquiera de tales métodos puede comprender la determinación de un periodo de supervivencia libre de enfermedad (DFS) para el sujeto de prueba basado en la puntuación de recurrencia del cáncer; por ejemplo, el periodo DFS puede ser de 1 año, 2 años, 3, años, 4 años, 5 años o 10 años.

En algunas realizaciones, el conjunto de información de secuencia comprende secuencias de regiones diana de secuencia variable, y la determinación de la puntuación de recurrencia del cáncer puede comprender la determinación de por lo menos una primera subpuntuación indicativa de la cantidad de SNV, inserciones/delecciones, CNV y/o fusiones presentes en secuencias de regiones diana de secuencia variable.

En algunas realizaciones, un número de mutaciones en las regiones diana de secuencia variable elegidas entre 1, 2, 3, 4, o 5 es suficiente para que la primera subpuntuación dé como resultado una puntuación de recurrencia de cáncer clasificada como positiva para recurrencia de cáncer. En algunas realizaciones, el número de mutaciones se elige entre 1, 2 o 3.

En algunas realizaciones, el conjunto de información de secuencia comprende secuencias de regiones diana epigenéticas, y la determinación de la puntuación de recurrencia de cáncer comprende la determinación de una segunda subpuntuación indicativa de la cantidad de lecturas de secuencia anormales en las secuencias de regiones diana epigenéticas. Las lecturas de secuencias anormales pueden ser lecturas indicativas de un estado epigenético diferente del ADN encontrado en una muestra correspondiente de un sujeto sano (por ejemplo, ADNcf encontrado en una muestra de sangre de un sujeto sano, o ADN encontrado en una muestra de tejido de un sujeto sano donde la muestra de tejido es del mismo tipo de tejido que se obtuvo del sujeto de prueba). Las lecturas anormales pueden ser consistentes con cambios epigenéticos asociados con el cáncer, por ejemplo, metilación de regiones diana variables de hipermetilación y/o fragmentación perturbada de regiones diana variables de fragmentación, donde "perturbada" significa diferente del ADN que se encuentra en una muestra correspondiente de un sujeto sano.

En algunas realizaciones, una proporción de lecturas correspondientes al conjunto de regiones diana variables de hipermetilación y/o conjunto de regiones diana variables de fragmentación que indican hipermetilación en el conjunto de regiones diana variables de hipermetilación y/o fragmentación anormal en el conjunto de regiones diana variables de fragmentación mayor o igual a un valor en el intervalo del 0,001%-10% es suficiente para que la segunda subpuntuación se clasifique como positiva para recurrencia de cáncer. El intervalo puede ser 0,001%-1%, 0,005%-1%, 0,01%-5%, 0,01%-2% o 0,01%-1%.

En algunas realizaciones, cualquiera de tales métodos puede comprender la determinación de una fracción de ADN tumoral a partir de la fracción de lecturas en el conjunto de información de secuencia que indican una o más características indicativas de origen de una célula tumoral. Esto puede hacerse para lecturas correspondientes a algunas o todas las regiones diana epigenéticas, por ejemplo, incluyendo una o ambas regiones diana variables de hipermetilación y regiones diana variables de fragmentación (la hipermetilación de una región diana variable de hipermetilación y/o la fragmentación anormal de una región diana variable de fragmentación pueden considerarse indicativas del origen de una célula tumoral). Esto puede hacerse para lecturas correspondientes a regiones diana variables de secuencia, por ejemplo, lecturas que comprenden alteraciones consistentes con el cáncer, como SNV, índices, CNV y/o fusiones. La fracción de ADN tumoral puede determinarse basándose en una combinación de lecturas correspondientes a regiones diana epigenéticas y lecturas correspondientes a regiones diana de secuencia variable.

La determinación de una puntuación de recurrencia de cáncer puede basarse por lo menos en parte en la fracción de ADN tumoral, en donde una fracción de ADN tumoral mayor que un umbral en el intervalo de 10^{-11} a 1 o 10^{-10} a 1 es suficiente para que la puntuación de recurrencia de cáncer se clasifique como positiva para recurrencia de cáncer. En algunas realizaciones, una fracción de ADN tumoral mayor o igual que un umbral en el intervalo de 10^{-10} a 10^{-9} , 10^{-9} a 10^{-8} , 10^{-8} a 10^{-7} , 10^{-7} a 10^{-6} , 10^{-6} a 10^{-5} , 10^{-5} a 10^{-4} , 10^{-4} a 10^{-3} , 10^{-3} a 10^{-2} , o 10^{-2} a 10^{-1} es suficiente para que la puntuación de recurrencia del cáncer se clasifique como positiva para la recurrencia del cáncer. En algunas realizaciones, la fracción de ADN tumoral mayor que un umbral de por lo menos 10^{-7} es suficiente para que la puntuación de recurrencia del cáncer se clasifique como positiva para la recurrencia del cáncer. Una determinación de que una fracción de ADN tumoral es mayor que un umbral, como un umbral correspondiente a cualquiera de las realizaciones anteriores, puede realizarse basándose en una probabilidad acumulativa. Por ejemplo, la muestra se considera positiva si la probabilidad acumulativa de que la fracción tumoral sea mayor que un umbral en cualquiera de los intervalos anteriores supera un umbral de probabilidad de por lo menos 0,5, 0,75, 0,9, 0,95, 0,98, 0,99, 0,995 o 0,999. En algunas realizaciones, el umbral de probabilidad es de por lo menos 0,95, por ejemplo 0,99.

