(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 908 347**

(51) Int. CI.:

G16B 20/00 (2009.01)
G16B 30/00 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **14.02.2016 PCT/CN2016/073753**

(87) Fecha y número de publicación internacional: **18.08.2016 WO16127944**

(96) Fecha de presentación y número de la solicitud europea: **14.02.2016 E 16748745 (3)**

(97) Fecha y número de publicación de la concesión europea: **09.02.2022 EP 3256605**

(54) Título: **Detección de mutaciones para cribado de cáncer y análisis fetal**

(30) Prioridad:

10.02.2015 US 201562114471 P
22.12.2015 US 201562271196 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
28.04.2022

(73) Titular/es:

THE CHINESE UNIVERSITY OF HONG KONG (100.0%)
Office of Research and Knowledge Transfer Services, (ORKTS), Room 301Pi Ch'iu Building, Shatin, New Territories
Hong Kong 999077, CN

(72) Inventor/es:

LO, YUK-MING DENNIS;
CHIU, ROSSA WAI KWUN;
CHAN, KWAN CHEE y
JIANG, PEIYONG

(74) Agente/Representante:

PONS ARIÑO, Ángel

DESCRIPCIÓN

Detección de mutaciones para cribado de cáncer y análisis fetal

Antecedentes

Se ha demostrado que el ADN derivado de tumor está presente en el plasma/suero sin células de pacientes con cáncer (Chen *et al.* Nat Med 1996; 2: 1033-1035). La mayoría de los métodos actuales se basan en el análisis directo de mutaciones que se sabe que están asociadas con el cáncer (Diehl *et al.* Proc Natl Acad Sci USA 2005; 102: 16368-16373; Forshew *et al.* Sci Transl Med 2012; 4: 136ra68). Pero, dicho análisis directo de un panel de mutaciones predeterminadas para analizar ha tenido una precisión baja en el cribado de cáncer, por ejemplo, analizando el ADN plasmático.

Además, un análisis directo de este tipo utilizando un panel de mutaciones predeterminadas proporciona una visión limitada de la composición genética de un tumor. Por lo tanto, normalmente se toman biopsias quirúrgicas para realizar la secuenciación de un tumor, para obtener información genética sobre el tumor. La necesidad de cirugía aumenta los riesgos y costes. Además, para encontrar una ubicación de un tumor, se necesitan costosas técnicas de escaneo antes de que se pueda realizar una biopsia quirúrgica.

Por lo tanto, es deseable proporcionar nuevas técnicas para realizar un cribado, detección o evaluación del cáncer amplios, particularmente de una manera no invasiva.

Breve sumario

Las realizaciones se refieren a la detección precisa de mutaciones somáticas en el plasma (u otras muestras que contienen ADN sin células) de pacientes con cáncer y para sujetos que se someten a pruebas de cribado de cáncer. La detección de estos marcadores moleculares sería útil para el cribado, detección, monitorización, control y pronóstico de pacientes con cáncer. Por ejemplo, se puede determinar una carga mutacional a partir de las mutaciones somáticas identificadas y la carga mutacional se puede usar para cribar cualquiera o varios tipos de cáncer, donde no se necesiten conocimientos previos sobre un tumor o posible cáncer del sujeto. Las realizaciones pueden ser útiles para guiar el uso de terapias (por ejemplo, terapia dirigida, inmunoterapia, edición genómica, cirugía, quimioterapia, terapia de embolización, terapia antiangiogénica) para los cánceres. Las realizaciones también se refieren a identificar mutaciones *de novo* en un feto mediante el análisis de una muestra materna que tiene ADN sin células del feto.

Otras realizaciones se refieren a sistemas y medios legibles por ordenador asociados con los métodos descritos en el presente documento.

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

Breve descripción de los dibujos

La figura 1 muestra una tabla 100 de las 28 principales mutaciones más comúnmente identificadas entre los cánceres.

La figura 2 es una tabla 200 que muestra un número esperado de mutaciones a detectar para diferentes fracciones de ADN tumoral, profundidades de secuenciación, número de mutaciones por genoma y la fracción del genoma buscado.

La figura 3 es un gráfico 300 que muestra la relación entre el porcentaje de lecturas de secuencias de réplicas de PCR y la profundidad de secuenciación.

Las figuras 4A y 4B muestran una comparación entre la profundidad de secuenciación necesaria para PCR y protocolos sin PCR para detectar mutaciones paraneoplásicas en el plasma de un sujeto con cáncer en varias fracciones de ADN tumoral según las realizaciones de la presente invención.

La figura 5 es un diagrama de Venn que muestra el número de ubicaciones de terminación frecuentes que son específicas para el caso HCC, específicas para la mujer gestante o compartidas por ambos casos según las realizaciones de la presente invención.

La figura 6 es un gráfico 600 que muestra aumentos, disminuciones o ningún cambio en los segmentos de 1 Mb para el paciente con HCC.

La figura 7 muestra un proceso 700 de filtrado, que utiliza valor de corte dinámico, realineación y fracción de mutación y los datos resultantes para las mutaciones identificadas a partir de una biopsia tumoral según las realizaciones de la presente invención.

La figura 8 muestra un gráfico 800 de tamaños de fragmentos de ADN plasmático identificados con un alelo mutante para el paciente con HCC en comparación con los tamaños de fragmentos de ADN plasmático identificados con el alelo de tipo silvestre.

La figura 9 muestra un proceso 900 de filtrado, que utiliza valor de corte dinámico, realineación y fracción de mutación y los datos resultantes para las mutaciones identificadas a partir de una biopsia de hígado normal adyacente según las realizaciones de la presente invención.

Las figuras 10A y 10B muestran una comparación del perfil de tamaño evaluado de los fragmentos de ADN plasmático que portan las 203 supuestas mutaciones identificadas a partir de la biopsia de hígado normal adyacente con el tamaño proporcionado por otras moléculas de ADN plasmático no informativas.

La figura 11 muestra un proceso 1100 de filtrado (que utiliza valor de corte dinámico, realineación, fracción de mutación y tamaño), y los datos resultantes para las mutaciones identificadas a partir del plasma según las realizaciones de la presente invención.

La figura 12 muestra un proceso 1200 de filtrado y los datos resultantes para las mutaciones identificadas a partir del plasma usando cortes dinámicos de fracciones mutantes inferiores según las realizaciones de la presente invención.

La figura 13 muestra un proceso 1300 de filtrado (que utiliza valor de corte dinámico, realineación y tamaño), y los datos resultantes para las mutaciones identificadas a partir del plasma según las realizaciones de la presente invención.

La figura 14 muestra un gráfico 1400 de tamaños de fragmentos de ADN plasmático identificados con un alelo mutante utilizando plasma en comparación con los tamaños de fragmentos de ADN plasmático identificados con el alelo de tipo silvestre.

La figura 15 muestra un proceso de filtrado 1500 y los datos resultantes para las mutaciones identificadas a partir del plasma utilizando una mayor profundidad de secuenciación según las realizaciones de la presente invención.

La figura 16 es un gráfico 1600 que muestra el número (densidad) de locus que tienen varios valores de fracción mutante.

La figura 17A muestra puntuaciones z para la distribución sobre los brazos cromosómicos 1p y 1q. La figura 17B muestra la fracción mutante aparente sobre los brazos cromosómicos 1p y 1q.

La figura 18 es una tabla que muestra las sensibilidades de detección de mutaciones predichas para varias fracciones de mutación y profundidades de secuenciación para determinados cortes dinámicos de recuento de alelos según las realizaciones de la presente invención.

La figura 19 es una tabla 1900 que muestra las sensibilidades predichas de detección de mutaciones para varias fracciones de mutación y profundidades de secuenciación para determinados cortes dinámicos de recuentos de alelos para una tasa de detección de falsos positivos del 0,1 % según las realizaciones de la presente invención.

La figura 20 muestra un proceso 2000 de filtrado y los datos resultantes para las mutaciones identificadas a partir del plasma usando valores de corte dinámicos menos rigurosos según las realizaciones de la presente invención.

La figura 21 es un gráfico 2100 que muestra las distribuciones del número de supuestas mutaciones para escenarios fetales y de cáncer.

La figura 22 es un gráfico 2200 que muestra las distribuciones del número de supuestas mutaciones para escenarios fetales y de cáncer cuando se usa realineación.

La figura 23 es una tabla 2300 que muestra los PPV y las tasas de recuperación para cortes dinámicos de varios tamaños sin realineación según las realizaciones de la presente invención.

La figura 24 es una tabla 2400 que muestra los PPV y las tasas de recuperación para cortes dinámicos de varios tamaños con realineación según las realizaciones de la presente invención.

La figura 25 muestra un proceso 2500 de filtrado (que utiliza valor de corte dinámico, realineación y tamaño), y los datos resultantes para las mutaciones identificadas a partir del plasma de sangre del cordón umbilical según las realizaciones de la presente invención.

La figura 26 es un gráfico 2600 de distribuciones de tamaño para fragmentos de ADN mutantes determinados a

partir del proceso 2500 y alelos de tipo silvestre según las realizaciones de la presente invención.

La figura 27 muestra un proceso 2700 de filtrado (que utiliza valor de corte dinámico, realineación y tamaño), y los datos resultantes para las mutaciones identificadas a partir del plasma de una muestra de HCC según las realizaciones de la presente invención.

La figura 28 es un gráfico 2800 de distribuciones de tamaño para fragmentos de ADN mutantes determinados a partir del proceso 2700 y alelos de tipo silvestre según las realizaciones de la presente invención.

La figura 29 muestra un proceso 2900 de filtrado que usa filtrado basado en SNP para mutaciones identificadas a partir de plasma de sangre de cordón umbilical según las realizaciones de la presente invención.

La figura 30 muestra un proceso 3000 de filtrado que usa filtrado basado en SNP para mutaciones identificadas a partir de plasma de HCC según las realizaciones de la presente invención

La figura 31 es una tabla 3100 que muestra correlaciones de tejido con modificaciones de histonas.

La figura 32 muestra la distribución de frecuencias de las fracciones fetales medidas en sitios con SNP individuales.

La figura 33A muestra una distribución de tamaño de ADN específico fetal y ADN compartido en plasma materno. La figura 33B muestra un gráfico de frecuencias acumulativas para el tamaño del ADN plasmático para el fragmento de ADN compartido y específico fetal. La figura 33C muestra la diferencia en frecuencias acumulativas, denominada ΔF.

La figura 34A muestra la distribución de tamaño de los fragmentos de ADN plasmático con el alelo mutante. La figura 34B muestra un gráfico de frecuencias acumulativas para el tamaño del ADN plasmático para el alelo mutante y el alelo de tipo silvestre. La figura 34C muestra la diferencia en frecuencias acumulativas, denominada ΔF

La figura 35 muestra un proceso 3300 de filtrado (que utiliza valor de corte dinámico, realineación, fracción de mutación y valor de corte por tamaño) y los datos resultantes para mutaciones *de novo* identificadas a partir de plasma según las realizaciones de la presente invención.

La figura 36A muestra perfiles de tamaño de fragmentos de ADN con las supuestas mutaciones identificadas en plasma utilizando criterios de filtrado de nivel A en comparación con el alelo de tipo silvestre. La figura 36B muestra perfiles de tamaño de fragmentos de ADN con las supuestas mutaciones identificadas en plasma usando criterios de filtrado de nivel B. La figura 36C muestra perfiles de tamaño de fragmentos de ADN con las supuestas mutaciones identificadas en plasma usando criterios de filtrado de nivel C. La figura 36D muestra perfiles de tamaño de fragmentos de ADN con las supuestas mutaciones identificadas en plasma usando criterios de filtrado de nivel D.

La figura 37 muestra los perfiles de los valores de ΔF correspondientes a supuestas mutaciones identificadas utilizando diferentes niveles de criterios de filtrado, concretamente, A, B, C y D.

La figura 38 muestra un recuento de frecuencias de varios tipos de mutaciones en una muestra de plasma materno y sangre del cordón umbilical.

La figura 39A muestra un gráfico del % de PPV y tasas de recuperación para filtros de diferentes tamaños según las realizaciones de la presente invención. La figura 39B muestra un gráfico del % de PPV y las tasas de recuperación para diferentes cortes dinámicos de fracción mutante.

Las figuras 40A-40D muestran gráficos del % de PPV y tasas de recuperación para filtros de varios tamaños en diferentes cortes dinámicos de fracción mutante.

La figura 41 es un gráfico que muestra las curvas de las tasas de recuperación y el % de PPV en diferentes cortes dinámicos de fracción mutante en función de los cortes dinámicos de tamaño.

Las figuras 42 y 43 muestran una tabla de las 47 mutaciones *de novo*.

La figura 44 muestra las tasas de recuperación y los PPV para la detección de las 47 mutaciones *de novo* y las 3.000 supuestas mutaciones somáticas

Las figuras 45A-45C y 46A-46C muestran simulaciones con cantidades variables de mutaciones para diversas profundidades de secuenciación y fracciones tumorales.

La figura 47 es un diagrama de flujo que ilustra un método 4700 para identificar mutaciones somáticas en un sujeto

humano mediante el análisis de una muestra biológica del sujeto humano según las realizaciones de la presente invención.

La figura 48 es un diagrama de flujo que ilustra un método 4800 para usar mutaciones somáticas identificadas para analizar muestras biológicas de un sujeto según las realizaciones de la presente invención.

La figura 49 es un diagrama de flujo que ilustra un método 4900 para identificar mutaciones *de novo* de un feto mediante el análisis de una muestra biológica de una mujer gestante del feto según las realizaciones de la presente invención.

La figura 50 muestra un diagrama de bloques de un sistema informático 10 ilustrativo que puede utilizarse con el sistema y los métodos según las realizaciones de la presente invención.

EXPRESIONES

La expresión "*muestra biológica*" se refiere a cualquier muestra que se toma de un sujeto (por ejemplo, un ser humano, una persona con cáncer, una persona que se sospecha que tiene cáncer, una persona para someter a cribado para determinar cáncer, una mujer gestante u otros organismos). Una muestra biológica puede incluir ADN sin células, algunos de los cuales pueden haberse originado a partir de células sanas y otros a partir de células tumorales. El ADN sin células se puede encontrar en la sangre o sus componentes (por ejemplo, plasma o plaquetas) o sus derivados (por ejemplo, suero) u otros fluidos, por ejemplo, orina, otros fluidos del tracto urogenital, sudor, líquido pleural, líquido ascítico, líquido peritoneal, saliva, lágrimas, secreción del pezón, líquido cefalorraquídeo, líquido intraocular, líquido amniótico y líquido de lavado cervical. Un ejemplo no fluido es una muestra de heces, que puede mezclarse con líquido diarreico. Para algunas de dichas muestras, la muestra biológica se puede obtener de forma no invasiva. En algunas realizaciones, la muestra biológica puede utilizarse como muestra constitutiva.

Tal como se usa en el presente documento, el término *"locus"* o su forma plural *"locus"* es una ubicación o dirección de cualquier longitud de nucleótidos (o pares de bases) que pueden tener una variación entre genomas de diferentes individuos o entre diferentes células dentro de un individuo (por ejemplo, entre células tumorales y células sanas).

La expresión *"secuenciación aleatoria"* tal como se usa en el presente documento, se refiere a la secuenciación en la que los fragmentos de ácido nucleico secuenciados no se han identificado o predeterminado específicamente antes del procedimiento de secuenciación. No se requieren cebadores específicos de secuencia para dirigirse a locus de genes específicos. En una realización, los adaptadores se añaden al final de un fragmento y los cebadores para la secuenciación se fijan a los adaptadores. Por lo tanto, cualquier fragmento se puede secuenciar con el mismo cebador y, por lo tanto, la secuenciación puede ser aleatoria. La secuenciación masiva en paralelo se puede realizar usando secuenciación aleatoria.

La expresión *"etiqueta de secuencia"* (también denominada lectura de secuencia) como se usa en el presente documento se refiere a una cadena de nucleótidos secuenciada de cualquier parte o de la totalidad de una molécula de ácido nucleico. Por ejemplo, una etiqueta de secuencia puede ser una cadena corta de nucleótidos (por ejemplo, ~ 30) secuenciados a partir de un fragmento de ácido nucleico, una cadena corta de nucleótidos en ambos extremos de un fragmento de ácido nucleico o la secuenciación de todo el fragmento de ácido nucleico que existe en la muestra biológica. Un fragmento de ácido nucleico es cualquier parte de una molécula de ácido nucleico más grande. Un fragmento (por ejemplo, un gen) puede existir por separado (es decir, no conectado) a las otras partes de la molécula de ácido nucleico más grande.

Una *"variante de secuencia"* (también llamada variante) corresponde a las diferencias de un genoma de referencia, que podría ser un genoma constitutivo de un organismo o genomas precursores. Los ejemplos de variantes de secuencia incluyen una variante de un solo nucleótido (SNV, por sus siglas en inglés) y variantes que implican dos o más nucleótidos. Los ejemplos de SNV incluyen polimorfismos de un solo nucleótido (SNP, por sus siglas en inglés) y mutaciones puntuales. Como ejemplos, las mutaciones pueden ser mutaciones *"de novo*" (por ejemplo, nuevas mutaciones en el genoma constitutivo de un feto) o "mutaciones somáticas" (por ejemplo, mutaciones en un tumor). Un alelo de tipo silvestre corresponde a un alelo en el genoma constitutivo. Un genoma constitutivo puede contener dos alelos de tipo silvestre si el sujeto es heterocigoto en ese locus. Una variante de secuencia de tipo silvestre corresponde a la secuencia en una ubicación particular en el genoma constitutivo. Un genoma constitutivo puede contener dos variantes de secuencia de tipo silvestre si el sujeto es heterocigoto en ese locus.

Una *"mutación somática"* se refiere a mutaciones en tejidos o células que se desarrollan después del nacimiento. Los organismos acumulan más mutaciones con la edad, debido a errores en la replicación del ADN o como resultado de la exposición a carcinógenos u otros factores ambientales. Normalmente, los seres humanos adquieren una mutación por célula por división celular. Pero individualmente, dichas mutaciones están presentes en una concentración extremadamente baja en el tejido porque no son clonales. Sin embargo, las mutaciones oncógenas se amplifican clonalmente y están presentes en una concentración fraccionaria más alta en un tejido tumoral. La concentración fraccionaria de diferentes mutaciones en un cáncer puede ser diferente debido a la heterogeneidad tumoral. Esto significa que un tumor normalmente se compone de muchos clones diferentes y cada clon tiene su propio perfil

mutacional.

Los "*cambios paraneoplásicos*" o "*cambios específicos del cáncer*" incluyen, pero sin limitación, mutaciones derivadas del cáncer (incluidas mutaciones de un solo nucleótido, deleciones o inserciones de nucleótidos, deleciones de segmentos genéticos o cromosómicos, translocaciones, inversiones), amplificación de genes, segmentos genéticos o cromosómicos, secuencias asociadas a virus (por ejemplo, episomas víricos e inserciones víricas), perfiles de metilación aberrantes o firmas de metilación específicas de tumores, perfiles de tamaño del ADN sin células aberrantes, marcas aberrantes de modificación de histonas y otras modificaciones epigenéticas, y ubicaciones de los extremos de los fragmentos de ADN sin células que son paraneoplásicos o específicos del cáncer.

Un "*fragmento de ADN canceroso informativo*" corresponde a un fragmento de ADN portador de uno o más de los cambios o mutaciones paraneoplásicos o específicos del cáncer. Un "*fragmento de ADN fetal informativo*" corresponde a un fragmento de ADN fetal portador de una mutación que no se encuentra en ninguno de los genomas de los progenitores. Un "*fragmento de ADN informativo*" puede referirse a cualquiera de los tipos de fragmentos de ADN mencionados anteriormente.

La expresión "*profundidad de secuenciación*" se refiere al número de veces que un locus está cubierto por una lectura de secuencia alineada con el locus. El locus puede ser tan pequeño como un nucleótido o tan grande como un brazo cromosómico o tan grande como todo el genoma. La profundidad de secuenciación puede expresarse como 50x, 100x, etc., donde "x" se refiere al número de veces que se cubre un locus con una lectura de secuencia. La profundidad de secuenciación también puede aplicarse a múltiples locus o a todo el genoma, en cuyo caso x puede referirse al número medio de veces que se secuencian los locus o el genoma completo, respectivamente. La secuenciación ultra profunda puede referirse a una profundidad de secuenciación de al menos 100x.

La expresión *"amplitud de secuenciación"* se refiere a qué fracción de un genoma de referencia en particular (por ejemplo, humano) o parte del genoma se ha analizado. El denominador de la fracción podría ser un genoma enmascarado repetido y, por lo tanto, el 100 % puede corresponder a todo el genoma de referencia menos las partes enmascaradas. Se puede enmascarar cualquier parte de un genoma y, por lo tanto, se puede enfocar el análisis en cualquier parte particular de un genoma de referencia. La secuenciación amplia puede referirse a al menos el 0,1 % del genoma que se analiza, por ejemplo, mediante la identificación de lecturas de secuencias que se alinean con esa parte de un genoma de referencia.

"Secuenciación exhaustiva" se refiere a la obtención de información molecular de casi todos los fragmentos de ácido nucleico clínica o biológicamente pertinentes analizables en la práctica en una muestra, por ejemplo, plasma. Debido a las limitaciones en las etapas de preparación de muestras, etapas de preparación de la biblioteca de secuenciación, secuenciación, asignación de bases y alineación, no todas las moléculas nucleicas del plasma (por ejemplo, ADN o ARN) en una muestra serían analizables o secuenciables.

Un *"molécula de ADN analizable"* se refiere a cualquier molécula de ADN que haya superado con éxito todas las etapas analíticas para analizarse y detectarse por cualquier medio adecuado, incluida la secuenciación. A *"molécula de ADN secuenciable"* se refiere a cualquier molécula de ADN que haya superado con éxito todas las etapas analíticas para secuenciarse y detectarse bioinformáticamente. Por lo tanto, la secuenciación exhaustiva puede referirse a procedimientos implementados para maximizar la capacidad de transformar tantas moléculas de ADN clínica o biológicamente pertinentes (por ejemplo, fragmentos de ADN informativos) en una muestra de plasma finita en moléculas secuenciables. Después de que haber creado una biblioteca de secuenciación de moléculas de ADN secuenciables utilizando dichos procedimientos, se puede secuenciar la totalidad o parte de la biblioteca. Si realmente se consumen completamente las moléculas de ADN secuenciables de la muestra finita para obtener información de secuencia, este acto podría denominarse *"secuenciación total de plantillas"* que corresponde a un espectro de secuenciación exhaustiva.

Una *"carga mutacional"* de una muestra es un valor medido basándose en cuántas mutaciones se miden. La carga mutacional se puede determinar de varias maneras, tal como un número en bruto de mutaciones, una densidad de mutaciones por número de bases, un porcentaje de locus de una región genómica que se identifican con mutaciones, el número de mutaciones observadas en una cantidad particular (por ejemplo, volumen) de muestra, y aumento proporcional o veces mayor en comparación con los datos de referencia o desde la última evaluación. Una *"evaluación de la carga mutacional"* se refiere a una medida de la carga mutacional de una muestra.

El *"valor predictivo positivo (PPV, por sus siglas en inglés)"* de una prueba de cribado se refiere al número de verdaderos positivos (TP, por sus siglas en inglés) identificados por una prueba expresado como una proporción de la suma de los verdaderos positivos y los falsos positivos (FP) clasificados por la prueba, por ejemplo, TP/(TP+FP). Un *"valor predictivo negativo (NPV)"* se refiere al número de verdaderos negativos (TN, por sus siglas en inglés) identificados por la prueba expresado como una proporción de la suma de verdaderos negativos y falsos negativos (FN) clasificados por la prueba, por ejemplo, TN/(TN+FN).

La expresión *"genoma constitutivo"* (también denominado CG, por sus siglas en inglés) está compuesto por los nucleótidos consenso en los locus dentro del genoma y, por lo tanto, puede considerarse una secuencia consenso. El

CG puede cubrir todo el genoma del sujeto (por ejemplo, el genoma humano) o solo partes del genoma. El genoma constitutivo (CG) se puede obtener a partir del ADN de las células, así como del ADN sin células (por ejemplo, como se puede encontrar en el plasma). De manera ideal, los nucleótidos consenso deben indicar que un locus es homocigoto para un alelo o heterocigoto para dos alelos. Un locus heterocigoto normalmente contiene dos alelos que son miembros de un polimorfismo genético. A modo de ejemplo, el criterio para determinar si un locus es heterocigoto puede ser un umbral de dos alelos que aparezcan cada uno en al menos un porcentaje predeterminado (por ejemplo, el 30 % o el 40 %) de lecturas alineadas con el locus. Si un nucleótido aparece en un porcentaje suficiente (por ejemplo, el 70 % o más), se puede determinar que el locus es homocigoto en el CG. Aunque el genoma de una célula sana puede diferir del genoma de otra célula sana debido a mutaciones aleatorias que se producen de manera espontánea durante la división celular, el CG no debe variar cuando se utiliza dicho consenso. Algunas células pueden tener genomas con reordenamientos genómicos, por ejemplo, linfocitos B y T, tal como los que implican a genes receptores de linfocitos T y anticuerpos, respectivamente. Dichas diferencias a gran escala seguirían siendo una población relativamente pequeña de la población total de células nucleadas en sangre y, por lo tanto, dichos reordenamientos no afectarían la determinación del genoma constitutivo con un muestreo suficiente (por ejemplo, profundidad de secuenciación) de las células sanguíneas. Otros tipos de células, incluidas las células bucales, células de la piel, folículos pilosos o biopsias de varios tejidos corporales normales, también pueden servir como fuentes de CG.