En algunas realizaciones, el conjunto de información de secuencia comprende secuencias de regiones diana de secuencia variable y secuencias de regiones diana epigenéticas, y la determinación de la puntuación de recurrencia de cáncer comprende determinar una primera subpuntuación indicativa de la cantidad de SNV, inserciones/deleciones, CNV y/o fusiones presentes en secuencias de regiones diana de secuencia variable y una segunda subpuntuación indicativa de la cantidad de lecturas de secuencia anormales en secuencias de regiones diana epigenéticas, y la combinación de la primera y la segunda subpuntuaciones para proporcionar la puntuación de recurrencia de cáncer. Cuando se combinan la primera y la segunda subpuntuaciones, pueden combinarse aplicando un umbral a cada subpuntuación de manera independiente (por ejemplo, mayor que un número predeterminado de mutaciones (por ejemplo, > 1) en regiones diana de secuencia variable, y mayor que una fracción predeterminada de lecturas anormales (por ejemplo, tumorales) en regiones diana epigenéticas), o entrenando un clasificador de aprendizaje automático para determinar el estado basándose en una pluralidad de muestras de entrenamiento positivas y negativas.

En algunas realizaciones, un valor para la puntuación combinada en el intervalo de -4 a 2 o de -3 a 1 es suficiente para que la puntuación de recurrencia del cáncer se clasifique como positiva para recurrencia del cáncer.

En cualquier realización donde una puntuación de recurrencia de cáncer se clasifica como positiva para recurrencia de cáncer, el estado de recurrencia de cáncer del sujeto puede estar en riesgo de recurrencia de cáncer y/o el sujeto puede ser clasificado como candidato para un tratamiento contra el cáncer posterior.

En algunas realizaciones, el cáncer es cualquiera de los tipos de cáncer descritos en otras partes de la presente, por ejemplo, cáncer colorrectal.

3. Terapias y administración relacionada

En ciertas realizaciones, los métodos divulgados en la presente se refieren a la identificación de terapias personalizadas para pacientes dado el estado de una variante de ácido nucleico como de origen somático o de línea germinal. Esencialmente, puede incluirse cualquier terapia contra el cáncer (por ejemplo, terapia quirúrgica, radioterapia, quimioterapia y/o similares). Típicamente, las terapias personalizadas incluyen por lo menos una inmunoterapia (o un agente inmunoterapéutico). La inmunoterapia se refiere de manera general a métodos para potenciar una respuesta inmunitaria contra un tipo de cáncer dado, incluyendo métodos para potenciar una respuesta de células T contra un tumor o cáncer.

En ciertas realizaciones, el estado de una variante de ácido nucleico de una muestra de un sujeto como de origen somático o de línea germinal puede compararse con una base de datos de resultados de comparadores de una población de referencia para identificar terapias personalizadas o dirigidas para ese sujeto. Típicamente, la población de referencia incluye pacientes con el mismo tipo de cáncer o enfermedad que el sujeto de la prueba y/o pacientes que están recibiendo, o que han recibido, la misma terapia que el sujeto de la prueba. Una terapia (o terapias) personalizada o dirigida puede identificarse cuando la variante nucleica y los resultados del comparador satisfacen ciertos criterios de clasificación (por ejemplo, son una coincidencia sustancial o aproximada).

La administración de la terapia se divulga pero no forma parte de la invención. Las terapias personalizadas descritas en la presente típicamente se administran por vía parenteral (por ejemplo, por vía intravenosa o subcutánea). Las composiciones farmacéuticas que contienen un agente inmunoterapéutico se administran típicamente por vía intravenosa. Ciertos agentes terapéuticos se administran por vía oral. Sin embargo, las terapias personalizadas (por ejemplo, agentes inmunoterapéuticos, etc.) también pueden administrarse por métodos como, por ejemplo, bucal, sublingual, rectal, vaginal, intrauretral, tópico, intraocular, intranasal y/o intraauricular, cuya administración puede incluir comprimidos, cápsulas, gránulos, suspensiones acuosas, geles, aerosoles, supositorios, bálsamos, pomadas o similares.

Aunque en la presente se han mostrado y descrito realizaciones preferidas de la presente invención, para los expertos en la técnica será obvio que dichas realizaciones se proporcionan únicamente a modo de ejemplo. No se pretende que la invención esté limitada por los ejemplos específicos proporcionados en la memoria descriptiva. Aunque la invención se ha descrito con referencia a la memoria descriptiva mencionada anteriormente, las descripciones e ilustraciones de las realizaciones de la presente no deben interpretarse en un sentido limitativo. A los expertos en la técnica se les ocurrirán numerosas variaciones, cambios y sustituciones sin apartarse de la invención tal como se define en las reivindicaciones adjuntas.

Aunque la divulgación anterior se ha descrito con cierto detalle a modo de ilustración y ejemplo con propósitos de claridad y comprensión, para un experto en la técnica a partir de la lectura de esta divulgación estará claro que pueden realizarse varios cambios en la forma y el detalle y pueden ponerse en práctica dentro del alcance de las reivindicaciones adjuntas.

Si diferentes versiones de una secuencia están asociadas con un número de registro en diferentes momentos,

se entenderá la versión asociada con el número de registro en la fecha de presentación efectiva de esta solicitud. Fecha de presentación efectiva significa la anterior de la fecha de presentación real o la fecha de presentación de una solicitud prioritaria que haga referencia al número de registro, si procede. De igual manera, si se publican diferentes versiones de una publicación, sitio web o similar en diferentes momentos, se entenderá la versión publicada más recientemente en la fecha de presentación efectiva de la solicitud, a menos que se indique lo contrario.

VII. EJEMPLOS

i) Caracterización de conjuntos de sondas de regiones diana con diferentes concentraciones de sondas para un conjunto de regiones diana de secuencia variable y sondas para un conjunto de regiones diana epigenéticas

Este ejemplo describe una evaluación del rendimiento de conjuntos de sondas que contienen sondas para un conjunto de regiones diana de secuencia variable y sondas para un conjunto de regiones diana epigenéticas como parte de un esfuerzo por combinar el análisis epigenético y genotípico del ADNcf de biopsia líquida.

Las muestras de ADNcf se procesaron antes de ponerlas en contacto con un conjunto de sondas de la región diana realizando la división basándose en el estado de metilación, la reparación de extremos, la ligadura con adaptadores y la amplificación por PCR (por ejemplo, usando cebadores dirigidos a los adaptadores).