La expresión *"ADN constitutivo"* se refiere a cualquier fuente de ADN que refleje la composición genética con la que nace un sujeto. Pueden producirse mutaciones aleatorias durante la división celular. A diferencia de las mutaciones paraneoplásicas, no hay amplificación clonal de las mutaciones aleatorias. Por lo tanto, el CG obtenido de la secuencia de consenso del ADN constitutivo refleja la composición genética con la que nace un sujeto. Para un sujeto, ejemplos de "muestras constitutivas", a partir de las que se puede obtener ADN constitutivo, incluyen ADN de glóbulos rojos sanos, ADN de células bucales, ADN de la raíz del cabello, ADN salival y ADN de raspados de piel. El ADN de estas células sanas define el CG del sujeto. Las células se pueden identificar como sanas de varias maneras, por ejemplo, cuando se sabe que una persona no tiene cáncer o la muestra se puede obtener de un tejido que probablemente no contiene células cancerosas o premalignas (por ejemplo, ADN de la raíz del cabello cuando se sospecha cáncer de hígado). Como otro ejemplo, se puede obtener una muestra de plasma cuando un paciente no tiene cáncer, y el ADN constitutivo determinado se compara con los resultados de una muestra de plasma posterior (por ejemplo, un año o más después). En otra realización, una sola muestra biológica que contenga <50 % de ADN tumoral se puede utilizar para deducir el genoma constitutivo y las alteraciones genéticas oncógenas. En dicha una muestra, las concentraciones de mutaciones de un solo nucleótido oncógenas serían más bajas que las de cada alelo de los SNP heterocigotos en el CG. Dicha muestra puede ser la misma que la muestra biológica utilizada para determinar un genoma de muestra, descrito a continuación.

La expresión *"genoma de muestra"* (también denominado SG, por sus siglas en inglés) es una colección de lecturas de secuencias que se han alineado con ubicaciones de un genoma (por ejemplo, un genoma humano). El genoma de muestra (SG) no es una secuencia consenso, pero incluye nucleótidos que pueden aparecer solamente en un número suficiente de lecturas (por ejemplo, al menos 2 o 3, o valores de corte dinámicos más altos). Si un alelo aparece un número suficiente de veces y no es parte del CG (es decir, no es parte de la secuencia consenso), entonces ese alelo puede indicar una "mutación de un solo nucleótido" (también conocida como SNM, por sus siglas en inglés). También se pueden detectar otros tipos de mutaciones, por ejemplo, mutaciones que implican dos o más nucleótidos (tales como las que afectan el número de unidades de repetición en tándem en un microsatélite o polimorfismo de repetición en tándem simple), translocación cromosómica (que puede ser intracromosómica o intercromosómica) e inversión de secuencia.

La expresión *"genoma de referencia"* (también conocido como RG, por sus siglas en inglés) se refiere a un genoma haploide o diploide cuyas lecturas de secuencias de la muestra biológica y la muestra constitutiva se puede alinear y comparar. Para un genoma haploide, solo hay un nucleótido en cada locus. Para un genoma diploide, se pueden identificar locus heterocigotos, teniendo dicho locus dos alelos, donde cualquiera de los alelos puede permitir una coincidencia para la alineación con el locus.

La expresión *"nivel de cáncer"* puede referirse a si existe cáncer, un estadio de un cáncer, un tamaño del tumor, la respuesta del cáncer al tratamiento y/u otra medida de la gravedad o progresión de un cáncer. La carga mutacional se puede utilizar para determinar el nivel de cáncer. Cuanto más avanzado esté el cáncer, mayor será la carga mutacional. El nivel de cáncer podría ser un número u otros caracteres, tal como letras u otros símbolos. El nivel podría ser cero. El nivel de cáncer también incluye las afecciones (estados) premalignas o precancerosas asociadas a mutaciones o a un número de mutaciones. El nivel de cáncer puede usarse de varias maneras. Por ejemplo, el cribado puede comprobar si el cáncer está presente en una persona que no se sabe que ha tenido cáncer con anterioridad. La evaluación puede investigar a alguien a quien se le ha diagnosticado cáncer. Detección puede significar "cribado" o puede significar comprobar si alguien, con características sugestivas de cáncer (por ejemplo, síntomas u otras pruebas positivas) o con factores de riesgo de cáncer (por ejemplo, hábitos como fumar o beber alcohol o antecedentes de infecciones víricas, por ejemplo, infección por el virus de la hepatitis), tiene cáncer.

El término "*clasificación*", como se usa en el presente documento, se refiere a cualquier número y uno o más caracteres distintos que se asocian a una propiedad particular de una muestra. Por ejemplo, un símbolo "+" (o la palabra "positivo") podría significar que una muestra se clasifica con un nivel particular de cáncer. La clasificación puede ser binaria (por

ejemplo, positiva o negativa) o tener más niveles de clasificación (por ejemplo, una escala de 1 a 10 o de 0 a 1). El término "*valor de corte*" y "*umbral*" se refieren a números predeterminados utilizados en una operación. Un valor umbral puede ser un valor por encima o por debajo del cual se aplica una determinada clasificación. Se puede predeterminar un valor de corte con o sin referencia a las características de la muestra o de la persona. Por ejemplo, los valores de corte se pueden elegir en función de la edad o el sexo del individuo analizado. Se puede elegir un valor de corte después y en función de la salida de los datos de prueba. Por ejemplo, se pueden usar determinados valores de corte cuando la secuenciación de una muestra alcanza una determinada profundidad.

Descripción detallada

La identificación de mutaciones en una muestra biológica de un organismo (por ejemplo, debidas a cáncer o en un feto) se ve obstaculizada por la prevalencia de errores de secuenciación y otras dificultades. Las realizaciones proporcionan técnicas para identificar con precisión mutaciones en un organismo mediante el análisis de moléculas (fragmentos) de ADN sin células del organismo. Para un análisis fetal de una muestra obtenida de forma no invasiva, las moléculas de ADN sin células del feto estarían en una muestra materna (por ejemplo, plasma materno) que también contiene moléculas de ADN sin células de la mujer gestante. Se puede identificar un número significativo de mutaciones verdaderas (a diferencia de los falsos positivos) o se puede potenciar sustancialmente la proporción de mutaciones verdaderas detectadas usando determinadas técnicas de secuenciación (por ejemplo, preparación de bibliotecas de secuenciación sin PCR) y determinados criterios de filtrado.

Cuando se utiliza una profundidad de secuenciación y una amplitud de secuenciación suficientes, se puede determinar una medida precisa de la carga mutacional de un sujeto, permitiendo, de este modo, una evaluación de un nivel de cáncer en el sujeto. A continuación, se describe la base teórica y la implementación práctica de los requisitos de los marcadores tumorales basados en ADN (por ejemplo, en plasma) para la detección, monitorización y pronóstico de cáncer.

I. MARCADORES MUTACIONALES DEL CÁNCER

No muchos cánceres tienen marcadores mutacionales o de otro tipo claros para identificar que el cáncer existe o es muy probable que esté presente en un individuo. E incluso si tales marcadores existen, por lo general, hay pocos marcadores conocidos que sean exclusivos para un cáncer específico. Por lo tanto, puede ser difícil detectar el cáncer en plasma u otra muestra similar con ADN sin células, donde dichos marcadores mutacionales no estarían en alta concentración. Una excepción es el ADN del virus del virus de Epstein-Barr (EBV, por sus siglas en inglés) en pacientes con carcinoma nasofaríngeo (NPC, por sus siglas en inglés). Por lo tanto, el ADN del EBV se puede encontrar en los núcleos de las células tumorales de NPC en la mayoría de los casos de NPC en China (Tsang *et al.* Chin J Cancer 2014; 33: 549-555). Asimismo, el ADN del EBV se puede encontrar en el plasma de pacientes con NPC (Lo *et al.* Cancer Res 1999; 59: 1188-1191).

Este ejemplo se utiliza para ilustrar la dificultad de obtener suficientes datos para cribar cáncer utilizando mutaciones puntuales de un panel para cribar un tipo particular de cáncer. Este ejemplo ilustra adicionalmente la necesidad de detectar muchas mutaciones en el plasma para alcanzar la sensibilidad para el cribado del cáncer.

A. ADN del EBV en pacientes con NPC

El NPC está estrechamente asociado con la infección por EBV. En el sur de China, el genoma de EBV se puede encontrar en los tejidos tumorales en casi todos los pacientes con NPC. El ADN del EBV en plasma derivado de tejidos con NPC se ha creado como un marcador tumoral para NPC (Lo *et al.* Cancer Res 1999; 59: 1188-1191). Este marcador tumoral ha demostrado ser útil para la monitorización (Lo *et al.* Cancer Res 1999; 59: 5452-5455) y pronóstico (Lo *et al.* Cancer Res 2000; 60: 6878-6881) de NPC. Se ha demostrado que el análisis del ADN del EBV en plasma mediante PCR en tiempo real es útil para la detección temprana de NPC en sujetos asintomáticos y puede ser potencialmente útil para el cribado de NPC (Chan *et al.* Cancer 2013;119:1838-1844). En este estudio anterior, el ensayo de PCR en tiempo real utilizado para el análisis de ADN del EBV en plasma se dirigió al fragmento*Bam*HI-W del genoma de EBV. Hay aproximadamente de seis a doce repeticiones de los fragmentos *Bam*HI-W en cada genoma de EBV y hay aproximadamente 50 genomas de EBV en cada célula tumoral de NPC (Longnecker *et al.* Fields Virology, 5ª edición, capítulo 61 "Epstein-Barr virus"; Tierney *et al.* J Virol. 2011; 85: 12362-12375). En otras palabras, habría del orden de 300-600 (por ejemplo, aproximadamente 500) copias de la diana de PCR en cada célula tumoral de NPC. Este elevado número de dianas por célula tumoral puede explicar por qué el ADN del EBV en plasma es tan sensible en la detección de NPC temprano.

B. Secuenciación dirigida para ADN del EBV

Como se ilustra en el ejemplo anterior, la alta sensibilidad del análisis de PCR en tiempo real del ADN del EBV en plasma está relacionada con la presencia de múltiples copias de la diana de PCR en cada genoma tumoral de NPC. Por lo tanto, los presentes inventores consideran que un aumento adicional en el número de dianas oncógenas que se buscaría detectar en el plasma del paciente con cáncer aumentaría adicionalmente la sensibilidad y la utilidad clínica del análisis de ADN plasmático. Las moléculas de ADN del EBV en el plasma de pacientes con NPC son

principalmente fragmentos cortos de menos de 180 pb (Chan *et al.* Cancer Res 2003; 63: 2028-2032). Como el tamaño de un genoma de EBV es de aproximadamente 172 kb, cada genoma de EBV se fragmentaría en aproximadamente 1.000 fragmentos de ADN plasmático. Por lo tanto, los 50 genomas de EBV en una célula tumoral de NPC se fragmentarían en unos 50.000 fragmentos de ADN plasmático y se liberarían en la circulación de un paciente con NPC.

Los presentes inventores consideran que a cuantos más de estos 50.000 fragmentos de ADN del EBV derivados de tumores se puedan dirigir, mayor será la sensibilidad para detectar un cáncer asociado con EBV que se podría conseguir. Se puede detectar un 5 %, 10 %, 20 %, 25 %, 30 %, 40 %, 50 %, 75 %, 90 % o 99 % del genoma de EBV para su uso en análisis. Se puede intentar dirigirse a las partes del genoma de EBV que se podrían diferenciar bioinformáticamente del genoma humano.

La alta sensibilidad de detección que ofrece la detección de una multiplicidad tan alta de dianas genómicas de EBV en plasma es particularmente importante en la detección de la recurrencia de la enfermedad en pacientes que reciben radioterapia con intención curativa. La tasa de detección de NPC recurrente en pacientes que recibieron radioterapia con intención curativa es inferior a la tasa de detección de NPC sin tratamiento previo (Leung *et al.* Clin Cancer Res 2003; 9: 3431-3134). Las tasas generales de detección para los dos grupos de cánceres utilizando PCR de ADN del EBV en tiempo real dirigidas al fragmento *Bam*HI-W fueron 62,5 % y 96,4 %, respectivamente. Dichas altas tasas de detección ilustran la necesidad de una alta multiplicidad en cualquier técnica de cribado. Dicha alta multiplicidad en una diana altamente correlacionada normalmente no está disponible para otros tipos de cáncer.

Se esperaría que la detección de una alta multiplicidad de dianas genómicas de EBV (o mutaciones deducidas como se describe más adelante) en plasma aumente la tasa de detección en el primer grupo. Otra utilidad de esta estrategia sería para el cribado de NPC. Para el cribado, es particularmente importante que se pueda detectar el cáncer en estadio temprano. Un sistema de detección de ADN del EBV en plasma altamente sensible permitiría este objetivo. Como se explica más adelante, las realizaciones pueden proporcionar una detección altamente sensible sin requerir el uso de un marcador mutacional o molecular diferente predeterminado.

II. CRIBADO DE CÁNCER

Un problema en el cribado de cáncer es que puede no saberse qué tipo de cáncer podría tener un sujeto o estar predispuesto al mismo. Otro problema es que un individuo puede ser susceptible a más de un tipo de cáncer. En consecuencia, las realizaciones pueden identificar mutaciones a partir de una muestra biológica del sujeto, por lo tanto, no es necesario cribar solamente un panel predeterminado de mutaciones. Los detalles sobre cómo identificar con precisión las mutaciones del ADN sin células en una muestra se describen en secciones posteriores. A continuación, se describen los procesos y las dificultades del cribado de cáncer.

Una vez que se identifican las mutaciones en una muestra biológica (por ejemplo, plasma), las mutaciones se pueden utilizar en el cribado de cáncer. El término cribado generalmente se refiere a la identificación de enfermedades a través del acto proactivo de realizar algún tipo de evaluación. Las herramientas de evaluación podrían incluir la evaluación del perfil demográfico de una persona, realizando análisis de sangre, pruebas de otros fluidos corporales (por ejemplo, orina, líquido ascítico, líquido pleural, líquido cefalorraquídeo), pruebas en biopsias de tejido, endoscopia (por ejemplo, colonoscopia) y pruebas de obtención de imágenes (por ejemplo, obtención de imágenes por resonancia magnética, tomografía computarizada, ultrasonografía o tomografía por emisión de positrones). Se puede utilizar una combinación de las modalidades de evaluación, por ejemplo, se pueden usar múltiples muestras y los resultados se pueden combinar para proporcionar una evaluación final.

A. Diferentes etapas de cribado y evaluación probabilística

El cribado de enfermedades generalmente se puede aplicar en diferentes estadios de la enfermedad, en concreto, pero sin limitación, cribado primario, secundario y terciario. El cribado primario se refiere a la identificación de la enfermedad antes de la aparición de los síntomas y, a veces, se denomina cribado asintomático. El cribado primario se puede realizar en la población general o en una población seleccionada con características que los hacen estar en mayor riesgo para la enfermedad a cribar. Por ejemplo, los fumadores tienen un mayor riesgo de carcinoma microcítico de los pulmones. Los portadores crónicos de HBV tienen un mayor riesgo de HCC. El cribado secundario se refiere a la identificación de la enfermedad cuando el sujeto presenta síntomas y sería necesario hacer la diferenciación entre un grupo de presuntos diagnósticos. El cribado terciario se refiere a la identificación temprana de la progresión de la enfermedad, aumento en el estadio o la gravedad de la enfermedad (por ejemplo, la aparición de metástasis), o recaída de la enfermedad. En cada etapa de cribado de enfermedad o cribado de cáncer, el objetivo es identificar o excluir la presencia de enfermedad o la progresión de la enfermedad, generalmente antes de que el curso natural de la enfermedad se presente en síntomas, ya que las opciones de tratamiento pueden verse comprometidas o ser menos eficaces en un momento posterior.

El acto de cribado es una evaluación probabilística. En general, el fin del cribado es descartar (es decir, excluir) o aceptar (es decir, confirmar) un presunto diagnóstico. La evaluación es para determinar si una persona tiene una probabilidad alta o baja (denominada alternativamente riesgo) de padecer la enfermedad, tener la enfermedad o tener

progresión de la enfermedad. En otras palabras, después de cada evaluación se hace una clasificación de si el sujeto tiene alto o bajo riesgo. Es posible que se necesiten etapas sucesivas de evaluación y que se realicen pruebas repetidas.

B. Ejemplos de EBV

EBV se utiliza como un ejemplo que ilustra el cribado. Un hombre del sur de China de mediana edad tiene un mayor riesgo de padecer NPC que las personas con un perfil demográfico diferente. La prueba de ADN del EBV en plasma podría entonces aplicarse como una herramienta de cribado primario de este individuo. Si la carga de ADN del EBV en plasma está por debajo del valor de corte utilizado para diferenciar a los individuos con NPC, se consideraría que esta persona tiene pocas posibilidades de tener NPC en este momento (Chan *et al.* Cancer 2013; 119: 1838-1844). La persona puede elegir hacerse o se le puede recomendar que se haga la prueba de ADN del EBV en plasma nuevamente más tarde (por ejemplo, después de uno o dos años).

Si se encuentra que la carga de ADN del EBV en plasma es más alta que el valor de corte utilizado para diferenciar a aquellos con NPC, o muestra un aumento progresivo de los valores anteriores de la persona, esta persona puede considerarse de alto riesgo de tener NPC. Es posible que se recomiende a esta persona a la siguiente etapa de la prueba para descartar o aceptar la enfermedad, por ejemplo, utilizando otras pruebas para confirmar la enfermedad. Por ejemplo, se podría realizar otra prueba de ADN del VEB en plasma 2 o 6 semanas más tarde para evaluar si persiste la elevación del ADN del VEB en plasma. Según el índice de sospecha, se puede recomendar a la persona que se haga una endoscopia para la inspección visual de la nasofaringe con y sin biopsia de tejido adicional y evaluación histológica para confirmar la presencia de NPC. Como alternativa, se pueden realizar obtención de imágenes (por ejemplo, imágenes por resonancia magnética) para visualizar la presencia o ausencia de tumor. Dichos ejemplos ilustran los beneficios de que la detección sea capaz de dictar qué pruebas adicionales deben realizarse.

La misma prueba podría aplicarse como herramienta para el cribado secundario y terciario. A modo ilustrativo, la prueba de ADN del EBV en plasma podría usarse para evaluar la probabilidad de NPC en un sujeto que presenta epistaxis recurrente (es decir, sangrado por la nariz) o ronquera, que son síntomas comunes de presentación de NPC. Si los resultados de la prueba muestran que la carga de ADN del EBV es más alta que el valor de corte utilizado para diferenciar las poblaciones con y sin enfermedad, se consideraría que esta persona tiene muchas posibilidades de tener NPC, determinando, de este modo, un mayor nivel de cáncer (Lo *et al.* Cancer Res 1999; 59: 1188-1191). A continuación, puede ser remitido para realizar pruebas de confirmación adicionales. Por otro lado, si la prueba de ADN del EBV en plasma muestra una carga de ADN del EBV inferior al valor de corte para discriminar las poblaciones con y sin enfermedad, la probabilidad de NPC se puede considerar baja y se pueden considerar otros presuntos diagnósticos.

En términos de cribado terciario, un sujeto con NPC con tratamiento curativo por radioterapia puede someterse a la prueba de ADN del EBV en plasma para la identificación temprana de una posible recurrencia de NPC, en otras palabras, recaída (Lo *et al.* Cancer Res 1999; 59: 5452-5455; Lo *et al.* Cancer Res 2000; 60: 6878-6881). La probabilidad de recurrencia de NPC se consideraría alta si los niveles de ADN del EBV en plasma aumentan más allá de un valor inicial estable posterior al tratamiento de los propios valores del sujeto o más allá del valor de corte utilizado para identificar la población con recurrencia de NPC.

C. Otras pruebas de cribado y características preferidas

El ejemplo de la prueba de ADN del EBV en plasma para el control de NPC solamente se proporciona como una ilustración de cómo se realiza el cribado de cáncer o enfermedad. Sería ideal si se pudieran crear otras pruebas o modalidades de cribado eficaces para otros tipos de cáncer. En la actualidad, las pruebas de cribado para otros tipos de cáncer no existen o tienen perfiles de rendimiento deficientes. Por ejemplo, la alfafetoproteína (AFP, por sus siglas en inglés) sérica es un marcador utilizado para la evaluación del HCC. Sin embargo, la AFP sérica muestra poca sensibilidad y especificidad. En términos de sensibilidad, menos del 50 % de los HCC son positivos para AFP. En cuanto a la especificidad, otras afecciones inflamatorias del hígado podrían estar asociadas con AFP sérica elevada.

Por tanto, la AFP sérica generalmente no se utiliza como una herramienta de cribado primario para individuos asintomáticos de bajo riesgo. Si se utilizara, habría muchas identificaciones falsas negativas y falsas positivas de HCC. En cambio, puede aplicarse a individuos de alto riesgo con un alto índice de sospecha de padecer HCC. Por ejemplo, un portador crónico de HBV con una sombra hipoecoica que se muestra en la ecografía del hígado puede someterse a pruebas de AFP sérica. Si es positivo, sirve como una prueba adicional para respaldar el presunto diagnóstico de HCC. Además, si un caso confirmado de HCC muestra AFP sérica positiva o elevada, la AFP sérica puede usarse como una herramienta posterior al tratamiento para el cribado de la recurrencia del HCC.

Otros ejemplos de herramientas de cribado del cáncer que se han implementado como parte de varias iniciativas de salud pública incluyen, mamografía para el cribado del cáncer de mama, evaluación de sangre oculta en heces para cribado colorrectal, prueba de antígeno prostático específico en suero para el cribado de cáncer de próstata y evaluación de frotis cervical para el cribado de cáncer de cuello uterino. Se han implementado muchos programas de cribado porque generalmente se percibe que la identificación temprana de la enfermedad o la progresión de la

enfermedad se traduciría en beneficios para la salud, tales como una mayor supervivencia sin enfermedad, años de mayor calidad de vida y ahorro económico en el control de las enfermedades. Por ejemplo, si los cánceres pudieran identificarse en una etapa temprana o incluso en una etapa asintomática, podrían aplicarse modalidades de tratamiento más sencillas o con menos efectos secundarios. Por ejemplo, el tumor aún puede estar en un estadio en el que se podría considerar la extirpación quirúrgica.

En general, es preferible adoptar herramientas que no sean invasivas y con pocos efectos secundarios para el cribado. Las modalidades invasivas o aquellas con alto potencial de complicaciones se reservan para personas cuya probabilidad previa a la prueba de las enfermedades es lo suficientemente alta como para justificar enfrentar dichos riesgos durante la evaluación. Por ejemplo, la biopsia hepática se realiza en individuos con muy alto índice de sospecha de HCC, tales como portadores crónicos de HBV o pacientes con cirrosis hepática con una sombra hipoecoica que se muestra en la ecografía hepática.

En cuanto al perfil de rendimiento de las pruebas de cribado, es preferible tener pruebas que tengan un alto valor predictivo positivo (PPV) o un alto valor predictivo negativo (NPV). El perfil de rendimiento preferido real para cualquier indicación de cribado depende del propósito del cribado. Las pruebas con PPV alto generalmente se usan para confirmar o "aceptar" la clasificación de una enfermedad. Las pruebas con un NPV alto generalmente se usan para excluir o "descartar" una clasificación de enfermedad. Algunas pruebas tienen tanto PPV como NPV altos. Suelen ser pruebas que podrían ofrecer una clasificación definitiva, por ejemplo, biopsias de tejido seguidas de examen histológico.

D. Identificación de dianas específicas de cáncer en tejidos tumorales para cribado

Se podría intentar detectar la presencia de cualquier mutación paraneoplásica que se origine en el genoma de una célula cancerosa entre el ADN plasmático para la detección de cánceres. Como se demuestra en el ejemplo de ADN del EBV en el NPC anterior, la alta sensibilidad clínica o tasa de detección de NPC usando la prueba de ADN del EBV en plasma está relacionada con la capacidad de detectar aproximadamente 500 fragmentos de ADN plasmático paraneoplásicos por célula de NPC, por ejemplo, 300-600. Para potenciar adicionalmente la sensibilidad de la prueba o para realizar una o más pruebas de cribado, se puede necesitar ser capaz de detectar 300 o más fragmentos paraneoplásicos por célula cancerosa (por ejemplo, 400, 500, 600, 800 o 1000 o más).

Una forma posible de tener más de 500 dianas específicas de cáncer para NPC, así como para generalizar esto a otros cánceres y neoplasias malignas, sería el análisis de un conjunto de mutaciones de nucleótido único específicas del sujeto, o mutaciones que implican más de un nucleótido. Para identificar dicha información específica de sujeto, se puede realizar una secuenciación masiva en paralelo del tejido tumoral de un sujeto con cáncer. El ADN constitutivo del sujeto puede secuenciarse como una referencia para la identificación de las mutaciones en el tejido tumoral. El ADN constitutivo se puede obtener de cualquier célula no maligna del sujeto, por ejemplo, pero sin limitación, células sanguíneas y células bucales. Además de las mutaciones de un solo nucleótido, otros cambios genéticos y epigenéticos específicos del cáncer o paraneoplásicos (por ejemplo, aberraciones del número de copias y metilación aberrante) también se pueden usar como dianas para la detección de cáncer.