Las muestras procesadas se pusieron en contacto con conjuntos de sondas de regiones diana que comprendían sondas para un conjunto de regiones diana de secuencia variable y sondas para un conjunto de regiones diana epigenéticas. Las sondas de regiones diana estaban en forma de oligonucleótidos biotinilados diseñados para cubrir con segmentos las regiones de interés. Las sondas para el conjunto de regiones diana de secuencia variable tenían una huella de aproximadamente 50 kb y las sondas para el conjunto de regiones diana epigenéticas tenían una huella de aproximadamente 500 kb. Las sondas para el conjunto de regiones diana de secuencia variable comprendían oligonucleótidos dirigidos a una selección de regiones identificadas en las Tablas 3-5 y las sondas para el conjunto de regiones diana epigenéticas comprendían oligonucleótidos dirigidos a una selección de regiones diana de hipometilación variable, regiones diana de hipometilación variable, regiones diana de unión a CTCF, regiones diana de sitios de inicio de transcripción, regiones diana de amplificación focal y regiones de control de metilación.

El ADNcf capturado y aislado de esta manera se preparó después para la secuenciación y se secuenció usando un secuenciador Illumina HiSeq o NovaSeq. Los resultados se analizaron con respecto a la diversidad (número de familias únicas de lecturas de secuencia) y el tamaño de la familia de lecturas (número de lecturas individuales en cada familia) de las lecturas de secuencia correspondientes a las sondas para el conjunto de regiones diana de secuencia variable y las sondas para el conjunto de regiones diana epigenéticas. Los valores que se indican a continuación se obtuvieron usando 70 ng de ADN de entrada. 70 ng de entrada se considera una cantidad relativamente alta y representa una condición desafiante para mantener los niveles deseados de diversidad y tamaño de familia.

Las proporciones de sonda de 2:1 y 5:1 (proporciones de concentración de masa por volumen de los conjuntos de sonda epigenética: de secuencia variable) dieron reducciones en la diversidad para las regiones diana de secuencia variable, indicando que la cantidad de regiones diana epigenéticas dio como resultado la interferencia con la generación del número esperado de familias de lectura distintas de las regiones diana de secuencia variable.

Las proporciones de sonda de 1:2 o 1:5 (conjuntos de sonda epigenética: de secuencia variable) dieron niveles más altos de diversidad para las regiones diana de secuencia variable, que en general estaban cerca del número esperado de familias de lectura distintas, lo que indica que en estas proporciones, la presencia de las regiones diana epigenéticas no estaban presentes en cantidades que interfirieran sustancialmente con la generación del número esperado de familias de lectura distintas de las regiones diana de secuencia variable.

Para las regiones diana epigenéticas, todas las proporciones dieron niveles de diversidad sustancialmente más bajos que el número esperado de familias de lecturas distintas. Sin embargo, esto no se considera problemático, dado que el análisis de metilación, número de copias y similares para las regiones diana epigenéticas no requiere una cobertura de secuenciación densa y profunda en la misma medida que la determinación de la presencia o ausencia de sustituciones de nucleótidos o indeles como se pretende para las regiones de secuencia variable.

ii) Detección del cáncer usando conjuntos combinados de regiones diana epigenéticas y de secuencia variable

Se analizan cohortes de muestras de ADNcf de pacientes con cáncer con diferentes estadios de cáncer de I a IVA (7 estadios en total) como se ha descrito anteriormente usando sondas en la proporción 1:5 (conjuntos de sondas epigenéticas: de secuencia variable). Las secuencias de regiones dianas de secuencia variable se analizan mediante la detección de alteraciones genómicas como SNV, inserciones, deleciones y fusiones que pueden llamarse con suficiente apoyo para discriminar las variantes tumorales reales de los errores técnicos. Las secuencias epigenéticas de la región diana se analizan de manera independiente para detectar fragmentos metilados en regiones que han

demostrado estar metiladas diferencialmente en el cáncer en comparación con las células sanguíneas. Por último, los resultados de ambos análisis se combinan para producir una llamada final de tumor presente/ausente para determinar si mostraron un perfil consistente con cáncer con una especificidad del 95%.

La FIG. 3 muestra la sensibilidad de la detección de cáncer basada en secuencias de regiones diana de secuencia variable y las secuencias de regiones diana epigenéticas solas y en combinación. La detección de cáncer fue un 100% sensible para cualquiera de los enfoques solos para las cohortes de estadio IIIA y IIIC. Para todas las demás cohortes, excepto una, la inclusión del análisis de las secuencias epigenéticas de la región diana aumentó la sensibilidad en aproximadamente un 10-30%. La única excepción fue la cohorte de estadio IIB, en la que cada muestra fue o bien un verdadero positivo de acuerdo con ambos enfoques o un falso negativo de acuerdo con ambos enfoques.

Por tanto, los métodos y composiciones divulgados pueden proporcionar ADNcf capturado utilizable para secuenciar concurrentemente regiones diana epigenéticas y de secuencia variable a diferentes profundidades de secuenciación en la detección sensible, combinada basada en secuencia y epigenética del cáncer.

iii) Identificación del nivel de riesgo de recurrencia del cáncer colorrectal

Se desarrolló y realizó un ensayo para identificar si los pacientes tratados por cáncer colorrectal (CCR) tenían un alto riesgo de recurrencia. Se tomaron muestras de plasma (de 3 a 4 ml) de 72 pacientes que se estaban sometiendo al tratamiento estándar para el CCR (cirugía +/- terapia neoadyuvante en 42 casos y terapia adyuvante +/- terapia neoadyuvante en 30 casos).