Dichos cambios pueden detectarse después en una muestra biológica del sujeto que puede contener ADN tumoral (por ejemplo, plasma o suero, los cuales contienen ADN sin células). En una realización, el objetivo es evaluar la carga mutacional del cuerpo a través del análisis de ADN plasmático. Para esta realización particular, la detección de mutaciones específicas de cáncer puede usarse para monitorizar el progreso del sujeto después del tratamiento porque sería necesario obtener los tejidos tumorales para la identificación de los cambios asociados al cáncer específicos para el sujeto. La detección de los cambios específicos del cáncer se puede realizar mediante PCR específica de alelo, secuenciación de amplicones usando secuenciación masiva en paralelo (por ejemplo, usando secuenciación profunda de amplicones etiquetados (Forshew *et al.* Sci Transl Med 2012; 4: 136ra68)), análisis de espectrometría de masas y análisis de micromatrices o secuenciación ultraprofunda, secuenciación exhaustiva y secuenciación de plantilla total como se describe en algunas realizaciones de la presente solicitud.

En una realización, la suma (ejemplo de una carga mutacional) de las cantidades de ADN plasmático que porta cada cambio específico de cáncer se puede determinar y utilizar para reflejar la cantidad de células cancerosas en el organismo. Esta última información sería útil para el pronóstico, monitorización y evaluación de la respuesta al tratamiento. En otras realizaciones, la carga mutacional se puede determinar como el producto o la media ponderada de las cantidades de las dianas específicas del cáncer.

En algunas realizaciones, la carga mutacional se puede determinar con poca o ninguna información sobre qué mutaciones pueden existir en la muestra, por ejemplo, durante un cribado inicial, como se describe a continuación. Además, se puede usar una proporción relativa de una mutación y el alelo de tipo silvestre en una posición para inferir la concentración fraccionaria de ADN derivado de tumor en la muestra de plasma.

III. EVALUACIÓN DE LA CARGA MUTACIONAL DEL ADN CIRCULANTE SIN CÉLULAS PARA EL CRIBADO DEL CÁNCER

Para identificar las mutaciones del cáncer y determinar la carga mutacional de un individuo, las realizaciones pueden analizar una muestra con ADN circulante sin células. Se sabe que los tumores, los cánceres y las neoplasias malignas liberan su contenido de ADN en la circulación (Bettegowda *et al.* Sci Transl Med 2014; 6: 224ra24). Por lo tanto, las mutaciones asociadas a tumores, cánceres y neoplasias malignas podrían detectarse en plasma y suero. Dichas mutaciones también podrían detectarse en otros fluidos corporales, tales como, pero sin limitación, orina, otros fluidos urogenitales, líquido de lavado cervical, secreción del pezón, saliva, líquido pleural, líquido ascítico y líquido cefalorraquídeo (Togneri *et al.* Eur J Hum Genet 2016; doi: 10.1038/ejhg.2015.281; De Mattos-Arruda *et al.* Nat Commun 2015; doi: 10.1038/ncomms9839; Liu *et al.* J Clin Pathol 2013; 66 :1065-1069.).

Las mutaciones podrían detectarse en estos fluidos corporales debido a la eliminación directa de células o ADN sin células en el fluido desde aquellos órganos que están en contacto directo con el fluido, por ejemplo, desde el tracto urinario (por ejemplo, del riñón o la vejiga) o genital (por ejemplo, de la próstata) hasta la orina, transrenalmente del plasma a la orina, del cerebro al líquido cefalorraquídeo, del páncreas al jugo pancreático, de la vesícula biliar a la bilis, de la orofaringe a la saliva, desde las células mamarias hasta el líquido de secreción del pezón, de los órganos abdominales al líquido ascítico o de los pulmones al líquido pleural. Además, las mutaciones podrían detectarse en los fluidos corporales porque proceden en parte de la filtración de plasma. Por lo tanto, el contenido en plasma, incluidas las mutaciones derivadas de tumores de otros órganos más distantes del lugar del fluido, podría detectarse en los fluidos corporales.

La detección de mutaciones entre los ácidos nucleicos sin células en plasma, suero y otros fluidos corporales es atractivo para la creación de pruebas de cribado del cáncer porque brindan acceso a los cambios genéticos y genómicos oncógenos de manera relativamente no invasiva y en lugar de la evaluación directa de una biopsia tumoral. Además, casi todas las formas de cambios genéticos y genómicos paraneoplásicos, cánceres o neoplasias malignas se han detectado entre la población de ácidos nucleicos sin células. En el presente documento se proporcionan ejemplos de cambios paraneoplásicos o cambios específicos del cáncer. Específico del cáncer generalmente se refiere a un cambio que proviene de una célula cancerosa, y paraneoplásico significa que el cambio puede provenir de una célula cancerosa, una lesión premaligna u otros tejidos debido a la proximidad anatómica, asociación fisiológica, asociación de desarrollo o una reacción a la presencia del cáncer.

Debido al acceso no invasivo al perfil genético y genómico oncógenos (determinado especialmente a partir de ácidos nucleicos sin células en plasma y suero), si se utiliza como prueba de cribado, el perfil oncógeno podría medirse de manera repetida, ya sea dentro de un intervalo más corto (por ejemplo, días o semanas) para "aceptar" o "descartar" la enfermedad o durante intervalos más largos, como bienalmente, anualmente, o semestralmente.

Las moléculas de ADN plasmático existen de manera natural en forma de fragmentos cortos de ADN (Yu *et al.* Proc Natl Acad Sci USA 2014; 111: 8583-8588). Por lo general, tienen < 200 pb de longitud y pueden fragmentarse en determinadas ubicaciones paraneoplásicas, como se analiza con más detalle a continuación. La mayoría de las moléculas de ADN en el plasma humano se originan a partir de células hematopoyéticas. Cuando una persona padece una neoplasia maligna no hematopoyética, especialmente durante los primeros estadios, el ADN derivado de tumor representa una fracción menor en plasma mezclado con un fondo de ADN hematopoyético que no deriva de tumor. La cantidad de ADN derivado de tumor en una muestra de plasma podría expresarse como una fracción del ADN total o el número de equivalentes genómicos o equivalentes celulares de células cancerosas. En el caso de una neoplasia maligna hematopoyética, se esperaría que la fracción de ADN asociado a neoplasia maligna en plasma fuera mayor que la de una neoplasia maligna no hematopoyética y podría detectarse usando las mismas realizaciones descritas en la presente solicitud.

En la presente solicitud, se describen protocolos que podrían aplicarse de forma genérica a la detección de cualquier tipo de cáncer siempre que el tumor aporte ADN al fluido corporal (Bettegowda *et al.* Sci Transl Med 2014; 6: 224ra24). El motivo es que las realizaciones descritas no dependen de la detección de biomarcadores que son típicos de un determinado tipo de cáncer. El esquema de clasificación utilizado para diferenciar individuos con y sin cáncer se basa en la evaluación de la carga mutacional que también podría aplicarse de forma genérica a efectos de detección de cualquier tipo de cáncer.

Para crear una prueba para el cribado de otros tipos de cáncer con alta sensibilidad y especificidad clínica, se necesitaría la capacidad de detectar una amplia gama y un gran número de mutaciones. Hay varias razones para justificar este requisito de prueba. A diferencia de la asociación de EBV con NPC, la mayoría de los otros cánceres no están asociados con un marcador genético no humano que podría distinguirse del ADN humano no canceroso con relativa facilidad. Por tanto, para crear una prueba de cribado para los cánceres no relacionados con el EBV, la prueba necesitaría detectar las otras variedades de cambios paraneoplásicos.

A. Requisitos de sensibilidad de la prueba (por ejemplo, *amplitud y profundidad)*

Basándose en los cálculos anteriores, para conseguir la misma sensibilidad que la prueba de ADN del EBV en plasma para la detección de NPC (Chan *et al.* Cancer 2013; 119: 1838-1844), la prueba necesitaría ser capaz de detectar preferentemente al menos ~500 copias de ADN plasmático portador de un cambio paraneoplásico para conseguir la detección del contenido de ADN equivalente de una célula tumoral en la circulación. Los datos de NPC se utilizan

como un sistema modelo para razonar a través de los principios para conseguir una prueba de cribado del cáncer clínicamente sensible y específica. Esto podría conseguirse detectando 500 copias de un cambio oncógeno, tal como en el caso de la prueba de ADN del EBV en plasma, o una copia de cada una de las 500 mutaciones oncógenas diferentes o una combinación, en concreto, múltiples copias de un conjunto de < 500 mutaciones. Debido a que los fragmentos de ADN plasmático generalmente tienen una longitud < 200 pb, se podría suponer que la detección de cualquier cambio paraneoplásico requeriría la detección de un fragmento de ADN plasmático portador de dicho cambio, denominado fragmento de ADN canceroso informativo.

Por lo tanto, algunos de los investigadores expertos en la materia han creado pruebas para detectar determinadas mutaciones en plasma como un medio para detectar cáncer. Por ejemplo, se ha utilizado la detección en plasma de mutaciones del *receptor del factor de crecimiento epidérmico* por reacción en cadena de la polimerasa digital (PCR) para la detección de cáncer de pulmón no microcítico (Yung *et al.* Clin Cancer Res 2009; 15: 2076-2084). Se han creado paneles que incluyen cientos de mutaciones paraneoplásicas diferentes, tal como en oncogenes y genes supresores de tumores, para la evaluación del ADN plasmático. Teóricamente, estas pruebas deberían haber alcanzado sensibilidades clínicas para la detección de esos otros tipos de cáncer que se acerquen al rendimiento como el de la prueba de ADN del EBV en plasma para NPC. Sin embargo, en la práctica, este no es el caso.

1. Amplitud

Ahora se aprecia que los cánceres son muy heterogéneos. El perfil de mutación varía mucho entre los cánceres de diferentes órganos, varía mucho entre diferentes sujetos con cánceres del mismo órgano o incluso entre diferentes focos tumorales en el mismo órgano del mismo sujeto (Gerlinger et al N Engl J Med 2012; 366: 883-892). Por tanto, cualquier mutación oncógena solamente es positiva en un pequeño subconjunto de cualquier sujeto con cáncer. Por ejemplo, la base de datos del Catalogue of Somatic Mutations in Cancer (COSMIC) documenta la variedad de mutaciones genéticas que se han detectado en los tejidos tumorales (cancer.sanger.ac.uk/cosmic).

La figura 1 muestra una tabla 100 de las 28 principales mutaciones más comúnmente identificadas entre los cánceres. Los datos muestran que la suma de las 28 principales mutaciones más prevalentes para los cánceres de cualquier órgano dado está lejos del 100 %. También es digno de mención que se podrían producir diferentes mutaciones con cada uno de los genes enumerados en la figura 1. Por tanto, si se evalúa la prevalencia de cualquier mutación específica entre los tumores, el número sería muy bajo. Debido a que la ubicación de las mutaciones del cáncer es tan variable e impredecible, para identificar 500 mutaciones diferentes en cualquier sujeto con cáncer, se podría considerar primero analizar una biopsia tumoral. Las mutaciones identificadas después se usarían para informar qué ensayos de ADN plasmático se usarían para la monitorización posterior. Sin embargo, la necesidad de una evaluación previa de una biopsia tumoral impediría aplicar la prueba de ADN plasmático para el cribado primario o asintomático.

Tal como se muestra en la figura 1, solamente una proporción de cada tipo de tumor puede mostrar cualquiera de las mutaciones principales. Los datos sugieren que una gran proporción de tumores no presenta ninguna de las principales mutaciones enumeradas en la base de datos COSMIC. En otras palabras, si se diseña una prueba de cribado del cáncer basada en la detección exclusiva de las mutaciones principales, muchos tumores no se detectarían debido a la ausencia de dichas mutaciones. Estos datos sugieren que la necesidad de detectar un gran número de mutaciones somáticas, como se demuestra por las realizaciones en la presente solicitud, es importante para realizar una prueba de cribado que sea genérica para diferentes tumores y, sin embargo, podría dar resultados positivos en una gran proporción de la población con cáncer.

Por lo tanto, para crear una prueba de ADN plasmático para la detección del cáncer o el cribado primario, se tendría que explorar a través de un espacio de búsqueda mucho más amplio dentro del genoma para recopilar suficientes mutaciones (por ejemplo, aberraciones en el número de copias y variantes de secuencia en relación con un genoma de referencia, tal como un genoma constitutivo o genomas precursores) u otros cambios específicos de cáncer o paraneoplásicos (por ejemplo, cambios de metilación) para formar la suma de 500 fragmentos de ADN plasmático específicos del cáncer por célula cancerosa. Observando los datos mostrados en la figura 1, asumiendo que la posibilidad de que se produzca una mutación paraneoplásica bien documentada en cualquier tumor es del 1 %, la prueba tendría que dirigirse a la detección de 50.000 supuestos sitios de mutación para tener al menos 500 mutaciones detectadas por tumor (basándose en la distribución de probabilidad de Poisson). Tendrían que analizarse 500.000 mutaciones o cambios paraneoplásicos supuestos para tener al menos 5.000 mutaciones o cambios paraneoplásicos representados para cualquier tumor. Por otro lado, si la probabilidad de que se produzcan mutaciones o cambios paraneoplásicos bien documentados en cualquier tumor es del 0,1 %, entonces, se necesitarían probar 50.000 mutaciones o cambios para tener al menos 50 mutaciones o cambios representados para cualquier tumor.

Por tanto, para maximizar la tasa de detección de cáncer, o la sensibilidad clínica, de la prueba de cribado del cáncer, la prueba necesitaría conseguir un estudio amplio de fragmentos de ADN plasmático en una muestra para identificar suficientes fragmentos portadores de cualquier tipo de mutación o cambio paraneoplásico. La amplitud del estudio podría conseguirse con el uso de estrategias de todo el genoma o estrategias dirigidas que cubran una gran fracción del genoma, por ejemplo, suficiente para cubrir al menos 50.000 dianas.

2. Profundidad

La profundidad del estudio también importa. Dependiendo del número de mutaciones detectadas por tumor, sería necesario detectar múltiples fragmentos de ADN plasmático que tuvieran esa mutación para alcanzar un umbral específico, por ejemplo, 500 fragmentos de ADN canceroso informativos para cada genoma equivalente de célula cancerosa. Por ejemplo, si solamente se identifica una mutación en un tumor en particular, entonces se necesitarían 500 fragmentos de ADN plasmático que cubran esa mutación. Por otro lado, si 50 mutaciones diferentes están presentes en el tumor, en promedio, se necesitaría detectar al menos 10 fragmentos de ADN canceroso informativos que cubran cada una de esas 50 mutaciones.

El ADN tumoral normalmente representa una población menor de ADN en el plasma. Asimismo, algunos cambios paraneoplásicos son de naturaleza heterocigota (es decir, con un cambio por genoma diploide). Por lo tanto, para detectar 10 copias de fragmentos de ADN canceroso informativos (es decir, fragmentos de ADN plasmático portadores de al menos un cambio paraneoplásico) por locus, sería necesario analizar al menos 100 moléculas del locus en una muestra de plasma con una fracción de ADN tumoral del 20 %. Por lo tanto, la capacidad de detectar múltiples fragmentos de ADN plasmático que cubren cualquier sitio de mutación individual depende de la profundidad con la que se examine la muestra de plasma. No obstante, solamente hay un número finito de genomas de células cancerosas en la muestra de plasma, lo que afecta tanto a la profundidad como a la amplitud requeridas del análisis de ADN plasmático.

Para ilustrar la detección de cánceres tempranos, supóngase que se pretende crear una prueba o protocolo que pueda detectar una fracción tumoral del 1 % en una muestra. Dado que normalmente hay 1.000 genomas equivalentes de ADN en cada mililitro de plasma, habría 10 equivalentes de células cancerosas de ADN en una muestra de un mililitro con una fracción de ADN tumoral del 1 %. Esto significa que incluso si se o pudiera detectar cada fragmento de ADN específico de cáncer en la muestra, solamente habría un máximo de 10 equivalentes del genoma de cualquier cambio paraneoplásico que estaría disponible para la detección. En consecuencia, incluso si se tiene conocimiento previo de que una mutación particular está presente en un tumor, su detección dirigida solo proporcionaría una señal de 10 equivalentes de genoma en el mejor de los casos, que puede carecer de la sensibilidad analítica para la detección fuerte de un cáncer a una concentración fraccionada del 1 %. Si la mutación a detectar es heterocigota, solo habría 5 fragmentos de ADN plasmático que mostrarían esta mutación.

En el mejor de los casos con una fracción de ADN tumoral del 1 %, se necesitaría cubrir la profundidad del análisis en este sitio de mutación al menos 1.000 veces para poder detectar los 10 equivalentes del genoma del ADN plasmático con la mutación. En esta situación, la amplitud del análisis necesitaría compensar el número relativamente bajo de copias detectadas por sitio de mutación. Es poco probable que la detección selectiva de un puñado o incluso cientos de sitios de mutación pueda conseguir la sensibilidad requerida para una prueba de cribado para detectar cáncer temprano.

3. Otros problemas

Además, en los análisis habituales, el rendimiento de detección de cualquier ensayo está lejos del mejor de los casos. Por ejemplo, podría haber pérdida o reducción en las plantillas de ADN plasmático y fragmentos de ADN del cáncer informativos durante las etapas de procesamiento de la muestra, las etapas de preparación de la biblioteca de secuenciación de ADN y el proceso de hibridación de captura de dianas basado en sondas. Algunas etapas pueden introducir sesgos en las proporciones relativas entre diferentes mutaciones y entre el ADN derivado del cáncer y el que no deriva del cáncer. Por ejemplo, la amplificación por PCR de bibliotecas de secuenciación diana, las bibliotecas de secuenciación de ADN genómico y la secuenciación de amplicones podrían introducir sesgos de CG y crear duplicados de PCR. Para la secuenciación masiva de ADN en paralelo, los errores en la identificación de un fragmento secuenciado pueden deberse a errores de secuenciación surgidos durante la amplificación por PCR o durante la secuenciación, durante la asignación de bases o debido a errores de alineación. Por último, el mecanismo de detección de señales de la plataforma de análisis puede tener un límite de detección antes de que se pueda proporcionar una lectura positiva segura para la detección de una mutación (por ejemplo, se pueden necesitar 5 fragmentos mutantes para una señal detectable). Todos estos factores hacen que, en la práctica, los requisitos de amplitud y profundidad del análisis del ADN plasmático pueden tener que ser incluso mayores que los escenarios ideales teóricos analizados.

Esencialmente, el análisis hasta el momento sugiere que los requisitos de sensibilidad de la prueba de cribado del cáncer están alcanzando las limitaciones de lo que las plataformas de análisis molecular podrían conseguir en la práctica. Biológicamente, se ha informado que el número de mutaciones somáticas albergadas por un tumor maligno varía entre aproximadamente 1.000 y varias decenas de miles (Lawrence *et al.* Nature 2013; 499: 214-218). Basándose en los datos de los presentes inventores, dependiendo de la concentración fraccionaria de ADN tumoral en la muestra de plasma, se podrían tener suficientes fragmentos informativos de ADN del cáncer en la muestra finita de plasma (normalmente, se obtendrían < 10 mililitros de plasma por extracción de sangre) para conseguir una detección temprana no invasiva del cáncer.

Por tanto, para alcanzar prácticamente los requisitos de sensibilidad de la prueba de cribado del cáncer, sería necesario maximizar el contenido de información sobre el cáncer que podría obtenerse en cada muestra de plasma. En la presente solicitud, se describen procesos que pueden conseguir la amplitud y profundidad eficaces necesarias

para alcanzar los requisitos de sensibilidad de la prueba de cribado del cáncer. En diversas realizaciones, se realiza secuenciación ultra profunda y amplia, exhaustiva o secuenciación de plantilla total. Se puede realizar una secuenciación masiva en paralelo sin PCR para aumentar la rentabilidad de la secuenciación ultraprofunda y amplia, exhaustiva o secuenciación de plantilla total. La secuenciación ultraprofunda y amplia, exhaustiva o la secuenciación de plantilla total se puede conseguir a través de la secuenciación de una sola molécula.

Algunas realizaciones pueden aumentar el número de fragmentos de ADN canceroso informativos accesibles mediante la detección combinada de una variedad de cambios específicos del cáncer o paraneoplásicos, por ejemplo, mutaciones de un único nucleótido, en combinación con firmas de metilación del ADN específicas para el cáncer o paraneoplásicas (por ejemplo, la localización de la 5-metilcitosina y la hidroximetilación), moléculas cortas de ADN plasmático específicas del cáncer o paraneoplásicas, marcas de modificación de las histonas específicas del cáncer o paraneoplásicas y ubicaciones finales del ADN plasmático específicas del cáncer o paraneoplásicas. Ciertos cambios específicos de cáncer o paraneoplásicos pueden utilizarse como criterios de filtrado en la identificación de mutaciones.

B. Requisitos de especificidad (por ejemplo, *criterios de filtrado)*

Como se describe anteriormente, es deseable detectar tantos fragmentos de ADN canceroso informativos como sea posible. Pero, puede ser difícil detectar con precisión tales fragmentos de ADN del cáncer informativos dado el nivel de ruido (por ejemplo, errores de diversas fuentes) presente en las técnicas de secuenciación actuales.

1. Especificidad de las mutaciones identificadas

Para conseguir un PPV alto o un NPV alto, la prueba de cribado del cáncer necesitaría mostrar un alto perfil de especificidad. Se podría conseguir una alta especificidad en varios niveles. La especificidad de las mutaciones y cualquier cambio paraneoplásico que se detecte necesitarían ser lo más específicos posible para el cáncer. Esto podría conseguirse mediante, pero sin limitación, puntuación de una firma genética o genómica como positiva solamente cuando existe un alto nivel de confianza de que es paraneoplásica. Esto podría conseguirse mediante la inclusión de firmas que se han informado previamente en otros tipos de cáncer. Por ejemplo, se puede centrar la atención particularmente en las firmas que prevalecen en el tipo de cáncer al que el individuo está predispuesto, según su perfil demográfico. O, se puede prestar especial atención a las firmas mutacionales que están asociadas con la exposición mutagénica a la que ha estado expuesto un sujeto (Alexandrov *et al.* Nature 2013; 500: 415-421). Esto también podría conseguirse minimizando el número de errores de secuenciación y alineación que pueden identificarse erróneamente como una mutación. Esto se puede conseguir mediante comparación con el perfil genómico de un grupo de controles sanos y/o se puede conseguir mediante comparación con el propio ADN constitutivo de la persona.

Podrían aplicarse estos criterios de filtrado para evaluar la probabilidad de que un fragmento de ADN plasmático proceda del tumor y, por tanto, pueda considerarse un fragmento de ADN canceroso informativo. Cada uno de los criterios de filtrado podría utilizarse individualmente, independientemente, colectivamente con igual ponderación o diferentes ponderaciones, o en serie en un orden especificado, o condicionalmente en función de los resultados de las etapas de filtrado anteriores. Para su uso condicional, puede utilizarse una estrategia bayesiana, así como una estrategia basada en árboles de clasificación o decisión. Un uso individual significa un criterio cualquiera. Un uso independiente puede implicar más de un criterio de filtrado, pero cada criterio de filtrado no depende de la aplicación de otro criterio de filtrado (por ejemplo, se puede aplicar en paralelo), en contraste con una aplicación en serie en órdenes específicos. Como ejemplo de uso colectivo utilizando ponderaciones, pueden utilizarse técnicas de inteligencia artificial. Por ejemplo, el aprendizaje supervisado puede utilizar cargas mutacionales medidas de muestras con clasificaciones conocidas para entrenar cualquier modelo. Los datos de secuenciación de un gran número de individuos (por ejemplo, cientos, miles o millones) pueden utilizarse para entrenar los modelos. Indicado de un modo más sencillo, dichas muestras conocidas pueden utilizarse para determinar los valores umbral de una o más puntuaciones determinadas a partir de los criterios de filtrado para determinar si una mutación es válida o no.

En una realización, si un fragmento de ADN plasmático cumple algunos o la totalidad de los criterios, se puede considerar que es un fragmento de ADN canceroso informativo, mientras que los otros que no cumplen algunos o la totalidad pueden considerarse un fragmento de ADN plasmático no informativo. En otra realización, a cada fragmento de ADN plasmático se le podría dar una ponderación del carácter informativo de que sea un fragmento de ADN canceroso informativo en función del grado de cumplimiento de la lista de criterios. Cuanto mayor sea la confianza de que el fragmento deriva de un tumor, mayor será la ponderación. En una realización, la ponderación se puede ajustar en función del perfil clínico del sujeto de prueba (por ejemplo, sexo, etnia, factor de riesgo para el cáncer, tal como el tabaquismo o el estado de hepatitis, etc.).