Se extrajo ADNcf de las muestras (cantidad mediana 27 ng) y se analizó usando un método esencialmente como se describe en la presente que se validó en CCR en estadio temprano y que integra evaluaciones de alteraciones genómicas y características epigenómicas indicativas de cáncer, incluyendo regiones diana variables de hipermetilación. El método diferencia las alteraciones derivadas de tumores de las no derivadas de tumores (por ejemplo, alteraciones de línea germinal o de hematopoyesis clonal de potencial indeterminado (CHIP)) en un enfoque no informado de tejido tumoral (ensayo LUNAR, Guardant Health, CA). El ensayo usa una única muestra de entrada e integra la detección de alteraciones genómicas con la cuantificación de señales epigenómicas asociadas al cáncer, y se validó usando 80 muestras de plasma de donantes de 50-75 años presuntamente libres de cáncer, y dio como resultado un único falso positivo (especificidad del 99%). La sensibilidad analítica (límite de detección) se estableció usando una serie de diluciones de 4 diferentes pts de CCR en estado tardío probados por triplicado a una entrada de ADN clínicamente relevante (30 ng) a través de múltiples lotes. Se mantuvo una sensibilidad del 100% incluso en el nivel más bajo probado (nivel tumoral estimado del 0,1%).

Las muestras de plasma posteriores a la finalización de la terapia SOC se recogieron una mediana de 31 días después de la resección quirúrgica (N = 42) o una mediana de 37 días después de la finalización de la terapia adyuvante (N = 27). La mediana de seguimiento fue de 515 días (33 - 938 días). Las muestras se consideraron positivas para ADNct si se detectaban alteraciones genómicas o epigenómicas indicativas de cáncer. Las alteraciones genómicas se detectaron usando la plataforma de Secuenciación Digital de Guardant Health para diferenciar las mutaciones reales de los errores de secuenciación. Se aplicó un filtro de variantes para diferenciar las mutaciones tumorales de las no tumorales (como CHIP). Las llamadas epigenómicas se basaron en medir si la tasa de metilación observada en las regiones hipermetiladas tumorales era superior a la esperada sobre la base de los niveles de metilación en sangre. Específicamente, en esta realización, los resultados genómicos se consideraron positivos si el número de alteraciones genómicas indicativas de cáncer detectadas superaba un umbral, donde el umbral era 1, 2 o 3 alteraciones. Los resultados epigenómicos incluyeron el análisis de metilación para determinar la proporción de lecturas que indican hipermetilación en el conjunto de regiones diana variables de hipermetilación. También se calculó una "fracción tumoral" global sobre la base de la proporción global de lecturas que presentaban características similares a las tumorales basadas en la metilación, y la muestra se consideró positiva si la probabilidad acumulada de que la fracción tumoral fuera mayor o igual que 10^{-7} superaba un umbral de probabilidad de 0,99. 14 muestras en total fueron positivas, de las cuales 10 lo fueron tanto para la vertiente epigenómica como para la genómica, 3 lo fueron sólo para la vertiente epigenómica y 1 lo fue sólo para la vertiente genómica.

7/111 pacientes con recurrencia la de 1 año de la cirugía fueron positivos para el ADNct detectado tras la resección del CCR. 30/31 pacientes sin recurrencia al de 1 año de la intervención quirúrgica fueron negativos para ADNct tras la resección del CCR. 20/22 pacientes sin recurrencia al de 1 año de la terapia adyuvante fueron negativos para ADNct tras la finalización de la terapia adyuvante SOC. 4/5 pacientes con recurrencia al de 1 año de la terapia adyuvante fueron positivos para ADNct tras la finalización de la terapia adyuvante SOC. En general, la detección de ADNct después de la finalización de la terapia estándar tuvo un valor predictivo positivo (PPV) de recurrencia del 100%, un valor predictivo negativo (NPV) del 76% y un cociente de riesgo de recurrencia de 9,22 ($p < 0,0001$) (FIG. 4).

En la siguiente tabla se resumen las estadísticas de rendimiento del ensayo para la rama genómica únicamente y para el análisis integrado usando las ramas genómica y epigenómica.

Tabla 6

Rendimiento del ensayo	Genómico (N)	Genómico + Epigenómico (N)
Valor predictivo positivo (N de los pacientes con ADNct detectado que recayeron)	10% (11/11)	100% (14/14)
Valor predictivo negativo (N de los pacientes con ADNct no detectado que se encontraban libres de recurrencia)	72% (42/58)	76% (42/55)
Sensibilidad para la recurrencia en el plazo de un año de la cirugía	56% (9/16)	69% (11/16)
Especificidad para estar libre de recurrencia en el plazo de un año de la cirugía	96% (51/53)	94% (50/53)

Resultados de la cohorte por secuenciación genómica frente a análisis epigenómico. De los 14 pacientes que dieron positivo en ADNct después de la finalización de la terapia con SOC, 10 dieron positivo tanto por evaluación genómica como epigenómica.

En la cohorte de cirugía, la detección de ADNct tuvo un valor predictivo positivo (PPV) de recurrencia del 100%, un valor predictivo negativo (NPV) del 76% y un cociente de riesgo de recurrencia de 8,7 ($p < 0,0001$). En la cohorte de terapia adyuvante, la detección de ADNct tuvo un PPV de recurrencia del 100%, un NPV del 76% y un cociente de riesgo de recurrencia de 9,3 ($p < 0,0001$).

Los pacientes negativos para ADNct después de la finalización de la terapia se clasificaron además en función de si eran positivos o negativos para ADNct antes de la terapia. Los pacientes que eran positivos antes y negativos después de la terapia se denominaron "depurados", mientras que los pacientes que eran negativos antes y después se denominaron "negativos". La población depurada contenía 6 individuos, de los cuales 3 tuvieron recurrencia y 3 no. La población negativa contenía 26 individuos, de los cuales 7 tuvieron recurrencia y 19 no.