Un fragmento de ADN podría recibir una mayor ponderación del carácter informativo o de la especificidad del cáncer si muestra más de un cambio específico de cáncer. Por ejemplo, muchos cánceres están globalmente hipometilados, especialmente en las regiones no promotoras. Se ha demostrado que el ADN canceroso es más corto que el ADN no canceroso en el plasma. Los fragmentos de ADN plasmático derivados de tumor tienden a fragmentarse en algunos lugares específicos. Por tanto, un fragmento de ADN plasmático de tamaño corto (por ejemplo, <150 pb) (Jiang *et al.* Proc Natl Acad Sci USA 2015; 112: E1317-1325), con uno o ambos extremos que se encuentran en ubicaciones finales

paraneoplásicas, muestra una mutación de un solo nucleótido, y se localiza en una región no promotora, y tiene un sitio de CpG hipometilado se consideraría con mayor probabilidades de que sea paraneoplásico. La detección del ADN hipometilado podría conseguirse con el uso de la conversión del ADN con bisulfito o la secuenciación directa de una sola molécula que podría distinguir la metilcitosina de la no metilcitosina. En la presente solicitud, los presentes inventores describen procesos, protocolos y etapas para aumentar la especificidad en la identificación de fragmentos de ADN canceroso informativos. Por ejemplo, se pueden utilizar uno o varios criterios de filtrado para aumentar la especificidad.

2. Especificidad de la carga mutacional

En otro nivel, la especificidad de la prueba de cribado del cáncer podría conseguirse evaluando si la cantidad (por ejemplo, el número) de cambios paraneoplásicos detectables en el plasma de pacientes con cáncer refleja una carga mutacional acorde con la esperada para el cáncer. En una realización, se podría comparar la carga mutacional en el plasma con la carga mutacional medida en el ADN constitutivo, por ejemplo, cuando la carga mutacional se determina con respecto a un genoma de referencia. En otras realizaciones, se podría comparar la carga mutacional en el plasma con la observada en el plasma del sujeto en un momento diferente, o de un paciente con cáncer con pronóstico (bueno o malo) o estadio del cáncer conocidos, o de una población sana sin cáncer. La población de referencia puede coincidir en edad, sexo o etnia, ya que se ha informado que la carga mutacional en el organismo o en los tejidos aumenta con la edad incluso en personas que no muestran tener cáncer (Slebos *et al.* Br J Cancer 2008; 98: 619-626). En la presente solicitud, se describe cuánto de amplio y profundo sería necesario realizar el análisis de ADN plasmático para capturar una carga mutacional adecuada para potenciar la diferenciación entre sujetos con cáncer de la población sana. Por lo tanto, no es necesario detectar todos los fragmentos de ADN en la muestra de plasma para conseguir la detección del cáncer, por ejemplo, si una muestra tiene suficiente información mutacional.

Si una carga mutacional observada sugiere cáncer podría, en una realización, estar basada en intervalos de referencia específicos del cáncer. Se ha informado que los cánceres de diferentes órganos tienden a albergar un intervalo esperado de carga de mutaciones. El número puede variar de 1.000 a varias decenas de miles (Lawrence *et al.* Nature 2013; 499: 214-218). Por lo tanto, si la prueba de cribado del cáncer de ADN plasmático muestra evidencia de que la carga mutacional de una persona se acerca a números en el intervalo de cualquier grupo de cáncer, se podría hacer una clasificación de alto riesgo de cáncer (figuras 44, 45A-45C y 46A-46C de la sección VIII). En otra realización, se podría hacer una clasificación para el cáncer si la carga mutacional en el plasma de una persona es significativamente más alta que un intervalo de referencia establecido a partir de una población sana sin cáncer.

La evidencia de una carga mutacional significativamente mayor podría basarse en distribuciones estadísticas, por ejemplo, más de tres desviaciones estándar de la media de los datos de referencia de control, o un número de múltiplos de la mediana de los datos de referencia de control, o mayor que un percentil particular (por ejemplo, el 99º percentil) de los datos de referencia de control o al menos 1 o 2 o 3 órdenes de magnitud mayor que la media, mediana, o 99ºpercentil de los datos de referencia de control. Los expertos en la materia podrían identificar varios medios estadísticos para identificar una carga mutacional significativamente aumentada desde el punto de vista estadístico. En otra realización, la clasificación podría tener en cuenta las variables que se ha demostrado que afectan los perfiles de sensibilidad y especificidad de la prueba de cribado del cáncer, tal como la fracción de ADN tumoral medida, supuesta o inferida de la muestra, profundidad de secuenciación, amplitud de secuenciación y tasas de error de secuenciación (figuras 44, 45A-45C y 46A-46C de la sección VIII).

La carga mutacional se puede determinar de varias maneras. La carga mutacional podría expresarse como el número de mutaciones detectadas. El número de mutaciones podría normalizarse a la cantidad de datos de secuenciación obtenidos, por ejemplo, expresada como un porcentaje de los nucleótidos secuenciados o una densidad de mutaciones detectadas por la cantidad de secuenciación realizada. El número de mutaciones también podría normalizarse al tamaño del genoma humano, por ejemplo,expresado como una proporción del genoma o una densidad por región dentro del genoma. El número de mutaciones podría informarse para cada ocasión cuando se realice una evaluación de la carga de mutaciones o podría integrarse a lo largo del tiempo, por ejemplo,el cambio absoluto, cambio porcentual o factor de cambio en comparación con una evaluación anterior. La carga mutacional podría normalizarse a la cantidad de la muestra (por ejemplo, volumen de plasma) analizada, a la cantidad de ADN obtenido de la muestra, o la cantidad de ADN analizable o secuenciable. En una realización, la carga mutacional se puede normalizar a un parámetro biométrico del sujeto evaluado, por ejemplo, peso, estatura o índice de masa corporal.

En la presente solicitud, se describe cuánto de amplio y profundo sería necesario que fuera el análisis de ADN plasmático para capturar una carga mutacional adecuada para potenciar la diferenciación entre un sujeto con cáncer de una población sin cáncer, por tanto, para conseguir una evaluación eficaz de la carga mutacional.

IV. SECUENCIACIÓN ULTRAPROFUNDA Y AMPLIA

Como se explica en detalle anteriormente, existe la necesidad de una secuenciación ultraprofunda y amplia para conseguir los perfiles de rendimiento necesarios para la prueba de cribado del cáncer o la identificación eficaz de mutaciones *de novo* fetales. En la presente solicitud, se muestra una serie de realizaciones para conseguir una secuenciación ultra profunda y amplia. Tales realizaciones incluyen, pero sin limitación, secuenciación exhaustiva,

secuenciación total de plantillas, secuenciación sin PCR, secuenciación de una sola molécula (un tipo de secuenciación sin PCR) y secuenciación dirigida. Se puede utilizar una combinación de estrategias para conseguir la profundidad y amplitud necesarias. Dicha combinación se puede usar para un programa de cribado como un todo, o para el cribado de un individuo o grupos de individuos en particular.

A efectos del cribado del cáncer, para detectar las mutaciones paraneoplásicas a partir de la secuenciación del ADN plasmático, la profundidad de la secuenciación afectaría a la capacidad de diferenciar las mutaciones verdaderas del cáncer y los falsos positivos debido a errores de secuenciación. Se requeriría una mayor profundidad de secuenciación cuando la fracción de ADN tumoral en el plasma fuera menor (figura 4B). Usando un análisis de valor de corte dinámico (descrito en una sección posterior), cuando la fracción de ADN tumoral es del 2 %, una profundidad de secuenciación de 200 veces podría ser capaz de detectar el 5,3 % de las mutaciones paraneoplásicas. El número de mutaciones detectadas sería mayor que el número esperado de falsos positivos, suponiendo que los errores de secuenciación aleatorios se produzcan con una frecuencia del 0,3 %. La porción del genoma a buscar dependería del número esperado de mutaciones en el tejido tumoral.

La porción del genoma que se va a buscar tendría que ser lo suficientemente grande como para obtener un número suficiente de mutaciones a detectar. Este parámetro de amplitud dependería del límite de detección inferior deseado de la fracción de ADN tumoral y del tipo de cáncer que se desea cribar. Por ejemplo, en melanoma, la frecuencia media de mutación es de aproximadamente 10 por 1 Mb. En otras palabras, habría aproximadamente 30.000 mutaciones en un genoma. Suponiendo que la fracción de ADN tumoral es del 2 % y se busca 1/10 del genoma, se espera que se detecten aproximadamente 159 mutaciones mediante la secuenciación del ADN plasmático a 200x. Por otro lado, si el tumor rabdoide es la diana a cribar, la frecuencia media de mutaciones es de solamente 0,2 por 1 Mb. Por lo tanto, la búsqueda de 1/10 del genoma arrojaría aproximadamente 3 mutaciones cancerosas cuando la fracción de ADN tumoral es del 2 %. Este número no es suficiente para diferenciarlo de los errores de secuenciación.

La figura 2 es una tabla 200 que muestra un número esperado de mutaciones a detectar para diferentes fracciones de ADN tumoral, profundidades de secuenciación, número de mutaciones por genoma y la fracción del genoma buscado. El número esperado de falsos positivos es <10 para todo el genoma en cada caso según un análisis de valor de corte dinámico (u otro análisis de filtrado adecuado) y una tasa de error de secuenciación del 0,3 %. Por tanto, cuando el número de mutaciones detectables (por ejemplo, en función de la profundidad y la amplitud) es superior a 10, las realizaciones serían útiles para diferenciar las mutaciones cancerosas reales de los falsos positivos.

Como se muestra en los datos de la tabla 200, la porción del genoma a analizar dependería de la fracción tumoral esperada y de la frecuencia de mutaciones somáticas en el tumor. Con el análisis del 5 % del genoma, el número de mutaciones sería mucho mayor que el número de falsos positivos cuando la fracción tumoral es del 10 %, la frecuencia de mutaciones es de 10 por Mb y la profundidad de secuenciación es de 200 veces. Usando el análisis de simulación, se dedujo que el número de mutaciones detectadas sería suficiente para discriminar los errores de secuenciación aleatorios incluso cuando se busca en el 0,1 % del genoma. Para otras frecuencias de mutaciones y profundidades de secuenciación, es posible que sea necesario analizar porciones mayores del genoma, por ejemplo, se puede analizar el 1 %, 5 %, 10 % y el 20 % del genoma alineando las lecturas de secuencias con un genoma de referencia.

A efectos del cribado del cáncer, no es necesario identificar el 100 % de las mutaciones paraneoplásicas. En una realización, solamente se tiene que demostrar que un individuo en particular tiene un mayor número de mutaciones detectadas en plasma (u otra muestra biológica) que las de una población de control de referencia sin cáncer. Sin embargo, para que esta estrategia sea altamente precisa, la proporción de mutaciones verdaderas detectadas mediante un protocolo de evaluación de la carga mutacional debería ser lo más alta posible (o la proporción de falsos positivos debe ser lo más baja posible), de manera que el elevado número de variantes detectadas por la evaluación sea un reflejo de la presencia de cáncer. Si esto no se puede conseguir, el alto número de supuestas mutaciones detectadas en una muestra puede ser simplemente un reflejo de un alto número de variantes falsas positivas y, por lo tanto, no permitiría la discriminación de un sujeto con cáncer y aquellos sin cáncer. Por tanto, las realizaciones en la presente solicitud describen cómo reducir la detección de falsos positivos y cómo aumentar la detección de mutaciones verdaderas para conseguir una evaluación eficaz de la carga mutacional.

La secuenciación ultraprofunda y amplia se puede conseguir mediante una secuenciación exhaustiva u otros medios, por ejemplo, secuenciación ligera (no exhaustiva) de múltiples paneles de secuenciación dirigida. La secuenciación ligera se puede utilizar para minimizar los duplicados de PCR para que se pueda obtener la profundidad requerida. Se pueden utilizar múltiples paneles de secuenciación dirigida para proporcionar una amplia cobertura en todo el genoma.

A. Secuenciación exhaustiva y secuenciación de plantilla total

Para producir una prueba de cribado de cáncer eficaz para la identificación temprana del cáncer y la identificación del cáncer en estadios tempranos, lo ideal sería obtener la mayor cantidad posible de información pertinente sobre el cáncer de la muestra de plasma. Hay una serie de problemas que dificultan la capacidad de obtener información pertinente sobre el cáncer de la muestra de plasma: (1) la muestra a analizar tiene un volumen finito; (2) la fracción tumoral en una muestra biológica particular puede ser baja durante el cáncer temprano; (3) la cantidad total de mutaciones somáticas por tumor disponibles para la detección es del orden de 1.000 a 10.000; y (4) las etapas

analíticas y los procesos técnicos darían lugar a una pérdida en el contenido de la información. Por tanto, se debe intentar minimizar la pérdida de cualquier contenido de información relacionado con el cáncer en la muestra de plasma que sea susceptible de detección.

Debido a las limitaciones en las etapas de preparación de muestras, etapas de preparación de la biblioteca de secuenciación, secuenciación, asignación de bases y alineación, no todas las moléculas de ADN plasmático en una muestra serían analizables o secuenciables. La secuenciación exhaustiva se refiere a los procedimientos implementados para maximizar la capacidad de transformar la mayor cantidad posible de moléculas de ADN informativas (por ejemplo, aquellas con mutaciones) en una muestra finita en moléculas analizables o secuenciables. Se podrían adoptar varios procesos para conseguir una secuenciación exhaustiva.

Lo que constituye la población de ADN informativo puede variar según lo que se esté analizando. Para las pruebas de cáncer, serían los fragmentos de ADN canceroso plasmático informativos. Para las pruebas prenatales, serían las moléculas de ADN procedentes del feto en el plasma materno. Para la monitorización del trasplante, serían las moléculas procedentes del donante en el plasma del receptor del trasplante. Para detectar otras enfermedades, serían aquellas moléculas de ADN plasmático derivadas del órgano o tejido o células con la patología. Para detectar un proceso biológico anormal que implica mutaciones, serían aquellas moléculas de ADN plasmático procedentes del órgano o tejido o células implicados en el proceso, por ejemplo, el cerebro en el envejecimiento. Los ejemplos de dichos procesos biológicos pueden incluir el envejecimiento, predisposición genética a las mutaciones (por ejemplo, xeroderma pigmentoso), influencias mutagénicas del medio ambiente (por ejemplo, radiación o exposición a los rayos UV), o toxinas y efectos de fármacos (por ejemplo, agentes citotóxicos). En cuanto al tipo de muestra, para la prueba de ADN en una muestra de orina, podrían ser moléculas de ADN del cáncer que han pasado por vía transrenal desde el sistema circulatorio (por ejemplo, desde el plasma) a la muestra de orina (Botezatu *et al.* Clin Chem 2000; 46: 1078-1084). Para otros tipos de cáncer, podrían ser moléculas de ADN canceroso que han pasado de un cáncer del tracto urogenital (por ejemplo, de la vejiga o los riñones) a la muestra de orina.

Para ser lo más exhaustivo posible, se podría adoptar uno cualquiera, todos o una combinación de procesos: (1) usar protocolos de preparación de ADN que reduzcan la pérdida de ADN o tengan una alta eficacia de secuenciación o eficacia de conversión de bibliotecas de ADN; (2) evitar el problema de los duplicados de PCR mediante el uso de protocolos de preparación de ADN sin PCR; (3) reducir los errores de secuenciación mediante el uso de protocolos de preparación de ADN sin PCR; (4) reducir los errores de alineación mediante la adopción de algoritmos de alineación eficaces, por ejemplo, una estrategia de realineación. Al adoptar algunas o todas estas medidas, se puede reducir el grado de pérdida en el contenido de información del ADN plasmático, así como el desperdicio de recursos de secuenciación, de manera que la secuenciación ultra profunda y amplia se pueda conseguir de manera más rentable.

Después de aplicar dichas medidas de intento de secuenciación exhaustiva, la cantidad de señal pertinente para el cáncer o de fragmentos de ADN canceroso informativos puede volverse tan eficaz que la información de tan solo una proporción de la muestra ya sea adecuada para alcanzar la clasificación para "aceptar" o "descartar" el cáncer. Por ejemplo, como se muestra en un ejemplo posterior de la comparación de la carga mutacional entre una muestra de plasma de un paciente con HCC y una muestra de plasma de sangre del cordón umbilical, los datos a una profundidad de 75x ya eran adecuados para distinguir claramente el caso de HCC del plasma de sangre del cordón umbilical de un recién nacido sin cáncer. Se generaron 220x de datos para la muestra de plasma de HCC. Pero 75x de datos ya era suficiente porque la cantidad de fragmentos de ADN canceroso informativos detectados utilizando los procedimientos para el intento de secuenciación exhaustiva ya era adecuada y de calidad adecuada para la clasificación positiva de cáncer.

Si realmente se consumen completamente las moléculas de ADN plasmático secuenciables de la muestra finita, este acto podría denominarse "secuenciación total de plantillas". Esto se refiere a un espectro de secuenciación exhaustiva. Por ejemplo, todas las bibliotecas de ADN plasmático se secuenciaron del caso de HCC para alcanzar la profundidad de 220x.

También se puede realizar una secuenciación exhaustiva utilizando un secuenciador de una sola molécula (Cheng *et al.* Clin Chem 2015; 61: 1305-1306). Ejemplos de dichos secuenciadores de ADN de una sola molécula, incluyen, pero sin limitación, un secuenciador fabricado por Pacific Biosciences utilizando la tecnología de secuenciación de ADN en tiempo real de una sola molécula (www.pacificbiosciences.com/) y un secuenciador de nanoporos (por ejemplo, uno fabricado por Oxford Nanopore (www.nanoporetech.com/)). Varias de estas plataformas de secuenciación de una sola molécula permitirían obtener directamente información epigenética de la molécula secuenciada (por ejemplo, patrones de metilación del ADN) (Ahmed *et al.* J Phys Chem Lett 2014; 5: 2601-2607). Como se han descrito aberraciones epigenéticas en cáncer, tener dicha información epigenética potenciaría aún más el cribado, detección, monitorización y pronóstico del cáncer. Por ejemplo, las técnicas de filtrado basadas en la metilación se describen a continuación.

Otra realización mediante la cual se puede obtener información epigenética a partir de los datos de secuenciación es realizar la conversión con bisulfito del ADN plantilla, seguido de la secuenciación del ADN. La conversión con bisulfito es un proceso mediante el cual una citosina metilada permanecería sin cambios, mientras que una citosina sin metilar se convertiría en uracilo. Este último se leería como un resto T durante la secuenciación del ADN. La secuenciación con bisulfito, una forma de secuenciación que reconoce la metilación, se puede realizar después en una biblioteca de

secuenciación para el ADN plantilla convertido con bisulfito. Después, la alineación se puede realizar usando estrategias conocidas por los expertos en la materia, por ejemplo, el método de Jiang *et al.* (PLoS One 2014; 9: e100360).

Cuando se utiliza la secuenciación del ADN sin células para el cáncer, se pueden combinar muchos tipos de información molecular a partir de los resultados de secuenciación, concretamente, secuencias genómicas víricas en plasma (para cáncer asociado con infecciones víricas, por ejemplo, EBV para NPC), variantes de un solo nucleótido oncógenas, aberraciones en el número de copias e información epigenética (por ejemplo, metilación del ADN (incluido el perfil de 5-metilcitosina y la hidroximetilación), cambios de acetilación/metilación de histonas, etc.). Dicha combinación de información puede hacer que el análisis sea más sensible, específico y clínicamente pertinente.

B. Protocolo sin PCR

Para la detección de cualquier cambio paraneoplásico en el plasma (u otro tipo de muestra que contenga ADN sin células) de un sujeto analizado, en teoría, la probabilidad de detectar dicho un cambio debería aumentar con el aumento del número de moléculas de ADN analizadas. En el presente caso se utilizó un ejemplo hipotético para ilustrar este principio. Supóngase que el 20 % del ADN plasmático en un sujeto con cáncer procede del tumor y que el tumor tiene una mutación puntual en una posición de nucleótido particular. La mutación se produce solamente en uno de los dos cromosomas homólogos. Como resultado, el 10 % del ADN plasmático que cubre esta posición particular de nucleótido portaría esta mutación. Si se analiza una molécula de ADN que cubre esta posición de nucleótido, la probabilidad de detectar la mutación sería del 10 %. Si se analizan diez moléculas de ADN plasmático que cubren este cambio de nucleótido, la probabilidad de detectar la mutación aumentaría al 65,1 % (probabilidad = $1 - 0{,}9^{10}$). Si se aumentara adicionalmente el número de moléculas que se analizan a 100, la probabilidad de detectar la mutación aumentaría al 99,99 %.

Este principio matemático se puede aplicar para predecir la probabilidad de detectar mutaciones paraneoplásicas cuando se usa la secuenciación masiva en paralelo para el análisis del ADN plasmático de sujetos con cáncer. Sin embargo, en las plataformas habituales de secuenciación masiva en paralelo utilizadas para la secuenciación de plasma (por ejemplo, el sistema de secuenciación HiSeq2000 de Illumina con el kit de preparación de bibliotecas TruSeq), se realizarían amplificaciones por PCR en el ADN plantilla antes de la secuenciación.

La amplificación se refiere a procesos que dan como resultado aumentos (más de 1 vez) en la cantidad de ADN plantilla en comparación con el ácido nucleico de entrada original. En la presente solicitud, los procesos de amplificación son etapas realizadas durante la preparación de la biblioteca antes de la etapa de análisis de la plantilla de ADN, por ejemplo, secuenciación. Con amplificación, la cantidad de plantilla de ADN disponible para el análisis aumentaría. En una realización, la amplificación se puede realizar usando PCR, que implica cambios cíclicos de temperatura. En otra realización, la amplificación se puede realizar utilizando procesos isotérmicos. En algunas realizaciones se muestra que el ADN plantilla amplificado disminuye la eficacia para conseguir la evaluación de la carga mutacional. Las etapas de expansión clonal que se producen durante la etapa de análisis, por ejemplo, amplificación de puentes durante la secuenciación por síntesis, no se consideran como una amplificación porque no da como resultado lecturas de secuencias adicionales o salida de secuencia.

Cuando se utiliza la PCR, la profundidad de secuenciación (es decir, el número de lecturas de secuencias que cubren un nucleótido en particular) no refleja directamente cuántas moléculas de ADN plasmático que cubren ese nucleótido en particular se analizan. Esto se debe a que una molécula de ADN plasmático puede generar múltiples réplicas durante el proceso de la PCR, y pueden originarse múltiples lecturas de secuencias a partir de una sola molécula de ADN plasmático. Este problema de duplicación sería más importante con i) un mayor número de ciclos de PCR para amplificar la biblioteca de secuenciación; ii) una mayor profundidad de secuenciación, y iii) un menor número de moléculas de ADN en la muestra de plasma original (por ejemplo, un menor volumen de plasma).

Además, la etapa de la PCR introduce más errores (Kinde *et al.* Proc Natl Acad Sci USA 2011; 108: 9530-9535) dado que la fidelidad de una ADN polimerasa no es del 100 %, y ocasionalmente, podría incorporarse un nucleótido erróneo en la cadena hija producida mediante la PCR. Si este error de PCR se produce durante los primeros ciclos de la PCR, se generarían clones de moléculas hijas que mostrarían el mismo error. La concentración fraccionaria de la base errónea puede alcanzar una proporción tan alta entre otras moléculas de ADN del mismo locus que el error se interpretaría de manera errónea como una mutación derivada de un tumor o derivada de un feto.

En el presente caso, los presentes inventores consideran que el protocolo sin PCR para la secuenciación masiva en paralelo permitiría el uso más eficaz de los recursos de secuenciación, y puede potenciar adicionalmente la obtención de información de la muestra biológica. En una realización, todas las moléculas de ADN en una muestra de plasma deben secuenciarse en un análisis de secuenciación utilizando un protocolo sin PCR durante el análisis de secuenciación masiva en paralelo. Un protocolo sin PCR que se puede utilizar es el creado por Berry Genomics (investor.illumina.com/mobile.view?c=121127&v=203&d=1&id=1949110). También se puede utilizar otro protocolo sin PCR como el comercializado por Illumina (www.illumina.com/products/truseq-dna-pcr-free-sample-prep-kits.html). En el presente caso se utiliza un ejemplo para ilustrar el principio.

A modo ilustrativo, primero se asume que todos los fragmentos de ADN plasmático tienen un tamaño de 150 pb, lo cual es coherente con los fragmentos de ADN plasmático que generalmente tienen menos de 200 pb, tal como se menciona anteriormente. Por tanto, cada genoma humano diploide se fragmentaría en 40 millones de fragmentos de ADN plasmático. Como hay aproximadamente 1000 genomas humanos diploides en un mililitro de plasma, habría 40 mil millones de fragmentos de ADN plasmático en 1 ml de plasma. Si se secuencian 40 mil millones de fragmentos de ADN de 1 ml de plasma, se esperaría que todas las moléculas de ADN se hubieran secuenciado. A modo ilustrativo, si se utiliza un sistema Illumina HiSeq 2000 que puede producir 2 mil millones de lecturas por ejecución, se necesitarían 20 ejecuciones para conseguir esta cantidad de secuenciación, que puede reducirse con plataformas de mayor rendimiento.

La concentración total de ADN en la muestra de plasma se puede determinar utilizando, por ejemplo, pero sin limitación, PCR digital o PCR en tiempo real antes del análisis de secuenciación. La concentración de ADN total se puede utilizar para determinar la cantidad de secuenciación requerida para secuenciar todas las moléculas de ADN analizables o secuenciables en la muestra. En otras realizaciones que implican otros grados de secuenciación exhaustiva, se puede secuenciar más del 20 %, 25 %, 30 %, 40 %, 50 %, 60 %, 75 %, 90 %, 95 %, o 99 % de las moléculas de ADN en una muestra de plasma, siendo todas ellas ejemplos de secuenciación exhaustiva.

Los determinantes clave para el porcentaje de moléculas de ADN que hay que secuenciar incluyen la cantidad de mutaciones, fracción tumoral en la muestra y rendimiento de la biblioteca de ADN. El número de moléculas potencialmente secuenciables en una biblioteca de secuenciación se puede determinar en función del volumen, concentración y eficacia de conversión de la biblioteca. El número de fragmentos de ADN necesarios para secuenciar puede determinarse en función del límite detectable deseado de la fracción tumoral y el número esperado de mutaciones en el tumor. Basándose en estos dos números, se puede determinar la parte de la biblioteca que se va a secuenciar.

Una ventaja de utilizar un protocolo sin PCR para la secuenciación exhaustiva es que se puede inferir directamente las cantidades absolutas de cualquier molécula diana en la muestra en lugar de determinar una cantidad con respecto a otras dianas de referencia que se secuencian en la misma reacción. Esto se debe a que cada lectura de secuencia representa la información de una molécula de ADN plasmático original. De hecho, si la amplificación por PCR se utiliza con secuenciación ultraprofunda y amplia, la cantidad de moléculas diana en relación con las demás se alejaría más de la representación real. El motivo se debe a la generación de duplicados de PCR como resultado de la amplificación por PCR, así como a los sesgos de amplificación en los que algunas regiones genómicas se amplifican mejor que otras.

La amplificación por PCR de las bibliotecas de secuenciación se lleva a cabo comúnmente en la mayoría de los protocolos existentes para la secuenciación masiva en paralelo porque esta etapa puede aumentar la cantidad de moléculas en las bibliotecas de secuenciación para que la etapa de secuenciación se pueda realizar más fácilmente. Un duplicado (réplica) de PCR es un producto clonal de una molécula de ADN plantilla original. La presencia de duplicados de PCR dificulta la consecución de una secuenciación ultraprofunda y amplia. La proporción de lecturas de secuencias provenientes de réplicas de PCR aumentaría con la cantidad de secuenciación realizada (profundidad de secuenciación). En otras palabras, habría un rendimiento decreciente en el contenido de información exclusiva a medida que se realiza una secuenciación más profunda. Por lo tanto, la secuenciación de las réplicas de PCR, en muchos escenarios, daría lugar a un desperdicio de recursos de secuenciación. En última instancia, esto significaría que se necesita mucha más secuenciación para alcanzar la misma amplitud y profundidad de cobertura genómica en comparación con un protocolo sin PCR. Por lo tanto, los costes serían mucho más altos. De hecho, en algunos casos, la proporción de duplicados de PCR puede ser tan alta que en la práctica nunca se podría alcanzar una amplitud y profundidad de cobertura preferidas.

Esto es contrario a la intuición para los expertos en la materia. Tradicionalmente, la amplificación por PCR, incluyendo la amplificación del genoma completo, se realiza para proporcionar más material genético de una muestra finita para realizar más análisis moleculares. Los datos de los presentes inventores muestran que dicha etapa de amplificación puede ser contraproducente. Esto es particularmente contrario a la intuición para el análisis de ADN plasmático.

Se sabe que el ADN plasmático contiene una baja abundancia de ADN a baja concentración, como también es cierto para otras muestras compuestas de ADN sin células. Por lo tanto, no se podría pensar que se podría obtener más información sin la amplificación de la escasa cantidad de ADN. De hecho, en el protocolo de preparación de bibliotecas de los presentes inventores basado en amplificación, normalmente se obtienen de 150 a 200 nM de biblioteca de ADN unido a adaptador por 4 ml de plasma. Pero como se muestra en los ejemplos de la presente solicitud, solamente se obtienen 2 nanomoles de bibliotecas de ADN unido a adaptador a partir de una cantidad equivalente de volumen de plasma. Se podría imaginar que cantidades tan bajas serían un obstáculo para obtener más información genómica y, por lo tanto, podrían verse inducidos a realizar una etapa de amplificación antes del análisis. Dicha una biblioteca amplificada crearía problemas significativos ya que una proporción significativa de dicha una biblioteca consistiría en duplicados de PCR.

Asimismo, con una biblioteca tan ampliada, prácticamente no se podría realizar una secuenciación de plantilla total para obtener la mayor cantidad de información posible de la muestra de plasma de 4 ml (porque se aplica una cantidad

fija de biblioteca por ejecución de secuenciación y se necesitaría una cantidad extrema de ejecuciones para consumir la biblioteca). Como se muestra en los datos de los presentes inventores, se necesitan aproximadamente 20 ejecuciones de secuenciación de Illumina para consumir completamente las bibliotecas sin PCR de los casos de HCC y gestantes que se han estudiado. Si en su lugar se utilizaran protocolos de construcción de bibliotecas basados en PCR o amplificación, habría que realizar 100 veces la cantidad de secuenciación, lo que significa unas 2000 ejecuciones. En otras palabras, con una biblioteca ampliada, se están creando moléculas duplicadas que consumirían una parte significativa del poder de secuenciación. En contraposición, los 2 nanomoles de la biblioteca del protocolo sin PCR se pueden consumir fácilmente, lo que equivale a agotar la información analizable de la muestra de plasma de 4 ml.

Es importante poder utilizar una proporción razonable de la muestra de plasma de 4 ml. Como se ilustra con algunos cálculos presentados anteriormente, el número de equivalentes del genoma del ADN canceroso en la muestra de plasma es bajo durante el cáncer temprano y es necesario poder aprovechar la detección de tantos equivalentes del genoma del cáncer en la muestra de plasma como sea posible. Supóngase que se puede conseguir la clasificación del cáncer realizando 10 ejecuciones de secuenciación Illumina de una muestra de ADN plasmático utilizando un protocolo de preparación de bibliotecas sin PCR. Estas 10 ejecuciones habrían consumido la mitad de la biblioteca de secuenciación. Esto se correlaciona con haber hecho uso del contenido analizable de la mitad de la muestra de plasma, en concreto, 2 ml, parara conseguir la clasificación del cáncer. Por otro lado, 10 ejecuciones realizadas en una biblioteca amplificada por PCR de la misma muestra equivaldrían a consumir solamente el 0,5 % de la biblioteca (porque generalmente hay una amplificación de 100 veces en el rendimiento de la biblioteca del protocolo amplificado por PCR). Esto se correlaciona con haber hecho uso del contenido analizable de solamente 0,02 ml de la muestra de plasma original de 4 ml, y la cantidad de datos obtenidos no sería suficiente para conseguir la clasificación del cáncer. Por lo tanto, es contrario a la intuición que con el uso de menos biblioteca de ADN producida sin amplificación por PCR se podría obtener más información pertinente sobre el cáncer por cantidad fija de secuenciación.

Los expertos en la materia han demostrado que los duplicados de PCR, también conocidos como réplicas de PCR, podrían eliminarse con un procedimiento bioinformático que identifique cualquier lectura de secuencia que muestre coordenadas de nucleótidos de inicio y fin idénticas. Sin embargo, como se verá en una sección posterior, se ha identificado que las ubicaciones finales de los fragmentos de ADN plasmático no son aleatorias y, por lo tanto, se produciría un filtrado erróneo. Usando un protocolo sin PCR sin aplicar una etapa bioinformática para filtrar las lecturas de secuencias con las mismas coordenadas de nucleótidos de inicio y fin, se identificó un pequeño porcentaje de lecturas de secuencias (normalmente < 5 %) con coordenadas de inicio o fin idénticas o ambas. Esta observación es el resultado de la naturaleza no aleatoria del corte del ADN plasmático. Las realizaciones pueden incorporar la identificación de ubicaciones finales específicas del cáncer como un criterio de filtrado para identificar fragmentos de ADN canceroso informativos. La adopción de un protocolo sin PCR facilitaría dicho análisis y el uso de este criterio. Asimismo, esto también significa que la práctica anterior de eliminar lecturas de secuencias con coordenadas de nucleótidos de inicio y final idénticas, de hecho, ha eliminado fragmentos de ADN canceroso informativos utilizables, dando como resultado la pérdida de contenido de información relacionada con el cáncer de la muestra de ADN plasmático.

La tasa de error de secuenciación de las plataformas de secuenciación de Illumina es de aproximadamente el 0,1 % al 0,3 % de los nucleótidos secuenciados (Loman *et al.* Nat Biotechnol 2012; 30: 434-439; Kitzman *et al.* Sci Transl Med 2012; 4: 137ra76). Las tasas de error informadas para algunas plataformas de secuenciación diferentes son incluso más altas. Como se ha demostrado una tasa de error de secuenciación del 0,3 % no es trivial y ha creado un obstáculo para que los investigadores identifiquen mutaciones *de novo* fetales (Kitzman *et al.* Sci Transl Med 2012; 4: 137ra76) o mutaciones somáticas específicas del cáncer en plasma con una precisión muy alta. Esta tasa de error es aún más pertinente para la secuenciación ultra profunda y amplia. Un 0,3 % de errores en un conjunto de datos de secuenciación con una profundidad de 200x se traduce en 200 millones de errores.

Una proporción de dichos errores de secuenciación se generan por las etapas de amplificación por PCR durante las etapas de preparación de la biblioteca de ADN previas a la secuenciación. Mediante el uso de un protocolo sin PCR para la preparación de bibliotecas, este tipo de errores podrían reducirse. Esto haría que la secuenciación fuera más rentable porque se podrían emplear menos reactivos en la secuenciación de estos artefactos y menos tiempo de procesos bioinformáticos en el procesamiento de estos errores. Además, las verdaderas mutaciones *de novo* fetales y las mutaciones somáticas paraneoplásicas podrían identificarse de manera más específica entre menos falsos positivos a una profundidad de secuenciación menor que si estuviera implicada de otro modo la amplificación por PCR. De hecho, estas ventajas no han sido evidentes para otros investigadores (véase la siguiente sección).

C. Resultados de secuenciación con y sin amplificación previa de bibliotecas de secuenciación

Se realizó un análisis de simulación para comparar la cantidad de secuenciación requerida para detectar mutaciones paraneoplásicas en plasma para protocolos con y sin amplificación previa de bibliotecas de secuenciación con PCR. Para determinar la proporción de lecturas de secuencias de las réplicas de PCR, es decir, secuenciar una molécula más de una vez, se han utilizado los siguientes supuestos: (1) hay 500 equivalentes de genoma de ADN en 1 ml de plasma; (2) el ADN se extrae de 2 ml de plasma con un rendimiento del 50 %; (3) el 40 % del ADN extraído se puede convertir con éxito en una biblioteca de secuenciación; (4) se realizaron 10 ciclos de PCR para la amplificación previa

y la eficacia de la PCR es del 100 %; (5) el patrón de fragmentación para las bibliotecas amplificadas y no amplificadas previamente es idéntico; (6) la longitud del ADN plasmático es de 166 pb.

La figura 3 es un gráfico 300 que muestra la relación entre el porcentaje de lecturas de secuencias de réplicas de PCR y la profundidad de secuenciación. El porcentaje de lecturas de secuencias provenientes de réplicas de PCR aumenta con la profundidad de la secuenciación. A una profundidad de secuenciación de 200x, el 44 % de las lecturas de secuencias serían de réplicas de PCR. Dichas lecturas de secuencias de réplicas de PCR no proporcionarían información adicional.

Las figuras 4A y 4B muestran una comparación entre la profundidad de secuenciación necesaria para PCR y protocolos sin PCR para detectar mutaciones paraneoplásicas en el plasma de un sujeto con cáncer en varias fracciones de ADN tumoral según las realizaciones de la presente invención. Según el porcentaje previsto de las réplicas de PCR, se realizó un análisis de simulación para determinar la cantidad de secuenciación necesaria para detectar mutaciones paraneoplásicas en el plasma de un sujeto con cáncer. Se realizaron simulaciones para cubrir fracciones de ADN tumoral en plasma del 1 % al 10 %. Se asumió que están presentes 30.000 mutaciones en el genoma de una célula cancerosa en este sujeto.

El protocolo con amplificación previa de PCR requeriría una mayor profundidad de secuenciación para detectar las mutaciones paraneoplásicas en cualquier fracción de ADN tumoral en el plasma. La diferencia en la profundidad de secuenciación requerida aumentaría exponencialmente con la reducción de la fracción de ADN tumoral. A una fracción de ADN tumoral en plasma del 10 %, los protocolos con y sin amplificación previa de PCR requieren profundidades de secuenciación de 37x y 25x, respectivamente. Sin embargo, a una fracción de ADN tumoral en plasma del 2 %, la profundidad de secuenciación correspondiente requerida sería 368x y 200x.

Por tanto, el uso de un protocolo sin PCR es muy ventajoso para la detección de cambios en el plasma paraneoplásicos, en particular cuando la fracción de ADN tumoral en el plasma es baja. Si el número de mutaciones presentes dentro del genoma tumoral del plasma es menor, se necesitarían mayores profundidades de secuenciación. La diferencia en la profundidad necesaria para los protocolos con o sin amplificación sería aún mayor, especialmente cuando la fracción de ADN tumoral en la muestra de plasma es baja.

D. Distinción de la "secuenciación profunda" convencional

Hay una serie de características que distinguen el uso de la secuenciación exhaustiva para conseguir una secuenciación ultra profunda y amplia de los métodos de secuenciación anteriores. En un aspecto, algunas de las estrategias de secuenciación anteriores denominadas "secuenciación profunda" normalmente implicarían la amplificación de una secuencia diana de interés, por ejemplo, mediante PCR. Posteriormente, el ADN amplificado, también llamado amplicón, se secuencia varias veces mediante secuenciación. Un ejemplo de esta estrategia es la secuenciación profunda de amplicones etiquetados (Forshew *et al.* Sci Transl Med 2012; 4: 136ra68). La secuenciación exhaustiva, por otro lado, se implementa de manera más eficaz sin ninguna etapa de amplificación, ya que entonces todos los fragmentos detectados son fragmentos originales y no datos replicados, lo que permite una mayor amplitud y profundidad real (en oposición a la profundidad aparente). Por profundidad aparente, los presentes inventores se refieren a la secuenciación de una biblioteca de secuenciación amplificada en la que una proporción de la potencia de secuenciación se consume en la secuenciación de duplicados de PCR y, por lo tanto, el rendimiento de información de la secuenciación no es proporcional a su profundidad.

Debido a que la secuenciación profunda normalmente utiliza una etapa de amplificación, una proporción de la potencia de secuenciación se gasta en la secuenciación de duplicados de PCR. La existencia de dichos duplicados de PCR dificultaría mucho el análisis exhaustivo de cada molécula de ADN plantilla dentro de la muestra mediante la secuenciación profunda de bibliotecas de secuenciación amplificada. Varios grupos han descrito métodos para proporcionar información sobre la tasa de duplicación, por ejemplo, mediante un código de barras de la biblioteca de secuenciación (Kinde *et al.* Proc Natl Acad Sci USA 2011; 108: 9530-9535). Por ejemplo, en el método descrito por Kinde et al, se tienen que realizar tres etapas: (i) asignación de un identificador exclusivo (UID, por sus siglas en inglés) a cada molécula plantilla, (ii) amplificación de cada molécula de plantilla etiquetada de forma exclusiva para crear familias de UID, y (iii) secuenciación redundante de los productos de amplificación. En contraposición, el uso de bibliotecas sin PCR para una secuenciación exhaustiva evitaría los problemas causados por los duplicados de PCR, y no sería necesario el método descrito por Kinde et al.

De hecho, la mayoría de las estrategias de secuenciación profunda puestas en práctica anteriormente no pueden conseguir la amplitud que podría conseguirse con el uso de secuenciación exhaustiva. Por ejemplo, la secuenciación de amplicones generalmente consigue una gran profundidad para una región genómica estrecha. Incluso con el uso de multiplexación, la amplitud total del genoma cubierto es limitada y está lejos de abarcar todo el genoma. Como se explica en la presente solicitud, para la prueba de cribado de cáncer, se prefiere lo más cercana a la cobertura del genoma completo para cubrir tantos supuestos sitios de mutación como sea posible. Por ejemplo, incluso si se aplica un grado extremo de secuenciación de amplicones multiplexada, por ejemplo, 3 millones de amplicones, cubriendo cada uno 1.000 bases, los duplicados de PCR se convertirían en un problema como se describe anteriormente.

De manera similar, los investigadores han aplicado la captura por hibridación para conseguir una secuenciación profunda de regiones genómicas selectivas, denominada secuenciación dirigida. Sin embargo, los protocolos de captura normalmente implican etapas de amplificación. Cuando el tamaño de la región diana es relativamente pequeño, se alcanzarían grandes proporciones de duplicados de PCR, un 50 % incluso hasta un 90 % (Nueva *et al.* J Clin Endocrinol Metab 2014; 99: E1022-1030) cuando se realizara la secuenciación dirigida en ADN plasmático. A niveles tan altos de duplicación de PCR, la profundidad eficaz de la secuenciación se reduce. La amplitud de la secuenciación está limitada por el tamaño de la región diana.

Estas observaciones ilustran que los investigadores no han estado motivados para conseguir una secuenciación que sea amplia y profunda al mismo tiempo. Sin embargo, adoptando los principios de secuenciación exhaustiva descritos en la presente solicitud, se pueden modificar los protocolos de secuenciación dirigida para garantizar que las tasas de duplicación de PCR se mantengan al mínimo mientras se necesita capturar una gran proporción del genoma humano. Por ejemplo, se puede utilizar la amplificación ligera para preparar la biblioteca de secuenciación diana para mantener los duplicados de PCR al mínimo. Posteriormente, la amplitud del análisis tendría que conseguirse agrupando los datos de múltiples paneles diana. Sin embargo, cuando se tienen en cuenta estas consideraciones, el estrategia dirigida puede no ser más rentable que la estrategia de secuenciación exhaustiva no dirigida. No obstante, puede haber otras razones por las que se prefiera el enriquecimiento dirigido de una gran parte del genoma. Por ejemplo, se puede justificar la necesidad de enfocar el esfuerzo de secuenciación exhaustiva a las regiones repetidas o no repetidas del genoma si una parte muestra agrupamiento para la aparición de mutaciones *de novo* o somáticas. A modo de ejemplo, se puede preferir centrar los esfuerzos en la heterocromatina en lugar de la región de eucromatina del genoma.

E. Para análisis fetal

La secuenciación exhaustiva de ADN plasmático puede ser útil para las pruebas prenatales no invasivas. El ADN fetal está presente en el plasma de una mujer gestante (Lo *et al.* Lancet 1997; 350: 485-487) y puede utilizarse para la prueba prenatal no invasiva de un feto (por ejemplo, para determinar aneuploidías cromosómicas y trastornos de un solo gen).

Hasta ahora, la detección de mutaciones fetales *de novo* por secuenciación del ADN plasmático materno se ven obstaculizadas por la tasa de error de secuenciación de la generación actual de secuenciadores de manera masiva en paralelo (Kitzman *et al.* Sci Transl Med 2012; 4: 137ra76 y publicación de patente de los Estados Unidos US 2015/0105261 A1). Por lo tanto, usando una estrategia previamente informada, se identificarían millones de mutaciones fetales *de novo* candidatas en el plasma materno, pero solo varias decenas de ellas serían mutaciones verdaderas a pesar de la incorporación de etapas bioinformáticas para filtrar posibles falsos positivos.

Sin embargo, utilizando una secuenciación exhaustiva del ADN plasmático materno, se podría superar este problema. Utilizando un proceso de preparación de bibliotecas sin PCR, una mutación fetal *de novo* candidata que se identifica en más de una molécula de ADN plasmático materno tendría una mayor probabilidad de ser una verdadera mutación. En otras realizaciones, se puede establecer un criterio de clasificación más riguroso, tal como la misma mutación identificada más de 2, 3, 4, 5 o más veces en la muestra de plasma materno.

Varios trabajadores han utilizado la secuenciación de una sola molécula, por ejemplo, utilizando la plataforma Helicos, para la prueba prenatal no invasiva del plasma materno para detectar aneuploidías cromosómicas fetales (van den Oever *et al.* Clin Chem 2012; 58: 699-706 y van den Oever *et al.* Clin Chem 2013; 59: 705-709). Sin embargo, dicho trabajo se realizó mediante la secuenciación de una pequeña fracción de las moléculas en el plasma y, por lo tanto, no consiguió una secuenciación amplia y profunda.

F. Aplicaciones adicionales de la secuenciación exhaustiva

En otra realización, se puede utilizar la secuenciación exhaustiva metilómica plasmática para identificar moléculas de ADN plasmático derivadas de diferentes órganos del organismo. Esto es posible porque diferentes tejidos dentro del cuerpo tienen diferentes perfiles de metilación. A través de un proceso de desconvolución, se pueden identificar las contribuciones relativas de diferentes tejidos en el plasma (Sun *et al.* Proc Natl Acad Sci USA 2015; 112: E5503-5512).

En otra realización de secuenciación exhaustiva de ADN plasmático, se pueden identificar mutaciones en el ADN plasmático que están asociadas con múltiples procesos fisiológicos o patológicos. En una realización, dichos procesos incluyen los asociados con el envejecimiento. En otra realización, dichos procesos incluyen aquellos asociados con agentes ambientales, por ejemplo, polución, radiación, agentes infecciosos, productos químicos tóxicos, etc. En esta última realización, diferentes procesos pueden tener sus propias firmas mutacionales (Alexandrov *et al.* Nature 2013; 500: 415-421).

La secuenciación exhaustiva de ácido nucleico plasmático también se puede aplicar a la secuenciación de ARNm y ARN no codificante (por ejemplo, microARN y ARN largo no codificante) en plasma. Los datos anteriores han demostrado que el perfil transcriptómico del plasma permitiría desconvolucionar las contribuciones de varios tejidos de la muestra de plasma (Koh *et al.* Proc Natl Acad Sci USA 2014; 111: 7361-7366). La secuenciación transcriptómica

exhaustiva del plasma potenciaría adicionalmente la solidez y la utilidad de dicha estrategia.

V. CRITERIOS DE FILTRADO PARA IDENTIFICAR MUTACIONES

Como se describe anteriormente en la sección III.B, la especificidad en la identificación de mutaciones y cualquier prueba que utilice dichas mutaciones (por ejemplo, el uso de la carga mutacional para determinar un nivel de cáncer) puede mejorarse aplicando criterios de filtrado a los locus en los que se hayan alineado una o más lecturas de secuencias que tengan una mutación. A modo de ejemplo para el cáncer, se puede conseguir una alta especificidad puntuando una firma genética o genómica como positiva solamente cuando existe una alta confianza en que esté asociada al cáncer. Esto podría conseguirse minimizando el número de errores de secuenciación y alineación que pueden identificarse erróneamente como una mutación, por ejemplo, mediante la comparación con el perfil genómico de un grupo de controles sanos, o puede conseguirse mediante la comparación con el propio ADN constitutivo de la persona o puede conseguirse mediante la comparación con el perfil genómico de la persona en un momento anterior.

Se podrían aplicar varios criterios como criterios de filtrado para evaluar la probabilidad de que un fragmento de ADN porte una mutación. Cada uno de los criterios de filtrado podría utilizarse individualmente, independientemente, colectivamente con igual ponderación o diferentes ponderaciones, o en serie en un orden especificado, o condicionalmente en función de los resultados de las etapas de filtrado anteriores, como se describe anteriormente. A continuación se proporcionan ejemplos de criterios de filtrado.

A. Valor de corte dinámico

Se pueden utilizar uno o más criterios de filtrado con valor de corte dinámico para distinguir variantes de un solo nucleótido, en concreto, mutaciones y polimorfismos, de cambios de nucleótidos debido a un error de secuenciación. Dependiendo del contexto, las mutaciones pueden ser mutaciones *"de novo*" (por ejemplo, nuevas mutaciones en el genoma constitutivo de un feto) o "mutaciones somáticas" (por ejemplo, mutaciones en un tumor). Se pueden determinar varios valores de parámetros para cada uno de una pluralidad de locus, donde cada valor de parámetro se compara con un valor de corte correspondiente. Se puede descartar que un locus tiene una mutación potencial si el valor de un parámetro no satisface un valor de corte.

Para la identificación de mutaciones somáticas en cáncer, los datos de secuenciación de alta profundidad del ADN constitutivo de una persona (por ejemplo, capa leucocitaria) y el ADN plasmático se pueden comparar para identificar sitios que son heterocigotos (AB) en el ADN plasmático y homocigotos (AA) en el ADN constitutivo. "A" y "B" indican los alelos mutantes y de tipo silvestre, respectivamente. En el presente caso, se ilustra una realización de la implementación de la estrategia de valores de corte dinámicos para la detección de mutaciones, donde, se utilizaron los modelos de distribución binomial y de Poisson para calcular tres parámetros.