Por tanto, en el CRC reseado, la detección de ADNct utilizando un ensayo genómico y epigenómico integrado, sin información sobre el tumor y sólo en plasma, tiene un PPV y un NPV de recurrencia altos después de la finalización de la terapia de tratamiento estándar. En el contexto posterior a la resección, la detección de ADNct identifica a los pacientes que pueden beneficiarse de la terapia adyuvante. Una vez finalizada la terapia adyuvante, la detección de ADNct identifica a los pacientes que pueden beneficiarse de una terapia adicional o modificada. Estos descubrimientos demuestran que el ADNct a partir de una única extracción de sangre después de la resección o después de la terapia adyuvante puede identificar a pacientes de alto riesgo e informar de la toma de decisiones terapéuticas. Por el contrario, los enfoques actuales de detección de la enfermedad residual por ADNct sólo evalúan las alteraciones genómicas, están limitados por los bajos niveles de ADNct y dependen de la secuenciación del tejido tumoral para diferenciar las alteraciones derivadas del tumor de las alteraciones confusas no derivadas del tumor (por ejemplo, hematopoyesis clonal de potencial indeterminado; CHIP).

REIVINDICACIONES

1. Un método para determinar la probabilidad de que un sujeto tenga cáncer, que comprende:

- 5 a. dividir el ADNcf obtenido de un sujeto de prueba en dos o más fracciones sobre la base del nivel de metilación usando un dominio de unión a metilo y, posteriormente, en cada fracción;
b. capturar una pluralidad de conjuntos de regiones diana a partir del ADNcf, en donde:

10 I. la pluralidad de conjuntos de regiones diana comprende un conjunto de regiones diana de secuencia variable y un conjunto de regiones diana epigenéticas, mediante lo cual se produce un conjunto capturado de moléculas de ADNcf;

II. las moléculas de ADNcf correspondientes al conjunto de regiones diana de secuencia variable se capturan con un mayor rendimiento de captura en el conjunto capturado de moléculas de ADNcf que las moléculas de ADNcf correspondientes al conjunto de regiones diana epigenéticas; y

15 III. el conjunto de regiones diana epigenéticas se selecciona entre una o más de

i) un conjunto de regiones diana variables de hipermetilación;

ii) un conjunto de regiones diana variables de hipometilación; y/o

20 iii) un conjunto de regiones diana variables de fragmentación, en donde el conjunto de regiones diana variables de fragmentación comprende regiones de sitios de inicio de transcripción y/o regiones de unión a CTCF;

opcionalmente, en donde el conjunto de regiones diana epigenéticas comprende además un conjunto de regiones diana de control de metilación;

25 c. secuenciar las moléculas de ADNcf capturadas, en donde las moléculas de ADNcf capturadas del conjunto de regiones diana de secuencia variable se secuencian con una mayor profundidad de secuenciación que las moléculas de ADNcf capturadas del conjunto de regiones diana epigenéticas;

30 d. obtener una pluralidad de lecturas de secuencia generadas por un secuenciador de ácidos nucleicos a partir de la secuenciación de las moléculas de ADNcf capturadas;

e. mapear la pluralidad de lecturas de secuencia para una o más secuencias de referencia para generar lecturas de secuencia mapeadas;

35 f. procesar las lecturas de secuencias mapeadas correspondientes al conjunto de regiones diana de secuencia variable y al conjunto de regiones diana epigenéticas para determinar la probabilidad de que el sujeto tenga cáncer.

2. El método de cualquier reivindicación anterior, en donde el ADNcf se amplifica antes y/o después del paso de captura, opcionalmente en donde la amplificación del ADNcf comprende los pasos de ligar adaptadores que contienen códigos de barras al ADNcf.

40 3. El método de la reivindicación 1 o de la reivindicación 2, en donde las dos o más fracciones

i) se etiquetan diferencialmente y se agrupan juntas para la preparación y/o secuenciación colectiva de muestras; y/o

45 ii) comprenden una división hipermetilada y otra hipometilada.

4. El método de cualquiera de las reivindicaciones anteriores, en donde la captura de la pluralidad de conjuntos de regiones diana de ADNcf comprende poner en contacto el ADNcf con sondas de unión a diana específicas para el conjunto de regiones diana de secuencia variable y sondas de unión a diana específicas para el conjunto de regiones diana epigenéticas, por ejemplo, en donde las sondas de unión a diana específicas para el conjunto de regiones diana de secuencia variable

50 i) están presentes en una concentración más alta que la de las sondas de unión a la diana específicas para el conjunto de regiones diana epigenéticas, por ejemplo (a) una concentración por lo menos 2 veces más alta, o (b) una concentración por lo menos 4 o 5 veces más alta; y/o

55 ii) tienen una afinidad de unión a la diana mayor que la de las sondas de unión a la diana específicas para el conjunto de regiones diana epigenéticas.

60 5. El método de cualquiera de las reivindicaciones anteriores, en donde las moléculas de ADNcf capturadas del conjunto de regiones diana de secuencia variable se agrupan con las moléculas de ADNcf capturadas del conjunto de regiones diana epigenéticas antes de la secuenciación y/o las moléculas de ADNcf capturadas del conjunto de regiones diana de secuencia variable y las moléculas de ADNcf capturadas del conjunto de regiones diana epigenéticas se secuencian en la misma celda de secuenciación.

65 6. El método de cualquiera de las reivindicaciones anteriores, en donde

- i) el conjunto de regiones diana de secuencia variable comprende regiones diana de por lo menos 10, 20, 30 o 35 genes relacionados con el cáncer; y/o
- ii) la huella del conjunto de regiones diana de secuencia variable tiene por lo menos 25 kB o 50 kB.

7. El método de cualquiera de las reivindicaciones anteriores, en donde la huella del conjunto de regiones diana epigenéticas es

- i) por lo menos 2 veces mayor que el tamaño del conjunto de regiones objetivo de secuencia variable; o
- ii) por lo menos 10 veces mayor que el tamaño del conjunto de regiones objetivo de secuencia variable.