En cuanto a un primer parámetro, la precisión de determinar los sitios homocigotos (AA) en el ADN constitutivo se ve afectada por el error de secuenciación. El error de secuenciación se puede estimar mediante una serie de métodos conocidos por los expertos en la materia. Por ejemplo, la tasa de error de secuenciación (denominada "ε") de las plataformas HiSeq de Illumina se ha estimado en 0,003. Suponiendo que los recuentos secuenciados siguen una distribución binomial, se calcula el primer parámetro, Score 1, como Score1 = 1-pbinom(c, D, ε). D representa la profundidad de secuenciación, que es igual a la suma de "c" y "a", "c" se refiere al número de lecturas de secuencias que cubren el alelo B mutante, "a" se refiere al número de lecturas de secuencias que cubren el alelo A de tipo silvestre. El término "pbinom" es la función de distribución acumulativa binomial, que se puede escribir como

$$\sum_{i=0}^{c} \binom{D}{i} \varepsilon^{i}(1-\varepsilon)^{D-i},$$

donde $\binom{D}{i}$ representa una función de combinación matemática, es decir, el número de combinaciones que seleccionan i veces del alelo mutante desde la profundidad de secuenciación D, que se puede escribir adicionalmente usando el factorial como $\frac{D!}{i!(D-i)!}$. Cuanto mayor sea el valor de Score1, más seguro será que el genotipo real es AA. Se podría utilizar un valor de corte superior a 0,01. Este parámetro se puede utilizar para controlar la influencia de los errores de secuenciación.

En cuanto a un segundo parámetro, existe la posibilidad de que el AA (homocigoto) de tipo silvestre observado en el genoma constitutivo esté mal asignado a partir del genotipo AB (heterocigoto) real debido a la profundidad de secuenciación insuficiente de los locus con SNP. Para minimizar la influencia de este tipo de error, se calcula el segundo parámetro, Score2, como Score2 = ppois(b, D/2), donde "b" es el número de recuentos secuenciados que cubren el alelo B, y "ppois" es la función de distribución acumulativa de Poisson, que se puede escribir como

$$\sum_{i=0}^{b} \frac{\lambda^i e^{-\lambda}}{i!},$$

donde λ es la profundidad de secuenciación promedio por cadena (es decir, D/2); e es la base de los logaritmos naturales (~2,717828). Cuanto menor sea el valor de Score2, más seguro será que el genotipo real es AA. Por ejemplo, puede utilizarse un valor de corte de <0,001, 0,0001, 10^{-10}, etc. Este parámetro se puede utilizar para controlar la eliminación de alelos o variantes, que se refiere a sitios heterocigotos que aparecen como sitios homocigotos porque un alelo o variante no se pudo amplificar y, por lo tanto, se ha eliminado este alelo o variante ausente. Determinados datos a continuación usan valores de corte de score 1 > 0,01 y score2 < 0,001, donde score1 y score2 pueden utilizarse para garantizar que la capa leucocitaria sea homocigota.

En cuanto a un tercer parámetro, existe la posibilidad de que el AB mutante observado esté mal asignado a partir del genotipo AA real debido a errores de secuenciación. Para minimizar la influencia de este tipo de error, se calcula el tercer parámetro, Score3, como Score 3 = $\binom{D}{b} \times \varepsilon \times \binom{\varepsilon}{3}^{(b-1)}$, donde $\binom{D}{b}$ representa una función de combinación matemática, es decir, el número de combinaciones que seleccionan b veces del alelo mutante desde la profundidad de secuenciación D, que se puede escribir adicionalmente usando el factorial como $\frac{D!}{b!(D-b)!}$; "ε" representa la tasa de error de secuenciación que se estimó en 0,003 en este ejemplo. Cuanto menor sea el Score3, más seguro será que el genotipo real es AB. Por ejemplo, puede utilizarse un valor de corte de <0,001, 0,0001, 10^{-10}, etc.

Score1 y Score2 se pueden aplicar al tejido constitutivo y Score 3 se puede aplicar a la mezcla (tumor o plasma). Por lo tanto, el análisis conjunto entre tejidos constitutivos y muestras de mezcla ajustando Score1, Score2 y Score3 se pueden realizar para determinar las posibles mutaciones.

Se pueden utilizar diferentes umbrales para el cálculo de cada puntuación en el punto de corte dinámico según el propósito previsto. Por ejemplo, podría usarse un valor menor para Score3 si se prefiere una alta especificidad en la identificación de mutaciones somáticas. De manera similar, se podría usar un valor más alto para Score3 si se prefiere detectar una mayor suma total de mutaciones somáticas. La especificidad de las mutaciones somáticas identificadas se puede mejorar utilizando otros parámetros de filtrado, por ejemplo, tal como se describe a continuación. También se pueden utilizar otros modelos matemáticos o estadísticos, por ejemplo, distribución de Chi cuadrado, distribución gamma, distribución normal, y otros tipos de modelos mixtos. El proceso podría aplicarse de manera similar para la identificación de mutaciones fetales *de novo*.

B. Realineación

Uno o más criterios de filtrado de realineación pueden reducir los efectos de los errores de secuenciación y alineación en la detección de variantes de secuencia a partir de datos de secuenciación y, por lo tanto, también reducen los falsos positivos en la identificación de mutaciones. A continuación se describen varias realizaciones que usan realineación.

En un (primer) procedimiento de alineación inicial, las lecturas de secuenciación se pueden alinear (mapear) con un genoma de referencia (por ejemplo, un genoma humano de referencia), por ejemplo, mediante cualquier técnica de alineación disponible para los expertos en la materia, por ejemplo, SOAP2 (Li *et al.* Bioinformatics 2009; 25: 1966-7). Después de la alineación con un locus, se puede realizar una comparación con un genoma (por ejemplo, un genoma de referencia, un genoma constitutivo del sujeto o asociado con el sujeto, o genomas de los progenitores del sujeto) para identificar si existe una variante de secuencia en las lecturas.

Las lecturas de secuencias que contienen las supuestas variantes se pueden realinear (mapear de nuevo) con el genoma humano de referencia mediante el uso de un (segundo) alineador independiente, por ejemplo, Bowtie2 (Langmead *et al.* Nat Methods 2012; 9: 357-9). El alineador independiente sería diferente del alineador inicial en cuanto al uso de algoritmos de coincidencia. Los ejemplos de algoritmos coincidentes utilizados por el alineador inicial y el realineador pueden incluir, por ejemplo, pero sin limitación, el algoritmo de Smith-Waterman (SW), algoritmo de Needleman-Wunsch, algoritmo Hashing y transformación de Burrows-Wheeler. La realineación puede identificar y cuantificar la calidad o certeza de las mutaciones identificadas. El alineador independiente puede diferir del alineador inicial en otros aspectos, también, tal como el umbral de notificación de una alineación válida, penalizaciones por inserciones/eliminaciones y faltas de coincidencia, el número de faltas de coincidencia permitidas, el número de nucleótidos que se utilizan como semillas para la alineación.

En algunas realizaciones, los siguientes criterios de realineación se pueden usar solos o en combinación para identificar una lectura mapeada como una lectura de secuencia de baja calidad: (1) la lectura de secuencia portadora de la mutación no se recupera mediante un alineador independiente, que no se alinea (mapea) con la lectura de secuencia; (2) la lectura de secuencia portadora de la mutación muestra resultados de mapeo inconsistentes cuando

se utiliza un alineador independiente para verificar la alineación original (por ejemplo, una lectura mapeada se coloca en un cromosoma diferente en comparación con el resultado de la alineación original); (3) la lectura de la secuencia portadora de la mutación alineada con la misma coordenada genómica muestra una calidad de mapeo inferior a un umbral especificado utilizando el alineador independiente (por ejemplo, calidad de mapeo ≤ Q20 (es decir, probabilidad de alineación errónea <1 %); otros ejemplos de umbrales pueden ser 0,5 %, 2 % y 5 % de probabilidad de alineación errónea; (4) la lectura de secuencia tiene la mutación ubicada dentro de las 5 pb de cualquiera de los extremos de lectura (es decir, extremos 5' o 3'). Esta última regla de filtrado puede ser importante porque los errores de secuenciación eran más frecuentes en ambos extremos de una lectura de secuencia. La calidad del mapeo es una métrica definida dentro de un alineador y especifica una probabilidad de que una lectura de secuencia esté alineada de manera errónea. Diferentes alineadores pueden usar diferentes métricas.

Si la proporción de lecturas de secuencias de baja calidad entre las lecturas de secuencias portadoras de la mutación es mayor que determinado umbral, (por ejemplo, 30 %, 35 %, 40 %, 45 % o 50 %), se puede descartar el sitio mutante candidato. Por lo tanto, si las lecturas de secuencias restantes son inferiores a un umbral, entonces el locus se puede descartar de un conjunto de locus que se identifican como portadores de una mutación en al menos algún tejido (por ejemplo, tejido de un tumor o tejido de un feto).

En trabajos anteriores, incluyendo los esfuerzos del algoritmo MuTect (Cibulskis *et al.* Nat Biotechnol 2013 y GATC (www.gatc-biotech.com); 31: 213-219), solo se realinearon los posibles sitios de inserción o eliminación. Esos otros esquemas no vuelven a calcular la puntuación de calidad de una lectura de secuencia utilizando datos de un alineador diferente. Asimismo, no se ha demostrado que se pueda utilizar una puntuación de calidad recalculada a efectos de filtrar supuestas variantes o mutaciones. Los datos se muestran a continuación para ilustrar la eficacia del uso de un procedimiento de realineación.

C. Fracción de mutación

Los expertos en la materia reconocerán que existen métodos disponibles para medir la concentración fraccionaria de ADN fetal en el plasma materno o la concentración fraccionaria de ADN tumoral en el plasma de un sujeto con cáncer. Por lo tanto, en una realización, para mejorar la posibilidad de identificar un verdadero fragmento de ADN informativo, solamente los alelos o variantes con un recuento fraccionario igual o superior a la concentración fraccionaria medida por otro método se considerarían como variantes o mutaciones verdaderas. El valor de corte de concentración fraccionaria se denomina umbral de fracción mutante (M%), o simplemente umbral de fracción. Otras implementaciones pueden usar un umbral inferior a la concentración fraccionaria medida, pero el umbral seleccionado puede depender del valor medido (por ejemplo, dentro de un porcentaje específico de la concentración fraccionaria medida).

En otra realización, podrían adoptarse otros valores como el umbral de la fracción mutante incluso sin tener en cuenta la fracción de ADN fetal o la fracción de ADN tumoral medidas. Puede usarse un M% más alto como valor de corte si se prefiere una mayor especificidad en la identificación de mutaciones. Se puede usar el M% más bajo como valor de corte si se prefiere una mayor sensibilidad en la identificación de mutaciones. Los ejemplos para el umbral de fracción incluyen un 5 %, 10 %, 15 %, 20 %, 25 % y 30 %.

En otra realización más, la varianza en la fracción alélica de supuestas mutaciones dentro de regiones cromosómicas contiguas podría proporcionar información sobre la probabilidad de que los fragmentos de ADN de la región sean fragmentos de ADN canceroso informativos. Por ejemplo, las regiones cromosómicas contiguas de interés pueden ser aquellas con aberraciones en el número de copias. En regiones con ganancias en el número de copias, habría un enriquecimiento en el ADN derivado del tumor. Por lo tanto, se esperaría que la fracción alélica de las mutaciones somáticas verdaderas fuera mayor en dichas regiones con ganancias, que en las regiones con pérdidas de número de copias (debido al agotamiento del ADN derivado del tumor en estas últimas regiones).

El intervalo o la varianza en las proporciones alélicas de las supuestas mutaciones verdaderas sería mayor en las regiones con ganancia de número de copias que en las regiones con pérdida de número de copias. Por lo tanto, se podrían establecer diferentes M% como valores de corte de filtrado para regiones con ganancias o pérdidas de número de copias para aumentar la probabilidad de identificar mutaciones somáticas verdaderas. Los valores de corte que especifican la varianza en la fracción mutante plasmática observada también podrían usarse para identificar moléculas de ADN que se han originado a partir de regiones cromosómicas que es más probable que se enriquezcan con (para regiones con aumento de número de copias) o se agoten de (para regiones con pérdidas de número de copias). pérdidas) ADN derivado de tumor. Entonces podría tomarse una decisión con respecto a la probabilidad de que los fragmentos de ADN sean fragmentos de ADN canceroso informativos.

D. Filtro de tamaño

Mientras que el ADN plasmático generalmente circula como fragmentos de < 200 pb de longitud, las moléculas de ADN plasmático derivadas de tumores y derivadas del feto son más cortas que las moléculas de ADN no fetales y no tumorales de fondo, respectivamente (Chan *et al.* Clin Chem 2004; 50: 88-92 y Jiang *et al.* Proc Natl Acad Sci USA 2015; 112: E1317-1325). Por tanto, el tamaño pequeño se puede utilizar como otra característica que aumenta la

probabilidad de que un fragmento de ADN plasmático derive del feto o derive de un tumor. Por lo tanto, en algunas realizaciones, podría aplicarse un criterio de filtrado por tamaño del ADN.

Se pueden utilizar varios criterios de tamaño. Por ejemplo, se puede requerir que una diferencia de umbral en los tamaños medianos entre fragmentos de ADN portadores de alelos mutantes y alelos de tipo silvestre sea al menos un determinado número de bases, que se puede denominar ΔS. Por lo tanto, ΔS≥10 pb se puede utilizar como criterio de filtro de tamaño. Ejemplos de otros umbrales de tamaño incluyen 0 pb, 1 pb, 2 pb, 3 pb, 4 pb, 5 pb, 6 pb, 7 pb, 8 pb, 9 pb, 11 pb, 12 pb, 13bp, 14 pb, 15 pb, 16 pb, 17 pb, 18 pb, 19 pb y 20 pb. También se pueden utilizar otras pruebas estadísticas, por ejemplo, prueba de la t, prueba de la U de Mann-Whitney, prueba de Kolmogorov-Smirnov, etc. Se puede determinar un valor de p utilizando estas pruebas estadísticas y compararlo con un umbral para determinar si los fragmentos de ADN portadores de la variante de secuencia serían significativamente más cortos que los portadores de los alelos de tipo silvestre. Los ejemplos del umbral para el valor de p pueden incluir, pero sin limitación, 0,05, 0,01, 0,005, 0,001, 0,0005 y 0,0001.

En consecuencia, en una realización, se puede obtener la información sobre el tamaño de las moléculas de ADN plasmático secuenciadas. Se puede hacer esto usando secuenciación de extremos emparejados, que incluye la secuenciación de toda la molécula de ADN. Para lo último, como las moléculas de ADN plasmático generalmente tienen menos de 166 pb, la secuenciación de toda la molécula de ADN podría realizarse fácilmente utilizando muchas plataformas de secuenciación masiva en paralelo de lectura corta. Como el ADN plasmático derivado de las células cancerosas es generalmente corto, mientras que el de los tejidos peritumorales o no tumorales es generalmente largo (Jiang *et al.* Proc Natl Acad Sci 2015; 112: E1317-1325), tener la información del tamaño del ADN plasmático ayudaría adicionalmente a la clasificación de los fragmentos secuenciados como derivados probables de células cancerosas o no cancerosas. Esta información ayudaría adicionalmente al cribado, detección, pronóstico y monitorización del cáncer.

Y, ya que el ADN fetal en el plasma materno es más corto que el ADN materno (Chan *et al.* Clin Chem 2004; 50: 88-92 y Yu *et al.* Proc Natl Acad Sci USA 2014; 111: 8583-8588), también se puede utilizar la información de tamaño del ADN plasmático al interpretar los resultados de la secuenciación exhaustiva del ADN plasmático. Por lo tanto, un fragmento más corto en el plasma materno tiene una mayor probabilidad de derivar del feto.

E. Estado de metilación

El perfil de metilación del ADN es diferente entre diferentes tejidos. Algunas firmas de metilación son relativamente específicas de tejido. Por ejemplo, el promotor de *SERPINB5* está hipometilado en la placenta (Chim *et al.* Proc Natl Acad Sci USA 2005; 102: 14753-14758) y el promotor de *RASSF1A* está hipermetilado en la placenta (Chiu *et al.* Am J Pathol 2007; 170: 941-950). Los promotores de determinados genes supresores de tumores, incluyendo *RASSF1A,* están hipermetilados en los diferentes tipos de cáncer. Sin embargo, en la placenta (Lun *et al.* Clin Chem 2013; 59: 1583-1594) y tejidos cancerosos (Chan *et al.* Proc Natl Acad Sci 2013; 110: 18761-18768) se muestra globalmente hipometilado, especialmente en las regiones no promotoras.

Como se ha demostrado que el ADN fetal en el plasma materno tiene diferentes patrones de metilación del ADN a partir del ADN derivado de la madre, la información sobre la metilación del ADN puede ayudar a predecir la probabilidad de que una molécula secuenciada sea de origen materno o fetal. En una realización, ya que la placenta es una fuente importante de ADN fetal en el plasma materno y el ADN placentario está más hipometilado que el ADN de las células sanguíneas maternas (Lun *et al.* Clin Chem 2013; 59: 1583-1594), es más probable que un fragmento de ADN hipometilado secuenciado a partir de plasma materno sea de origen fetal. De manera similar, en una realización, ya que el ADN tumoral está más hipometilado que el ADN de las células sanguíneas (Chan *et al.* Proc Natl Acad Sci 2013; 110: 18761-18768), es más probable que un fragmento de ADN hipometilado que contiene una supuesta mutación (candidata) secuenciada a partir del plasma de un individuo analizado para detectar cáncer sea paraneoplásico o sea específico del cáncer que uno que no tenga hipometilación.

El estado de metilación se puede utilizar de varias formas para determinar si un locus presenta una mutación. Por ejemplo, se puede requerir una cantidad umbral de densidad de metilación de los fragmentos de ADN que se alinean con el locus con la mutación antes de que el locus se considere una mutación. Como otro ejemplo, se puede utilizar una puntuación binaria de un sitio CpG, por ejemplo, donde solamente hay un sitio CpG por fragmento de ADN. Se puede descartar un sitio CpG si el fragmento de ADN no tiene el estado de metilación esperado. La decisión de descartar un fragmento de ADN puede depender de otros criterios de filtrado. Por ejemplo, si el fragmento de ADN es suficientemente corto, entonces el fragmento de ADN se puede conservar. Este es un ejemplo del uso de varios criterios de filtrado en combinación con diferentes ponderaciones o en combinación como parte de un árbol de decisión.

El análisis de metilación del ADN plasmático podría conseguirse mediante estrategias con reconocimiento de la metilación, incluyendo la conversión con bisulfito, digestión con enzimas de restricción sensibles a la metilación o tratamiento con proteínas de unión a metilo. Todos estos procesos con reconocimiento de la metilación podrían seguirse mediante una secuenciación masiva en paralelo, secuenciación de una sola molécula, micromatriz, PCR digital o análisis PCR. Además, algunos protocolos de secuenciación de una sola molécula podrían leer directamente el estado de metilación de una molécula de ADN sin tratamiento previo por parte de otros procesos con reconocimiento

de la metilación (Ahmed *et al.* J Phys Chem Lett 2014; 5: 2601-2607).

Además de la metilación de la citosina, hay otras formas de metilación del ADN, tales como, pero sin limitación, hidroximetilcitosina (Udali *et al.* Hepatology 2015; 62: 496-504). Los tejidos cerebrales (Sherwani y Khan. Gene 2015; 570: 17-24) y el melanoma (Lee *et al.* Lab Invest 2014; 94: 822-838) muestran una mayor proporción de hidroximetilcitosinas.

F. Ubicación final del ADN plasmático

También se puede realizar el filtrado de posibles mutaciones específicas del cáncer o paraneoplásicas o fetales, basándose en la coordenada del nucleótido terminal o de la ubicación final. Los presentes inventores han identificado ubicaciones terminales de fragmentos de ADN que no son aleatorias y que varían en función del tejido de origen. Por lo tanto, la ubicación terminal puede utilizarse para determinar la probabilidad de que una lectura de secuencia con una supuesta mutación proceda realmente de tejido fetal o de tejido tumoral.

Recientemente, se ha demostrado que el patrón de fragmentación del ADN plasmático no es aleatorio (Snyder *et al.* Cell 2016; 164: 57-68 y el documento PCT WO 2016/015058 A2). El patrón de fragmentación del ADN en el plasma está influenciado por el posicionamiento nucleosómico, los sitios de unión de factores de transcripción, los sitios de corte o hipersensibles a DNasas, los perfiles de expresión (Snyder *et al.* Cell 2016; 164: 57-68 y el documento PCT WO 2016/015058; Ivanov *et al.* BMC Genomics 2015; 16 Supl. 13:S1) y los perfiles de metilación del ADN (Lun *et al.* Clin Chem 2013; 59: 1583-1594) en el genoma de las células que han aportado las moléculas de ADN plasmático. Por lo tanto, los patrones de fragmentación son diferentes para las células de diferentes orígenes tisulares. Aunque hay regiones genómicas que muestran fragmentos más frecuentes, los lugares de corte del ADN plasmático real dentro de la región podrían seguir siendo aleatorios.

Los presentes inventores barajan la hipótesis de que los diferentes tejidos están asociados a la liberación de fragmentos de ADN plasmático que tienen diferentes sitios de corte o ubicaciones finales. En otras palabras, incluso los sitios de corte específicos no son aleatorios. De hecho, los presentes inventores han demostrado que las moléculas de ADN plasmático en pacientes con cáncer muestran ubicaciones finales diferentes a las de los pacientes sin cáncer. Algunas realizaciones pueden utilizar moléculas de ADN plasmático con dichas ubicaciones finales asociadas al cáncer como fragmentos de ADN canceroso informativos o utilizar dicha información de ubicación final como criterio de filtrado, por ejemplo, junto con uno o más criterios de filtrado diferentes. Por lo tanto, con la identificación de esas ubicaciones finales del ADN plasmático asociadas al cáncer, se podría puntuar el fragmento de ADN plasmático como un fragmento de ADN canceroso informativo o atribuir una ponderación diferencial basada en la naturaleza de la ubicación final de dicho fragmento. Estos criterios pueden utilizarse para evaluar la probabilidad de que los fragmentos tengan su origen en un cáncer, determinados órganos, o el cáncer de determinados órganos.

En consecuencia, la posibilidad de que un fragmento de ADN plasmático sea un fragmento de ADN canceroso informativo sería mucho mayor si muestra una supuesta mutación así como ubicaciones finales paraneoplásicas. Varias realizaciones pueden también tener en cuenta el estado de dicho fragmento y su longitud, o cualquier combinación de dichos y otros parámetros. Dado que un fragmento de ADN plasmático tiene dos extremos, se puede modificar aún más la ponderación para identificarlo como un fragmento derivado del cáncer teniendo en cuenta si uno o ambos extremos están asociados al cáncer o provienen de un tipo de tejido asociado al cáncer. El uso de un proceso de preparación de bibliotecas que aumente la probabilidad de que un fragmento de ADN monocatenario se convierta en la biblioteca de secuenciación potenciaría la eficacia de esta última realización (para un ejemplo de dicho proceso de preparación de bibliotecas, véase Snyder *et al.* Cell 2016; 164: 57-68), como se analiza en la siguiente sección. En una realización, puede utilizarse una estrategia similar basada en las ubicaciones finales para la detección de mutaciones asociadas con otras patologías o procesos biológicos (por ejemplo, mutaciones causadas por el proceso de envejecimiento o mutaciones causadas por factores mutagénicos ambientales).

También se puede utilizar una estrategia similar para identificar mutaciones *de novo* de un feto mediante la secuenciación del ADN en el plasma de una mujer gestante portadora del feto. Por lo tanto, tras la identificación de las ubicaciones finales que son específicas o relativamente específicas para la placenta, se puede atribuir una mayor ponderación a que una supuesta mutación fetal *de novo* sea verdadera si dicho fragmento de ADN en el plasma materno también porta una ubicación final específica de la placenta o enriquecida por la misma. Dado que un fragmento de ADN plasmático tiene dos extremos, se puede modificar aún más la ponderación para identificarlo como un fragmento derivado del feto teniendo en cuenta si uno o ambos extremos están asociados con la placenta.

A fin de ilustrar la viabilidad de esta estrategia, se analizaron los datos de secuenciación del ADN plasmático de un paciente con HCC y una mujer gestante. A efectos ilustrativos, el análisis se centró en el cromosoma 8. La misma estrategia se puede aplicar al genoma completo o cualquier otro cromosoma o cualquier región genómica o combinaciones de los mismos.

Se determinaron las coordenadas de los nucleótidos terminales en ambos extremos de cada fragmento de ADN plasmático secuenciado. Posteriormente, se contó el número de fragmentos que terminaban en cada nucleótido del cromosoma 8. Se determinó 1 millón de nucleótidos principales que tenían el mayor número de fragmentos de ADN

que terminaban en ellos para cada muestra de plasma del caso de HCC y la mujer gestante.

La figura 5 es un diagrama de Venn que muestra el número de ubicaciones de terminación frecuentes que son específicas para el caso HCC, específicas para la mujer gestante o compartidas por ambos casos según las realizaciones de la presente invención. Después se identificaron las coordenadas de los 463.228 nucleótidos que eran las posiciones finales frecuentes compartidas por los dos casos. Para el caso de HCC, los 463.228 nucleótidos compartidos se restaron del millón principal para obtener las coordenadas de los 536.772 nucleótidos que se identificó que eran las posiciones finales frecuentes específicas para el caso de HCC. De manera similar, los 463.228 nucleótidos compartidos se restaron de 1 millón de posiciones finales más comunes para el caso de gestación para obtener las coordenadas de los 536.772 nucleótidos que también se identificó que eran las posiciones finales frecuentes específicas para la mujer gestante.

Los fragmentos de ADN plasmático con nucleótidos terminales que terminan exactamente en las 536.772 posiciones finales específicas del HCC tendrían más probabilidades de proceder del tumor. En contraposición, los fragmentos de ADN plasmático con nucleótidos terminales que terminan exactamente en las posiciones finales específicas de la gestación o en las posiciones compartidas por los dos casos tendrían menos probabilidades de proceder del tumor, siendo las posiciones finales específicas de la gestación potencialmente menos probables y a las que se les da una menor ponderación en cualquier realización que utilice ponderaciones.

Por tanto, la lista de posiciones finales principales que son específicas para el caso del HCC puede utilizarse para seleccionar las mutaciones asociadas al cáncer, y la lista de posiciones finales principales que son específicas para el caso de la gestante o compartidas por ambos casos puede utilizarse para filtrar las mutaciones falsas positivas. Se puede utilizar un procedimiento similar para identificar las mutaciones fetales y filtrar las mutaciones falsas positivas para pruebas prenatales no invasivas.

En general, para identificar dichas ubicaciones finales del ADN plasmático biológicamente pertinentes, las muestras de ADN plasmático de grupos de individuos con diferentes enfermedades o antecedentes epidemiológicos o perfiles fisiológicos podrían compararse con muestras de otro grupo de individuos sin dichas enfermedades o antecedentes o perfiles. En una realización, cada una de estas muestras se pudo secuenciar en profundidad para poder identificar las posiciones finales comunes de los fragmentos de ADN plasmático dentro de cada muestra. En otra realización, los datos de la secuencia del grupo de personas con perfil complementario podrían agruparse para la identificación de ubicaciones finales comunes representativas de la enfermedad o del perfil fisiológico.

Un objetivo de este análisis es identificar las ubicaciones finales del ADN plasmático que son comunes a las personas con la enfermedad o perfil biológicamente pertinente, pero no en individuos sin la enfermedad ni perfil biológicamente pertinente. Por ejemplo, las comparaciones podrían implicar a personas con y sin cáncer, individuos con y sin cáncer de órganos o tejidos particulares, gestantes y no gestantes, gestantes con y sin determinadas enfermedades fetales o asociadas a la gestación e individuos de diferentes edades. Las ubicaciones finales del ADN plasmático específicas del tejido o pertinentes para la enfermedad después de haberse identificado en un grupo de muestras de referencia se convierten en el conjunto de referencia para la interpretación de las muestras de prueba.

Cada fragmento de ADN plasmático de una muestra podría interrogarse individualmente y se le asignaría una puntuación de probabilidad en función de la ubicación final. La puntuación de probabilidad para una determinada ubicación final puede depender de la separación en una cantidad de lecturas de secuencia (por ejemplo, un porcentaje de lecturas de secuencia u otro valor normalizado por la profundidad de la secuenciación en las muestras) que terminan en la ubicación final para los individuos diana (por ejemplo, el cáncer) en relación con la cantidad de lecturas de secuencia que terminan para el grupo de control. Una mayor separación conllevaría una mayor especificidad y, por tanto, se puede aplicar una mayor puntuación de probabilidad. Por tanto, podría realizarse la clasificación de fragmentos de ADN plasmático con ubicaciones finales específicas en probablemente asociado con la enfermedad o no, fetal o materno, etc.

Como alternativa, los fragmentos de ADN plasmático procedentes de la misma región podrían interpretarse colectivamente, en concreto, la frecuencia de finalización en un nucleótido concreto puede calcularse normalizando a la profundidad de secuenciación. De esta manera, se pueden identificar determinados nucleótidos como ubicaciones finales comunes en relación con otras ubicaciones del genoma, por ejemplo, solamente basándose en el análisis de una muestra de un tipo concreto, aunque pueden usarse más muestras. Por tanto, podría realizarse la clasificación de fragmentos de ADN plasmático con ubicaciones finales específicas en probablemente asociado con la enfermedad o no, fetal o materno, etc. Para los locus que muestran altas frecuencias de fragmentos de ADN plasmático con dichas ubicaciones finales de ADN plasmático biológicamente pertinentes, se podría determinar que dichos locus están enriquecidos con el ADN biológicamente pertinente y se incluirían como un grupo de fragmentos de ADN plasmático que tienen alta probabilidad como asociados al cáncer o específicos del feto o asociados a otras enfermedades o procesos biológicos. El nivel de probabilidad puede basarse en lo alta que sea la frecuencia de un determinado nucleótido en relación con otros nucleótidos, de manera similar a las comparaciones entre los distintos grupos, como se describe anteriormente.

A fin de ilustrar la eficacia de esta estrategia, las posibles mutaciones asociadas al cáncer se identificaron directamente

a partir de los datos de secuenciación del ADN plasmático del paciente con HCC. Los cambios de un solo nucleótido que estaban presentes en las lecturas de secuencia de al menos dos fragmentos de ADN plasmático se consideraron como posibles mutaciones asociadas al cáncer. También se secuenció el tejido tumoral y las mutaciones que estaban presentes en el tejido tumoral se consideraron verdaderas mutaciones asociadas al cáncer.

En el cromosoma 8, se identificó un total de 20.065 posibles mutaciones a partir de los datos de secuenciación del ADN plasmático del paciente con HCC sin utilizar el análisis de valor de corte dinámico. Una variante de secuencia se consideraría una posible mutación si la variante de secuencia estuviera presente en al menos dos fragmentos de ADN secuenciados. Se identificaron 884 mutaciones somáticas verdaderas a partir del resultado de la secuenciación del tejido tumoral. Las 20.065 supuestas mutaciones incluían 802 (91 %) de las 884 mutaciones reales. Por lo tanto, solo el 4 % de las supuestas mutaciones eran verdaderas mutaciones somáticas en el tejido tumoral, lo que dio un PPV del 4 %.

Para potenciar la precisión de la detección de las mutaciones somáticas, los presentes inventores utilizaron los siguientes algoritmos de filtrado basados en las posiciones de los nucleótidos terminales de las lecturas de secuencia portadoras de las supuestas mutaciones. (1). Para cualquier supuesta mutación, si hay al menos una lectura de secuencia portadora de la mutación y que termine en posiciones finales específicas de HCC, la mutación se calificaría para el análisis mutacional posterior. (2). Se eliminaría una lectura de secuencia que portase una supuesta mutación pero que terminara en cualquier posición final específica de la gestación o en las posiciones compartidas por ambos casos. Una mutación se calificaría para el análisis mutacional posterior solo si hubiera dos o más lecturas de secuencia que mostraran la misma mutación tras la eliminación de las lecturas basadas en este algoritmo.

Aplicando los algoritmos de filtrado 1 y 2 indicados anteriormente, se obtuvieron los resultados en la tabla 1. Los efectos de la aplicación de diferentes algoritmos de filtrado basándose en la posición de los nucleótidos terminales, o ubicaciones finales, de los fragmentos de ADN portadores de las supuestas mutaciones.

Tabla 1

	Sin filtro	Inclusión de mutaciones con extremos específicos de HCC (filtro 1)	Eliminación de lecturas con extremos compartidos o específicos de la gestación (filtro 2)	Aplicando ambos algoritmos de filtrado
N.º de supuestas mutaciones identificadas	20.065	1.526	2.823	484
Porcentaje de mutaciones verdaderas detectadas	91 %	29 %	88 %	40 %
PPV	4 %	17 %	28 %	71 %

Hubo una mejora sustancial en el PPV al adoptar uno cualquiera de los tres algoritmos que requerían que las ubicaciones finales fueran específicas del HCC o que el algoritmo filtrara las posiciones específicas de la gestación o las compartidas. Aplicando ambos algoritmos, el PPV aumentó al 71 %.

Se puede identificar otro número de ubicaciones finales asociadas al HCC y a la gestación para cada cromosoma, o incluso para otra región genómica, o incluso para todo el genoma, por ejemplo, pero sin limitación, 0,5 millones, 2 millones, 3 millones, 4 millones, 5 millones, 6 millones, 7 millones, 8 millones, 9 millones o 10 millones. En diversas realizaciones, se pueden determinar las ubicaciones finales más frecuentes en las moléculas de ADN plasmático en una o más cohortes de pacientes con cáncer, siendo cada cohorte de un tipo de cáncer. Además, las ubicaciones finales más frecuentes en las moléculas de ADN plasmático pueden determinarse para los sujetos sin cáncer. En una realización, estos pacientes con cáncer y los sujetos sin cáncer pueden subdividirse en grupos con diferentes parámetros clínicos, por ejemplo, el sexo, estado de tabaquismo, la salud previa (por ejemplo, el estado de hepatitis, diabetes, peso), etc.

Como parte de la utilización de estos criterios de filtrado, el análisis estadístico puede utilizarse para identificar las posiciones que tienen mayor probabilidad de ser nucleótidos terminales o ubicaciones finales del ADN circulante para diferentes condiciones fisiológicas y patológicas. Los ejemplos de los análisis estadísticos incluyen, pero sin limitación, la prueba de la t de Student, la prueba de Chi-cuadrado, y pruebas basadas en la distribución binomial o en la distribución de Poisson. Para estos análisis estadísticos, se pueden utilizar diferentes valores de corte de valores de p, por ejemplo, pero sin limitación, 0,05, 0,01, 0,005, 0,001 y 0,0001. Los valores de corte de valores de p también pueden ajustarse para las comparaciones múltiples.

G. Secuenciación monocatenaria

En una realización, la secuenciación se puede realizar en las dos cadenas complementarias de cada molécula plantilla denominada secuenciación de cadena sencilla (Snyder *et al.* Cell 2016; 164: 57-68). Las variaciones que están

presentes en las lecturas de secuenciación de ambas cadenas se utilizan para el análisis posterior, mientras que las variaciones que solo aparecen en la lectura de secuenciación de una cadena se descartan, o al menos se pueden descartar los datos de un fragmento de ADN. Esto puede reducir aún más exponencialmente los errores de secuenciación de las moléculas de ADN plasmático.

Debido a que cada cadena de los fragmentos de ADN plasmático podría analizarse de forma independiente, las coordinadas de las ubicaciones finales o de nucleótidos terminales de los fragmentos de ADN plasmático podrían determinarse con mayor precisión y exactitud. La secuenciación de cadena sencilla también permite la detección de fragmentos de ADN plasmático que circulan en forma monocatenaria en lugar de forma bicatenaria. Al incluir las moléculas de ADN plasmático monocatenario en el análisis (por ejemplo, mediante el uso de un protocolo de preparación de bibliotecas que facilitaría el análisis de ADN monocatenario (Snyder *et al.* Cell 2016; 164: 57-68)), una población adicional de fragmentos de ADN canceroso potencialmente informativos se vuelve susceptible de detección.

Asimismo, el uso de protocolos de preparación de bibliotecas que favorecen el ADN monocatenario (por ejemplo, véase Snyder *et al.* Cell 2016; 164: 57-68), también permitiría identificar ubicaciones adicionales que pueden usarse para el criterio de filtrado basado en la ubicación final. Por ejemplo, si después de las alineaciones de las dos lecturas de secuencia para las dos cadenas, las dos cadenas no se alinean en la misma ubicación final específica del tejido, entonces a la lectura de secuencia se le puede dar un peso más bajo como si tuviera una mutación.

VI. DETECCIÓN DE MUTACIONES SOMÁTICAS EN PLASMA DE PACIENTES CON CÁNCER

A continuación se describen varios ejemplos para la detección de mutaciones somáticas en sujetos sometidos a pruebas para detectar cáncer. Los datos se muestran para varios criterios de filtrado. Y, se ilustra la eficacia de los protocolos sin PCR.

A. Preparación de muestras

Se obtuvieron muestras clínicas de un paciente con HCC. Se recogió una muestra de sangre antes de la intervención. Se recogieron una biopsia de tumor de HCC y una biopsia del tejido hepático normal adyacente en el momento de la resección del tumor. Las bibliotecas de ADN se prepararon a partir de las muestras utilizando protocolos de preparación de bibliotecas sin PCR y se secuenciaron utilizando la serie HiSeq de Illumina de secuenciadores en paralelo masivos. Las profundidades de secuenciación conseguidas para la capa leucocitaria, biopsia tumoral, biopsia del tejido hepático normal adyacente y plasma fueron 45x, 45x, 40x y 220x del genoma haploide humano, respectivamente.

1. Información del paciente

El paciente con HCC era un hombre chino de 58 años, que era portador del HBV sin cirrosis. El tamaño del tumor era de 18 cm. Ingresó en el Departamento de Cirugía, del Prince of Wales Hospital para la resección del tumor, y se reclutó con consentimiento informado. El estudio fue aprobado por el Chinese University of Hong Kong and New Territories East Cluster Clinical Research Ethics Committee. Se recogieron 9 ml de sangre periférica en tubos con EDTA antes de la cirugía. El tejido tumoral y el tejido normal adyacente se recogieron después de la resección del tumor.

2. Procesamiento de muestras

Todas las muestras de sangre se procesaron mediante un protocolo de doble centrifugación (Chiu et al Clin Chem 2001; 37: 1607-1613). Brevemente, después de la centrifugación a 1.600 g durante 10 min a 4 °C, la porción de plasma se volvió a centrifugar a 16.000 g durante 10 min a 4 °C para eliminar las células sanguíneas. La porción de células sanguíneas se volvió a centrifugar a 2.500 g y se eliminó el plasma residual. El ADN de las células sanguíneas y el del plasma se extrajo con el protocolo de sangre y fluidos corporales del QIAamp DNA Blood Mini Kit y el QIAamp DSP DNA Blood Mini Kit (Qiagen), respectivamente (Qiagen). El ADN del tumor y de los tejidos normales adyacentes se extrajeron con el QIAamp DNA Mini Kit (Qiagen) según el protocolo de tejido del fabricante.

3. Cuantificación de ADN plasmático

El ADN se extrajo de 3,7 ml de plasma y se eluyó en 110 microlitros de agua. La concentración de ADN fue de 0,629 nanogramos por microlitro (fluorómetro Qubit, Thermo Fisher Scientific), produciendo 69 ng de ADN. Después se utilizaron 30 ng de ADN para la construcción de bibliotecas. Debido a que cada genoma de 3 Mb se divide en fragmentos de 166 pares de bases (pb), debe haber aproximadamente $1,81 \times 10^7$ fragmentos de ADN plasmático por genoma. Los 30 ng de ADN deben contener $[(30 \times 1.000)/3,3] \times 1,81 \times 10^7$ fragmentos = $1,64 \times 10^{11}$ fragmentos totales.

4. Construcción de bibliotecas de ADN

Las bibliotecas de ADN para las muestras de ADN genómico y la muestra de plasma materno se construyeron con el kit de preparación de bibliotecas sin PCR de ADN TruSeq (Illumina) según el protocolo del fabricante, excepto que se usó una quinta parte del adaptador indexado para la construcción de la biblioteca de ADN plasmático. Había tres

muestras de ADN genómico, en concreto, el ADN de la capa leucocitaria del paciente, el ADN del tejido tumoral y el ADN del tejido normal adyacente. Para cada muestra de ADN genómico, se trató con ultrasonidos un microgramo de ADN hasta obtener fragmentos de 200 pb (Covaris) para la construcción de bibliotecas. Las concentraciones de la biblioteca variaron de 17 a 51 nM en 20 µl de biblioteca.

Para la muestra de ADN plasmático de 30 ng (1,64 x 10^{11} fragmentos), el rendimiento de la biblioteca fue de 2242 pM en 20 µl de biblioteca, que equivalía a 44.854 átomos, es decir, 2,70 × 10^{10} fragmentos de ADN plasmático de 166 pb. La conversión de ADN a biblioteca fue del 16,4 %. Este nivel de conversión es mucho más alto que la experiencia previa de los presentes inventores con otros kits de preparación de bibliotecas de ADN en los que solo alrededor del 1 % del ADN de entrada podía convertirse en biblioteca.

5. Secuenciación de bibliotecas de ADN

Todas las bibliotecas de ADN se secuenciaron en las plataformas de secuenciación HiSeq 1500, HiSeq 2000 o HiSeq 2500 (Illumina) para 75 pb x 2 (extremo emparejado). Se secuenciaron múltiples carriles para cada biblioteca de ADN genómico. Las profundidades de secuenciación de las bibliotecas de ADN de la capa leucocitaria, del tejido tumoral y del tejido normal adyacente fueron 45x, 45x y 40x, respectivamente. Se secuenciaron 30,7 carriles para la biblioteca de ADN plasmático y se obtuvieron aproximadamente 4.400 millones de lecturas con extremos emparejados mapeadas no duplicadas. La profundidad de secuenciación fue de 220x.

Para calcular la recuperación de la biblioteca de ADN plasmático después de la secuenciación, se secuenciaron 120 µl de biblioteca de ADN a 10 pM por carril como entrada. El número total de fragmentos de entrada fue de 120 x 10 x 30,7 x 6,02 x 10^{23} / 10^{18} = 2,22 x 10^{10} fragmentos. Después de la secuenciación, se obtuvieron 4,40 x 10^{9} fragmentos. La recuperación de la biblioteca de ADN después de la secuenciación fue del 19,9 %.

Las secuencias de ADN plasmático se alinearon o mapearon con el genoma humano de referencia. El número de lecturas mapeadas a cada segmento de 1 Mb (bin) como proporción de todas las lecturas de secuencia se determinó en todo el genoma. Las proporciones o representaciones genómicas por segmentos de 1 Mb se compararon con datos de secuenciación de ADN plasmático obtenidos de un grupo de control sano para identificar regiones genómicas con un aumento o disminución estadísticamente significativos en las representaciones genómicas, como se describió anteriormente en la publicación de patente de Estados Unidos 2009/0029377.

La figura 6 es un gráfico 600 que muestra aumentos, disminuciones o ningún cambio en los segmentos de 1 Mb para el paciente con HCC. Las regiones con un aumento estadísticamente significativo en la representación genómica indican la presencia de una ganancia de número de copias, mientras que las regiones con una disminución estadísticamente significativa en la representación genómica indican la presencia de una pérdida de número de copias. Los bins con aumento, disminución estadísticamente significativos, o ningún cambio significativo en las representaciones genómicas se muestran como puntos verdes, rojos y grises, respectivamente. Al cuantificar el grado de pérdida de número de copias en segmentos genómicos consecutivos que mostraron dichas pérdidas (por ejemplo, como se describe en la solicitud de patente de Estados Unidos 14/994.023), se determinó que la concentración fraccionaria de ADN derivado de tumor en plasma era del 15 %.

B. Mutaciones presentes en biopsia tumoral y tejido adyacente

A continuación, se identificaron mutaciones somáticas presentes en la biopsia tumoral en comparación con los datos de secuenciación de la capa leucocitaria del paciente. Este análisis se realizó para determinar cuántas mutaciones somáticas tenía este tumor en particular y sirvió como el conjunto de mutaciones de referencia que se pretendía detectar en el ADN plasmático. Para cualquier alelo detectado en la biopsia tumoral pero no en el ADN de la capa leucocitaria, se aplicó una serie de criterios de filtrado para identificar las mutaciones somáticas. El análisis inicial se realizó en la mitad de los datos de secuencia, en concreto, 110x.

La figura 7 muestra un proceso 700 de filtrado, que utiliza valor de corte dinámico, realineación y fracción de mutación y los datos resultantes para las mutaciones identificadas a partir de una biopsia tumoral según las realizaciones de la presente invención. Tal como se muestra en la figura 7, primero se aplicó la estrategia de valor de corte dinámico para minimizar la detección de las variantes de un solo nucleótido falsas positivas, que son principalmente el resultado de errores de secuenciación. Los números que se muestran en cada recuadro representan el número de supuestas mutaciones identificadas en cada etapa.

A continuación, se aplicó la estrategia de realineación como criterio de filtrado de nivel A a las 16.027 supuestas mutaciones identificadas utilizando la estrategia de valor de corte dinámico para eliminar adicionalmente las variantes debidas a errores de secuenciación y errores de alineación. A continuación, se aplicaron dos valores de corte de concentración fraccionaria diferentes de forma independiente. Utilizando al menos una fracción de ADN tumoral del 20 % (M%) como valor de corte (criterio de nivel B), se identificaron 12.083 mutaciones somáticas. Utilizando al menos una fracción de ADN tumoral del 30 % como valor de corte (criterio de nivel C), se identificaron 11.903 mutaciones somáticas. Se consideran estas 11.903 variantes como las verdaderas mutaciones somáticas presentes en este tumor. El número es compatible con el número promedio informado de mutaciones presentes por tumor.

Se espera que las moléculas de ADN plasmático derivadas de tumores sean más cortas que las moléculas no derivadas de tumores. Como un medio para evaluar si estas variantes son verdaderas mutaciones somáticas derivadas de tumores, se buscaron fragmentos de ADN plasmático que cubrieran estos 11.903 locus y se evaluó el perfil de tamaño de estos fragmentos.

La figura 8 muestra un gráfico 800 de tamaños de fragmentos de ADN plasmático identificados con un alelo mutante para el paciente con HCC en comparación con los tamaños de fragmentos de ADN plasmático identificados con el alelo de tipo silvestre. Estos fragmentos de ADN plasmático identificados como portadores de una mutación son de hecho más cortos que aquellos otros fragmentos de ADN plasmático que no fueron informativos para estas mutaciones somáticas. Dicho análisis de tamaño confirma la eficacia de la identificación de las mutaciones y también confirma la capacidad de utilizar el tamaño como criterio de filtrado.

La figura 9 muestra un proceso 900 de filtrado, que utiliza valor de corte dinámico, realineación y fracción de mutación y los datos resultantes para las mutaciones identificadas a partir de una biopsia de hígado normal adyacente según las realizaciones de la presente invención. Se aplicó el mismo conjunto de criterios para cribar mutaciones en la biopsia de la biopsia de hígado normal adyacente, como se utiliza para la biopsia tumoral. Tal como se muestra en la figura 9, solamente se identificaron 203 mutaciones cuando el filtro final se basó en requerir al menos un 20 % de fracción de ADN tumoral (criterio de nivel B). Solo se identificaron 74 mutaciones cuando el filtro final se basó en requerir al menos un 30 % de fracción de ADN tumoral (criterio de nivel C).

Las figuras 10A y 10B muestran una comparación del perfil de tamaño evaluado de los fragmentos de ADN plasmático que portan las 203 supuestas mutaciones identificadas a partir de la biopsia de hígado normal adyacente con el perfil de tamaño de otras moléculas de ADN plasmático no informativas. La figura 10A muestra una frecuencia de fragmentos de ADN plasmático en un intervalo de tamaño para el supuesto alelo mutante y el alelo de tipo silvestre. La figura 10B muestra una frecuencia acumulativa de los fragmentos de ADN plasmático en función del tamaño para el supuesto alelo mutante y el alelo de tipo silvestre. Como se muestra en las figuras 10A y 10B, no hay diferencia en los perfiles de tamaño de los dos grupos de ADN expresados en forma de una curva de distribución de frecuencia de tamaño así como en los gráficos de diferencia de tamaño acumulado. El perfil de tamaño de estas moléculas sugiere que es probable que las variantes sean falsos positivos.

C. *Análisis mutacional del plasma*

A continuación, el objetivo de los presentes inventores fue aplicar varios criterios de filtrado para identificar mutaciones somáticas o fragmentos de ADN canceroso informativos en plasma.

La figura 11 muestra un proceso 1100 de filtrado (que utiliza valor de corte dinámico, realineación, fracción de mutación y tamaño), y los datos resultantes para las mutaciones identificadas a partir del plasma según las realizaciones de la presente invención. En la figura 11, el número de supuestas mutaciones somáticas se muestra en cada recuadro para cada etapa de filtrado. El número de mutaciones somáticas verdaderas recuperadas en cada etapa de filtrado, entre las 11.903 identificados a partir de la biopsia tumoral, se muestra como un número absoluto, así como un porcentaje. Los PPV para cada etapa de filtrado se calculan y también se muestran. Se pueden conseguir PPV de más del 85 % cuando se utilizaron los criterios de nivel B, C o D en combinación con el valor de corte dinámico y el filtrado de nivel A.

La figura 12 muestra un proceso 1200 de filtrado y los datos resultantes para las mutaciones identificadas a partir del plasma usando cortes dinámicos de fracciones mutantes inferiores según las realizaciones de la presente invención. El dato de la figura 12 muestra que el PPV se pudo mantener mientras que el número de mutaciones somáticas verdaderas recuperadas fue mucho mayor cuando se aplicaron valores de corte de concentración fraccionaria más bajos en el nivel B o el nivel C.

D. Tamaño

A continuación, se exploró el efecto de omitir los valores de corte de concentración fraccionarios (niveles B y C).

La figura 13 muestra un proceso 1300 de filtrado (que utiliza valor de corte dinámico, realineación y tamaño), y los datos resultantes para las mutaciones identificadas a partir del plasma según las realizaciones de la presente invención. Los datos mostrados en la figura 13 indican que se podría conseguir la misma recuperación y PPV con el uso del valor de corte dinámico, realineación y el requisito de tamaño (es decir, con una preferencia por moléculas de ADN cortas), como se consiguió al usar también el criterio de filtrado de fracción mutante.

La figura 14 muestra un gráfico 1400 de tamaños de fragmentos de ADN plasmático identificados con un alelo mutante utilizando plasma en comparación con los tamaños de fragmentos de ADN plasmático identificados con el alelo de tipo silvestre. Los perfiles de tamaño muestran que las mutaciones identificadas usando las etapas de filtrado mostraron un tamaño de ADN corto como se esperaba para el ADN derivado de tumor.

E. Aumento de la profundidad de secuenciación

Se aumentó adicionalmente la profundidad de secuenciación de la muestra de plasma de 110x a 220x.