8. El método de cualquiera de las reivindicaciones anteriores, en donde las moléculas de ADNcf capturadas del conjunto de regiones diana de secuencia variable se secuencian para

- i) una profundidad de secuenciación por lo menos 2 veces mayor;
- ii) una profundidad de secuenciación por lo menos 3 veces mayor;
- iii) una profundidad de secuenciación por lo menos 4-10 veces mayor; o
- iv) una profundidad de secuenciación por lo menos 4-100 veces mayor

que la de las moléculas de ADNcf capturadas del conjunto de regiones diana epigenéticas.

9. El método de cualquiera de las reivindicaciones anteriores, en donde hay por lo menos 10 regiones en el conjunto de regiones diana de secuencia variable y por lo menos 100 regiones en el conjunto de regiones diana epigenéticas.

10. El método de cualquiera de las reivindicaciones anteriores, que comprende además determinar si

- i) las moléculas de ADNcf correspondientes al conjunto de regiones diana de secuencia variable comprenden mutaciones asociadas al cáncer; y/o
- ii) las moléculas de ADNcf correspondientes al conjunto de regiones diana epigenéticas comprenden o indican modificaciones epigenéticas asociadas al cáncer, por ejemplo en donde las modificaciones epigenéticas asociadas al cáncer comprenden hipermetilación en una o más regiones diana variables de hipermetilación, una o más perturbaciones de la unión a CTCF, y/o una o más perturbaciones de los sitios de inicio de la transcripción.

11. El método de cualquiera de las reivindicaciones anteriores, en donde el dominio de unión a metilo está inmovilizado sobre un soporte sólido.

12. Un sistema, que comprende:

una interfaz de comunicaciones que recibe, a través de una red de comunicaciones, una pluralidad de lecturas de secuencias generadas por un secuenciador de ácidos nucleicos a partir de la secuenciación de un conjunto capturado de moléculas de ADNcf, en donde el conjunto capturado de moléculas de ADNcf se obtiene de acuerdo con los pasos (a) y (b) de la reivindicación 1, en donde las moléculas de ADNcf capturadas correspondientes a la región diana de secuencia variable se secuencian a una mayor profundidad de secuenciación que las moléculas de ADNcf capturadas correspondientes al conjunto de regiones diana epigenéticas; y

un controlador que comprende, o es capaz de acceder a, medios legibles por ordenador que comprenden instrucciones ejecutables por ordenador no transitorias que, cuando son ejecutadas por al menos un procesador electrónico realizan un método que comprende:

- (i) recibir, a través de la red de comunicaciones, las lecturas de secuencias generadas por el secuenciador de ácidos nucleicos;
- (ii) mapear la pluralidad de lecturas de secuencia para una o más secuencias de referencia para generar lecturas de secuencia mapeadas;
- (iii) procesar las lecturas de secuencia mapeadas correspondientes al conjunto de regiones diana de secuencia variable y al conjunto de regiones diana epigenéticas para determinar la probabilidad de que el sujeto tenga cáncer.

13. El método de cualquiera de las reivindicaciones 1-11 o el sistema de la reivindicación 12, en donde al sujeto de prueba se le ha diagnosticado con anterioridad un cáncer, opcionalmente en donde el sujeto de prueba recibió uno o más tratamientos previos contra el cáncer, además opcionalmente en donde el ADNcf se obtiene en uno o más puntos temporales preseleccionados después del uno o más tratamientos previos contra el cáncer.

14. El método de cualquiera de las reivindicaciones 1-11, y 13, y/o el sistema de la reivindicación 12 o de la reivindicación 13, en donde los resultados del método y/o del sistema se usan como entrada para generar un informe, opcionalmente en donde:

- (i) el informe está en papel o en formato electrónico;

(ii) en el informe se muestra la información sobre, y/o información derivada de, la información de la secuencia determinada por el método y/o sistema, como en donde esta información es el estado del cáncer del sujeto, según se determina por el método y/o sistema; y/o

5 (iii) el método y/o sistema comprende además un paso de comunicar el informe a un tercero, como el sujeto del que se derivó la muestra o un profesional sanitario.

10

15

20

25

30

35

40

45

50

55

60

65

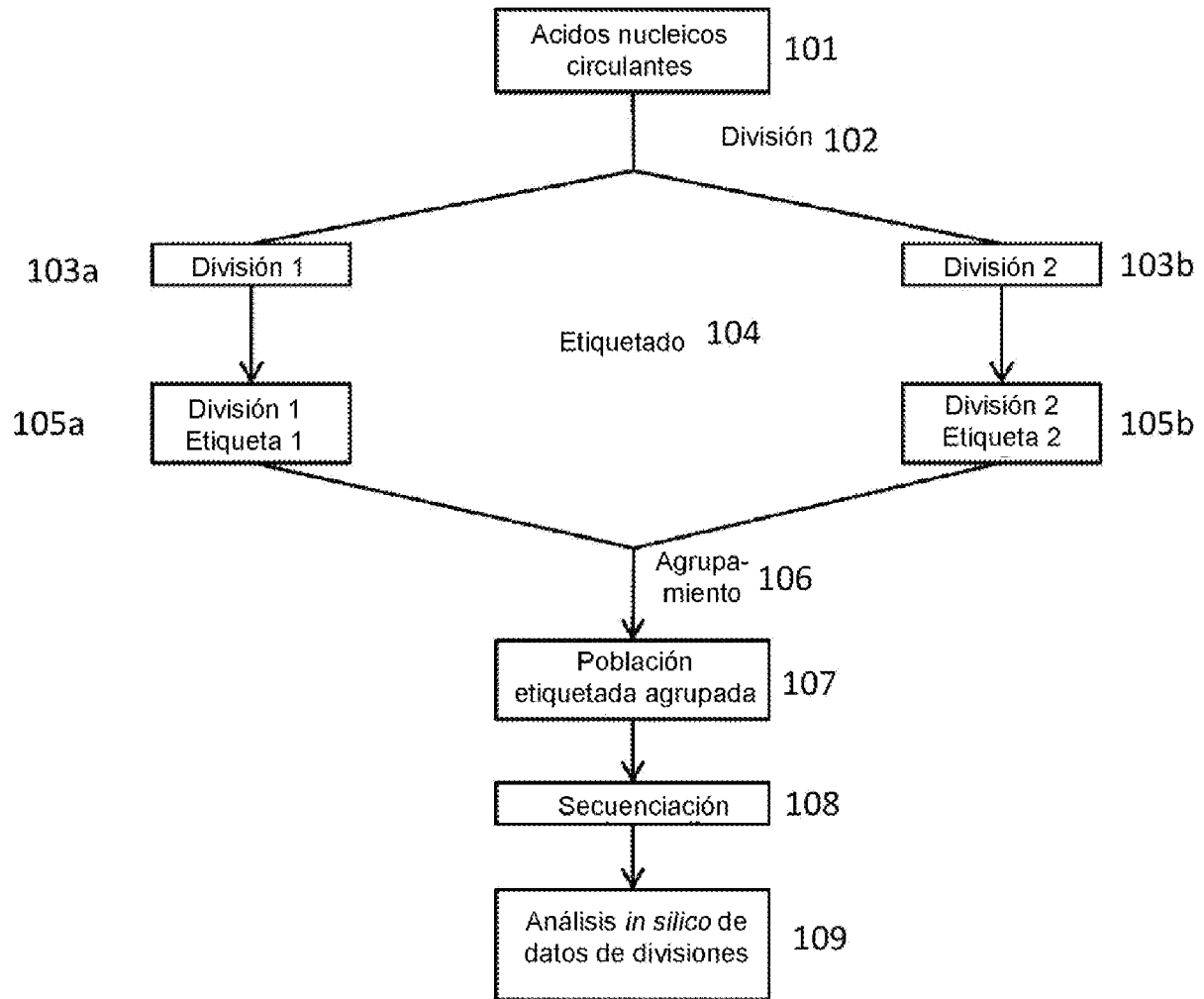


Fig. 1

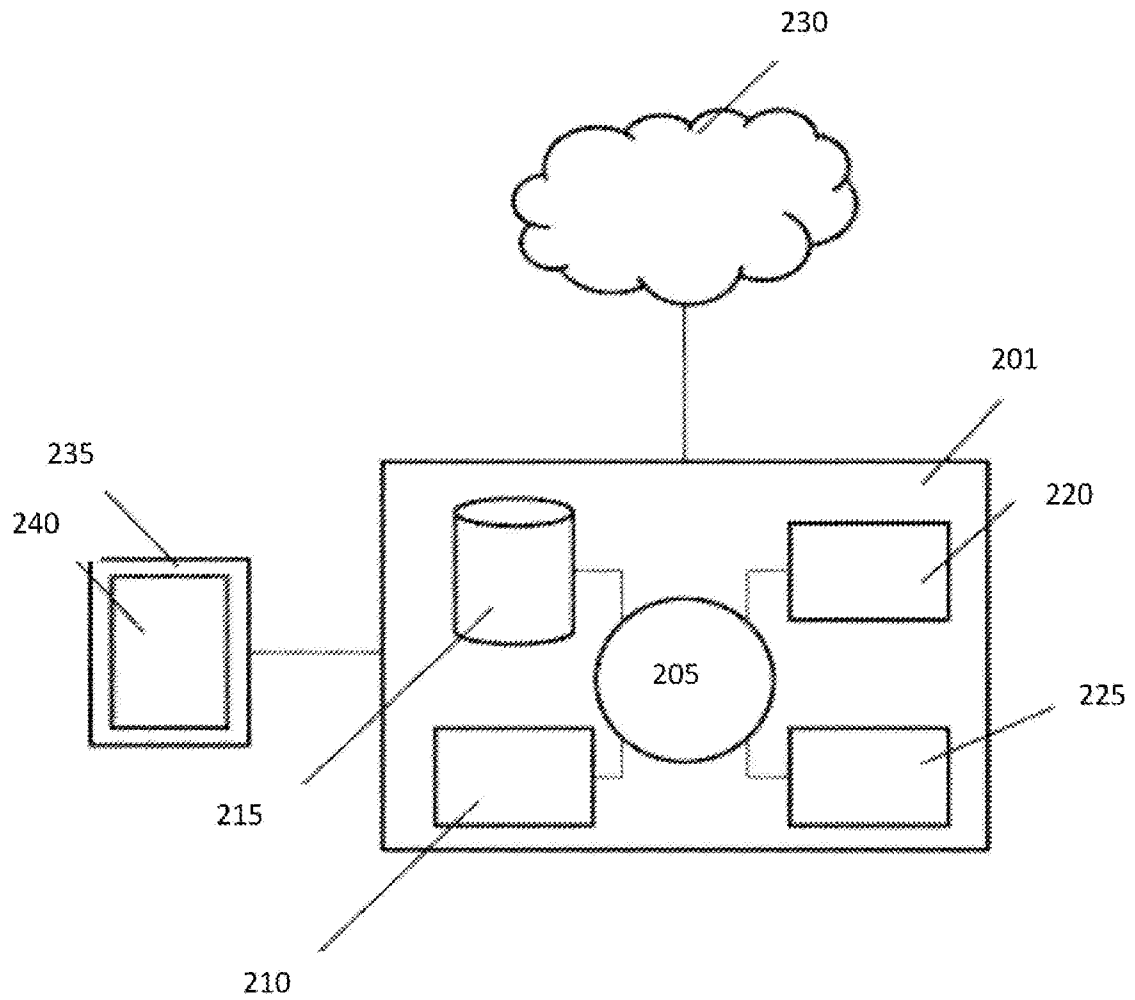


Fig. 2

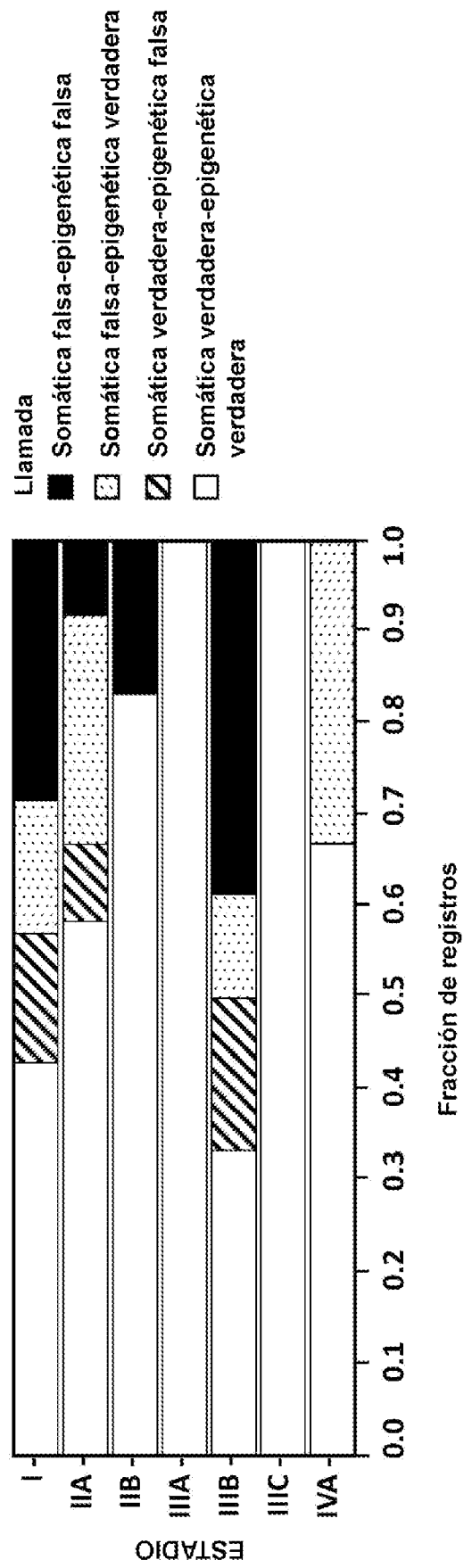


Fig. 3

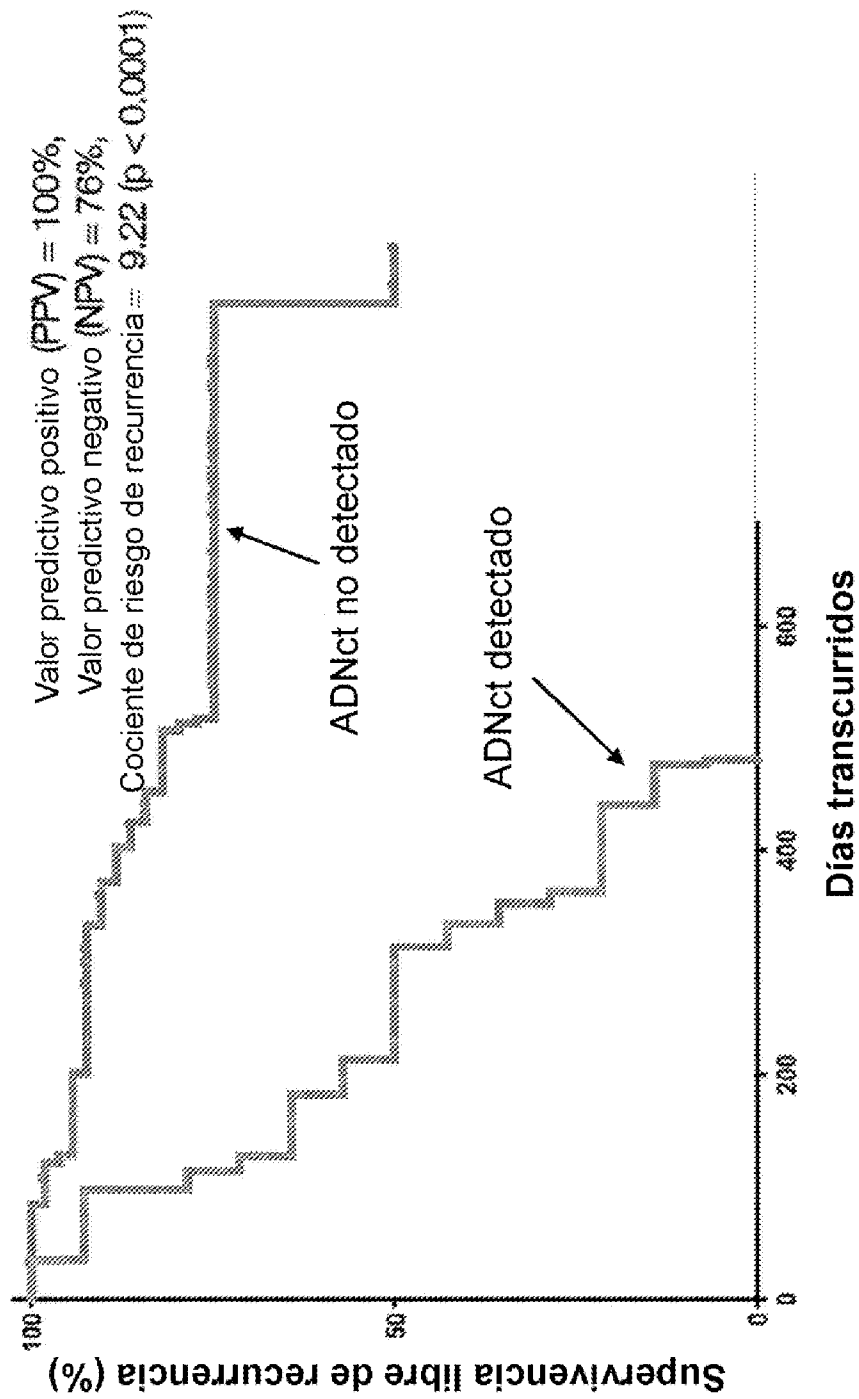


Fig. 4