La figura 15 muestra un proceso de filtrado 1500 y los datos resultantes para las mutaciones identificadas a partir del plasma utilizando una mayor profundidad de secuenciación según las realizaciones de la presente invención. El proceso 1500 usa el mismo conjunto de criterios de filtrado que el que se muestra en la figura 12. Con la profundidad de secuenciación aumentada (220x), la proporción de mutaciones somáticas verdaderas recuperadas fue mucho mayor. De las 10.915 mutaciones detectadas en la etapa de filtrado de nivel B, 93 mutaciones se ubicaron dentro de los exones. Únicamente una mutación, es decir, un cambio no sinónimo en el exón 3 de *CTNNB1* (c.C98G, P.S33C), se informó como una de las principales 28 mutaciones cancerosas prevalentes en la base de datos COSMIC.

F. Fracción mutante

La figura 11 mostró los efectos en el PPV y la tasa de recuperación cuando los valores de corte del nivel B y nivel C eran 20 % y 30 %, respectivamente. Se puede usar un M% más bajo como valor de corte si se prefiere una mayor sensibilidad en la identificación de mutaciones. La figura 12 muestra los efectos en el PPV y la tasa de recuperación cuando el valor de corte del nivel B era del 5 % y el valor de corte del nivel C era del 10 %.

Como se describe anteriormente, también se puede utilizar una varianza en la fracción mutante como criterio de filtrado. Se estudió la fracción alélica plasmática de la fracción mutante somática, procedentes de diferentes regiones cromosómicas. Tal como se muestra en la figura 6, el tumor del paciente con HCC demostró una pérdida de número de copias en el cromosoma 1p y una ganancia de número de copias en el cromosoma 1q. Se representó gráficamente la distribución de frecuencias de las fracciones mutantes en el cromosoma 1p y el cromosoma 1q.

La figura 16 es un gráfico 1600 que muestra el número (densidad) de locus que tienen varios valores de fracción mutante. Como se ve en el gráfico 1600, se observaron valores más altos de fracciones mutantes para la región de ganancia de número de copias (cromosoma 1q) y valores más bajos de fracciones mutantes para la región de pérdida de número de copias (cromosoma 1p).

También se estudió el intervalo de valores y la varianza de los valores de las fracciones mutantes en las dos regiones.

La figura 17A muestra puntuaciones z para la distribución sobre los brazos cromosómicos 1p y 1q. La figura 17B muestra la fracción mutante aparente sobre los brazos cromosómicos 1p y 1q. Las puntuaciones z de la distribución de valores fueron mayores (figura 17A) y los valores reales fueron más variables (figura 17B) en la región de ganancia de número de copias (cromosoma 1q) que en la región de pérdida de número de copias (cromosoma 1p).

Estos datos sugieren que se podrían establecer diferentes M% como valores de corte de filtrado para regiones con ganancias o pérdidas de número de copias para aumentar la probabilidad de identificar mutaciones somáticas verdaderas. Los valores de corte que especifican la varianza en la fracción mutante plasmática observada también podrían usarse para identificar moléculas de ADN plasmático que se han originado a partir de regiones cromosómicas que es más probable que se enriquezcan con (para regiones con aumento de número de copias) o se agoten de (para regiones con pérdidas de número de copias) ADN derivado de tumor. Entonces podría tomarse una decisión con respecto a la probabilidad de que el fragmento de ADN sea un fragmento de ADN canceroso informativo.

G. Criterios menos rigurosos

Se exploró si se pueden utilizar criterios menos rigurosos en el valor de corte dinámico. En los ejemplos mostrados anteriormente, el umbral de valor de corte dinámico (Score3) utilizado fue para minimizar el cambio de identificación de mutación somática falsa positiva. Para el análisis de valor de corte dinámico, una variante de secuencia se calificaría como mutación candidata cuando la variante de secuencia está presente en un número (N) de fragmentos de ADN secuenciados, donde el número (N) depende del número de locus secuenciados, el número de nucleótidos en el espacio de búsqueda y la probabilidad de tener la tasa de falsos positivos prevista. En el ejemplo anterior, la tasa de falsos positivos prevista se fijó en $<10^{-10}$, y el espacio de búsqueda es el genoma completo ($3x10^9$ nucleótidos).

La figura 18 es una tabla que 1800 muestra las sensibilidades de detección de mutaciones predichas para varias fracciones de mutación y profundidades de secuenciación para determinados cortes dinámicos de recuento de alelos según las realizaciones de la presente invención. Cada fila corresponde a una profundidad de secuenciación diferente. El valor de corte en plasma se utiliza para determinar si el número de fragmentos de ADN con la mutación en plasma es suficiente para considerarse como una mutación. Usando estos valores, las columnas restantes proporcionan la sensibilidad prevista, TP/(TP+FN), de detección de mutaciones en plasma para diversos porcentajes tumorales. La capa leucocitaria también se somete a un valor de corte para filtrar los errores de secuenciación en la capa leucocitaria. Sin ese filtro, las realizaciones pueden pasar por alto la inclusión del locus como un sitio homocigoto para la detección de variantes en plasma, ya que algunas realizaciones solamente detectan variantes que se encuentran en ubicaciones donde la capa leucocitaria es homocigota. Los datos de la tabla 1800 sirven como datos de referencia para interpretar el siguiente gráfico cuando se utilizan valores de corte dinámicos menos rigurosos.

Se exploraron los efectos de relajar el umbral para permitir una tasa de detección de falsos positivos del 0,1 %.

La figura 19 es una tabla 1900 que muestra las sensibilidades predichas de detección de mutaciones para varias fracciones de mutación y profundidades de secuenciación para determinados cortes dinámicos de recuentos de alelos para una tasa de detección de falsos positivos del 0,1 % según las realizaciones de la presente invención. Estos datos muestran datos para un valor de corte dinámico menos riguroso.

La figura 20 muestra un proceso 2000 de filtrado y los datos resultantes para las mutaciones identificadas a partir del plasma usando valores de corte dinámicos menos rigurosos según las realizaciones de la presente invención. Se utilizó una profundidad de secuenciación de 220x. Cuando se usó el valor de corte dinámico menos riguroso, el PPV en la primera etapa se redujo del 12 % al 3,3 %. Cuando se combina con las otras etapas de filtrado, es decir, los niveles A, B, C y D, se podría conseguir una mayor recuperación de las mutaciones somáticas verdaderas con PPV similares a los algoritmos basados en valores de corte dinámicos rigurosos.

Estos datos sugieren que cada criterio de filtrado desempeña un papel diferente. La utilidad de cada criterio podría cambiarse modificando la rigurosidad de los umbrales utilizados. En este ejemplo, el valor de corte dinámico menos riguroso permitió la identificación más sensible de mutaciones somáticas. La especificidad del esquema general se mantuvo debido a la eficacia de los otros criterios para filtrar los falsos positivos.

A continuación, se evaluó adicionalmente la eliminación completa de la etapa de valor de corte dinámico. En cambio, se aplicaron valores de corte fijos. Por ejemplo, se determinó el número de supuestas mutaciones identificadas si un alelo heterocigoto no presente en el ADN de la capa leucocitaria se ve al menos un número específico de veces (por ejemplo, 1, 2, 3, etc.) en plasma. Se aplicó este análisis para analizar los datos de ADN plasmático del paciente con HCC, así como una muestra de plasma materno secuenciada a más de 200x. No se conocía que madre que aportó la muestra de plasma materno tuviera cáncer y, por lo tanto, es probable que la mayoría de las supuestas mutaciones identificadas en esta muestra sean alelos específicos del feto heredados del padre o falsos positivos.

La figura 21 es un gráfico 2100 que muestra las distribuciones del número de supuestas mutaciones para escenarios fetales y de cáncer. El eje vertical corresponde a un recuento del número de locus con una supuesta mutación (alelo mutante). El eje horizontal corresponde al número de fragmentos de ADN necesarios para que se identifique que un locus tiene una mutación.

Ambas muestras se secuenciaron a una profundidad similar utilizando protocolos de preparación de bibliotecas sin PCR. Por lo tanto, las mutaciones falsas positivas aportadas por los errores de secuenciación y los errores de alineación deben ser similares en ambas muestras. Se observa que el número de supuestas mutaciones disminuyó a medida que aumentó el número de lecturas de secuencia utilizadas como valor de corte para la puntuación de una mutación. Debido a que las mutaciones falsas positivas tienden a producirse al azar y, por lo tanto, están presentes en proporciones alélicas más bajas, es probable que los falsos positivos se estén filtrando con el aumento progresivo del número de lecturas requeridas como valor de corte.

Por otro lado, se pudo observar que el número de supuestas mutaciones identificadas en el paciente con cáncer comenzaba a demarcarse y era mayor que el detectado en el plasma de la mujer gestante a partir de un valor de corte de aproximadamente 18 lecturas de secuencia y en adelante. Esto significa que la carga mutacional en el paciente con HCC es mayor que el número de alelos fetales heredados por vía paterna en la muestra de plasma materno.

A continuación, se aplicaron los criterios de filtrado de realineación (nivel A) al mismo conjunto de datos.

La figura 22 es un gráfico 2200 que muestra las distribuciones del número de supuestas mutaciones para escenarios fetales y de cáncer cuando se usa realineación. El número total de supuestas mutaciones disminuyó sustancialmente incluso en los números de valores de corte de lecturas de secuencia fijos correspondientes en comparación con los datos mostrados en la figura 21 cuando no se aplicó la realineación. La demarcación en el número de supuestas mutaciones entre el plasma de HCC y el plasma materno fue aún más obvia. Estos datos sugieren que la etapa de realineación es un proceso poderoso para la eliminación de falsos positivos.

Además, se evaluó el valor del filtrado por tamaño. De nuevo, la estrategia de valor de corte dinámico no se utiliza en este análisis. En cambio, se usó un número mínimo fijo de lecturas de secuencia que mostraban el mismo alelo menor como primera etapa para identificar supuestas mutaciones.

La figura 23 es una tabla 2300 que muestra los PPV y las tasas de recuperación para cortes dinámicos de varios tamaños sin realineación según las realizaciones de la presente invención. Tal como se muestra en la figura 23, los PPV para la identificación de mutaciones somáticas usando los valores de corte fijos solos fueron subóptimos. Cuando se utilizaron valores de corte de diferentes tamaños en cada nivel de valor de corte fijo, los PPV mejoraron.

La figura 24 es una tabla 2400 que muestra los PPV y las tasas de recuperación para cortes dinámicos de varios tamaños con realineación según las realizaciones de la presente invención. Para los datos mostrados en la figura 24,

la realineación se aplicó después de la identificación inicial de supuestas mutaciones mediante los valores de corte fijos. Los PPV mejoraron sustancialmente. Después se aplicaron valores de corte de diferentes tamaños para un filtrado adicional, se observó alguna mejora en el PPV.

H. Detección de carga mutacional elevada en cáncer

Se realizó una evaluación de la carga mutacional utilizando el criterio de filtrado descrito para la muestra de plasma del paciente con HCC y el plasma de una muestra de sangre del cordón umbilical de un recién nacido. El genoma constitutivo para la muestra de sangre del cordón umbilical fue la capa leucocitaria de la sangre del cordón umbilical. El plasma de sangre de cordón umbilical funciona bien como control ya que la mayoría de los bebés nacen sin cáncer y aún no han adquirido mutaciones somáticas ni han estado expuestos a carcinógenos.

El plasma de sangre del cordón umbilical se secuenció a 75x utilizando un protocolo de preparación de bibliotecas sin PCR.

La figura 25 muestra un proceso 2500 de filtrado (que utiliza valor de corte dinámico, realineación y tamaño), y los datos resultantes para las mutaciones identificadas a partir del plasma de sangre del cordón umbilical según las realizaciones de la presente invención. La figura 25 muestra el número de supuestas mutaciones detectadas en el plasma de la sangre del cordón umbilical cuando se utilizó un valor de corte dinámico riguroso seguido de los criterios de los niveles A a D que se muestran en la figura. Se identificó un pequeño número de supuestas mutaciones.

La figura 26 es un gráfico 2600 de distribuciones de tamaño para fragmentos de ADN mutantes determinados a partir del proceso 2500 y alelos de tipo silvestre según las realizaciones de la presente invención. Cuando se evaluó el perfil de tamaño de estas mutaciones, no eran particularmente cortas, a diferencia del ADN derivado del cáncer.

A continuación, se seleccionaron aleatoriamente 75x de datos de secuencia de ADN plasmático de la muestra de HCC para poder realizar una evaluación comparable. Se aplicó el mismo conjunto de criterios de filtrado. De aproximadamente 5.000 a 6.000 de las mutaciones derivadas de tumor se recuperaron con PPV del 89 % o más.

La figura 27 muestra un proceso 2700 de filtrado (que utiliza valor de corte dinámico, realineación y tamaño), y los datos resultantes para las mutaciones identificadas a partir del plasma de una muestra de HCC según las realizaciones de la presente invención. Se utilizó una profundidad de secuenciación de 75x.

La figura 28 es un gráfico 2800 de distribuciones de tamaño para fragmentos de ADN mutantes determinados a partir del proceso 2700 y alelos de tipo silvestre según las realizaciones de la presente invención. Los fragmentos de ADN plasmático con estas mutaciones eran, de hecho, más cortos que los fragmentos de ADN no informativos.

Sin embargo, se observó que el 84 % de las supuestas mutaciones identificadas en el plasma de la sangre del cordón umbilical se produjeron en sitios de polimorfismo de un solo nucleótido informados públicamente, mientras que esta proporción fue solamente del 3 % en la muestra de plasma de HCC. Por lo tanto, se plantea la hipótesis de que los alelos informados públicamente en el plasma de la sangre del cordón umbilical pueden ser moléculas de ADN materno que han ingresado en la circulación fetal y permanecieron detectables en la sangre neonatal (Lo *et al.* Clin Chem 2000; 46:1301-1309). Después de eliminar cualquier sitio de los sitios conocidos de polimorfismo de un solo nucleótido, el número de supuestas mutaciones en el plasma de sangre de cordón umbilical disminuyó a solo 8 (figura 29) mientras que los datos para el plasma de HCC permanecieron prácticamente sin cambios (figura 30).

La figura 29 muestra un proceso 2900 de filtrado que usa filtrado basado en SNP para mutaciones identificadas a partir de plasma de sangre de cordón umbilical según las realizaciones de la presente invención. La figura 30 muestra un proceso 3000 de filtrado que usa filtrado basado en SNP para mutaciones identificadas a partir de plasma de HCC según las realizaciones de la presente invención. La incorporación de una etapa de filtrado para eliminar polimorfismos de un solo nucleótido corresponde al filtrado de nivel E. Por consiguiente, el número de supuestas mutaciones (que en su mayoría son falsos positivos) detectadas en el plasma de la sangre del cordón umbilical se redujo en un 84 % (8 de cada 49). Por otro lado, el número de supuestas mutaciones en la muestra de HCC solo se ha reducido en un 3 %.

Los datos de los presentes inventores muestran que al usar el protocolo de preparación de bibliotecas sin PCR seguido de una secuenciación ultraprofunda y amplia con la incorporación del conjunto descrito de criterios de filtrado, se pudo identificar de manera sensible y específica mutaciones derivadas de tumores en el plasma de un paciente con cáncer en función de la cantidad de supuestas mutaciones identificadas. La carga mutacional identificada en el plasma del paciente con cáncer superó la observada en el plasma de sangre del cordón umbilical no canceroso de control en 3 órdenes de magnitud. Por lo tanto, se podría hacer la clasificación entre cáncer y no cáncer.

Además, se demostró que una submuestra (75x) del total de datos secuenciados (220x) ya era adecuada para conseguir la discriminación entre cáncer y no cáncer. Como se muestra en los datos de simulación a continuación (figuras. 44, 45A-45C y 46A-46C de la sección VIII), mientras que en estas realizaciones se necesitan datos de secuencia ultra profundos y amplios, el grado de amplitud y profundidad depende de la fracción de ADN tumoral en la

muestra de plasma y del número de mutaciones albergadas por el tumor que son susceptibles de detección en ADN plasmático.

I. Tejido de Origen

Existen datos (Snyder *et al.* Cell 2016; 164: 57-68; documento PCT WO 2016/015058 A2; Ivanov *et al.* BMC Genomics 2015; 16 Suppl 13:S1) para sugerir que la ubicación genómica de dichas mutaciones somáticas puede mostrar patrones de agrupamiento dependiendo del tejido de origen del tumor. La bibliografía sugería que las mutaciones somáticas tendían a localizarse junto con ubicaciones genómicas con modificaciones específicas de histonas. Las ubicaciones específicas de tejido de las modificaciones de histonas podrían obtenerse a través de bases de datos públicas como la base de datos Epigenomics Roadmap (www.roadmapepigenomics.org).

Se obtuvieron las ubicaciones específicas de tejido de las modificaciones de histonas a través de la base de datos Epigenomics Roadmap (www.roadmapepigenomics.org). En tejidos sanos, se informa que H3K4me1 está asociado con regiones potenciadoras activas/equilibradas. H3K27ac está asociado con regiones potenciadoras activas. H3K9me3 está altamente correlacionado con la heterocromatina constitutiva. En otras palabras, en tejidos sanos, H3K4me1 y H3K27ac están asociados con regiones genómicas con expresión génica activa en el tejido, mientras que H3K9me3 está asociado con regiones reprimidas del genoma. Sin embargo, se ha informado en el cáncer que el número de mutaciones somáticas está más representado en las regiones genómicas reprimidas. Ningún dato hasta la fecha ha informado de la existencia de dicha correlación en ADN plasmático.

Se realizó un análisis de correlación de Spearman entre el número de cada una de las tres modificaciones de histonas por bin de 1 Mb y el número de mutaciones somáticas en el mismo bin de 10 Mb.

La figura 31 es una tabla 3100 que muestra correlaciones de tejido con modificaciones de histonas. La figura 31 utiliza SNV para determinar el tejido de origen de la predicción del tumor. El coeficiente de correlación más fuerte se obtuvo para el patrón de modificación de histonas del tejido hepático. Esto es coherente con el hecho de que los datos de ADN plasmático se obtuvieron de un paciente con HCC. Por lo tanto, si se analiza otra muestra de prueba, podrían identificarse fragmentos de ADN plasmático que se originan a partir de locus que están asociados con modificaciones de histonas que se sabe que están asociadas con cáncer. Dichos locus estarían enriquecidos con fragmentos de ADN plasmático derivados de cáncer. Por lo tanto, los fragmentos de ADN plasmático de estos locus podrían clasificarse como fragmentos de ADN canceroso informativos. También se puede realizar un estrategia similar para identificar mutaciones fetales utilizando modificaciones de histonas que se sabe que están asociadas con tejidos fetales (por ejemplo, la placenta).

La correlación de Spearman se calcula entre la densidad de SNV por megabase en plasma y la densidad de marcadores de histonas por megabase en varios órganos o tejidos. La correlación más alta sugeriría el tejido de origen del tumor.

VII. DETECCIÓN DE MUTACIÓN DE NOVO EN FETOS

La mayor parte del análisis anterior se ha relacionado con el cáncer, pero las realizaciones también se pueden usar para identificar mutaciones *de novo* en fetos.

Las mutaciones congénitas pueden dar como resultado enfermedades que pueden manifestarse durante el período prenatal, durante la niñez o más adelante en la vida. Las mutaciones congénitas se refieren a mutaciones que están presentes en el genoma fetal. Algunas enfermedades son susceptibles de tratamiento temprano, mientras que otras pueden estar asociadas con un deterioro significativo de la función. Por lo tanto, el diagnóstico prenatal de algunas de estas enfermedades está justificado. Podría realizarse diagnóstico prenatal de enfermedades asociadas con anomalías genéticas, genómicas o cromosómicas analizando el material genético fetal antes del nacimiento. El material genético fetal podría obtenerse mediante procedimientos invasivos, tal como la amniocentesis o la muestra de vellosidades coriónicas. Estos procedimientos están asociados con riesgos de aborto espontáneo fetal. Por lo tanto, se prefiere realizar una evaluación prenatal mediante estrategias no invasivas, incluso a través del análisis de ácidos nucleicos fetales sin células que están presentes en el plasma materno.

La mayoría de las mutaciones congénitas se heredan de los padres y dan lugar a enfermedades hereditarias. Anteriormente se informaron estrategias para la detección no invasiva de mutaciones heredadas mediante análisis de ADN fetal sin células circulante en plasma materno (publicaciones de patentes de Estados Unidos 2009/0087847 y 2011/0105353). Las supuestas mutaciones fetales podrían confirmarse conociendo o probando las mutaciones maternas y/o paternas.

Sin embargo, las enfermedades también están causadas por mutaciones *de novo*. Las mutaciones *de novo* son mutaciones presentes en el genoma constitutivo de un feto que no se heredan del padre o de la madre. Las mutaciones *de novo* representan una proporción significativa de la carga de enfermedad para determinadas enfermedades, por ejemplo, acondroplasia, neoplasia endocrina múltiple. Se ha estimado que cada persona tiene entre 20 y 30 mutaciones *de novo* en el genoma constitutivo (Kong *et al.* Nature 2012; 488: 471-475). Dichas mutaciones pueden

causar enfermedades si se producen en regiones del genoma que afectarían a la función genética, epigenética o reguladora del genoma. Actualmente no existe un método eficaz para la detección prenatal de mutaciones *de novo* a menos que exista un riesgo conocido a priori. Una sospecha a priori de una mutación *de novo* podría surgir si, por ejemplo, una ecografía del feto revela características sospechosas de acondroplasia. Si ninguno de los padres tiene mutaciones para acondroplasia, entonces se buscará una mutación *de novo* en el gen del *receptor 3 del factor de crecimiento de fibroblastos*.

Para la mayoría de las otras enfermedades causadas por mutaciones *de novo*, normalmente, no hay signos estructurales o físicos que puedan detectarse prenatalmente para sugerir qué gen investigar. Actualmente no existe un método eficaz para detectar mutaciones *de novo* prenatalmente porque la búsqueda de 30 de esos cambios dentro de los 3 mil millones de nucleótidos del genoma del haplotipo es como buscar una aguja en un pajar. Conseguir la detección de mutaciones *de novo* mediante el análisis de ADN fetal sin células circulante se asocia con una dificultad mucho mayor debido al ADN plasmático de fondo de la madre que diluye adicionalmente las mutaciones *de novo* fetales de 5 a 10 veces. En el presente caso se describen realizaciones que permitirían la detección eficaz de mutaciones *de novo* fetales mediante el análisis del ADN fetal sin células circulante en el plasma materno.

A. Ejemplo para la detección de una mutación de novo en el feto

1. Información familiar

Gestación única con feto masculino programado para cesárea en la semana 38ª de gestación. La familia se reclutó en el departamento de obstetricia y ginecología, del Prince of Wales Hospital con consentimiento informado. El estudio fue aprobado por el Chinese University of Hong Kong and New Territories East Cluster Clinical Research Ethics Committee. Durante el ingreso se recogieron 20 ml de sangre materna y 10 ml de sangre paterna. Se recogieron muestras de tejido placentario y 3 ml de sangre de cordón umbilical después del parto.

2. Procesamiento de muestras

Todas las muestras de sangre se procesaron mediante un protocolo de doble centrifugación como se describe anteriormente (Chiu et al Clin Chem 2001; 37: 1607-1613). Brevemente, después de la centrifugación a 1.600 g durante 10 min a 4 °C, la porción de plasma se volvió a centrifugar a 16.000 g durante 10 min a 4 °C para eliminar las células sanguíneas. La porción de células sanguíneas se volvió a centrifugar a 2.500 g y se eliminó el plasma residual. El ADN de las células sanguíneas y el del plasma materno se extrajo con el protocolo de sangre y fluidos corporales del QIAamp DNA Blood Mini Kit y el QIAamp DSP DNA Blood Mini Kit (Qiagen), respectivamente (Qiagen). El ADN de la placenta se extrajo con el QIAamp DNA Mini Kit (Qiagen) según el protocolo de tejidos del fabricante.

3. Cuantificación de ADN plasmático

El ADN se extrajo de 5 ml de plasma materno. Utilizando el ensayo de PCR digital *ZFX/Y* (Lun et al Clin Chem 2008; 54: 1664-1672), la concentración de *ZFX* y *ZFY* fue de 1.038 copias/ml de plasma y de 103 copias/ml de plasma, respectivamente. Después se utilizaron 4,5 ml de equivalentes de ADN plasmático para la construcción de bibliotecas. Suponiendo que cada genoma se divide en fragmentos de 166 pares de bases (pb), debe haber aproximadamente $1{,}81 \times 10^7$ fragmentos de ADN plasmático por genoma. Los 4,5 ml de ADN plasmático deben contener (1038+103) x $4{,}5 \times 1{,}81 \times 10^7$ fragmentos = $9{,}28 \times 10^{10}$ fragmentos totales.

4. Construcción de bibliotecas de ADN

Las bibliotecas de ADN para las muestras de ADN genómico y la muestra de plasma materno se construyeron con el kit de preparación de bibliotecas sin PCR de ADN TruSeq (Illumina) según el protocolo del fabricante, excepto que se usó una quinta parte del adaptador indexado para la construcción de la biblioteca de ADN plasmático. Había cuatro muestras de ADN genómico, en concreto, el ADN de la capa leucocitaria de la madre, el ADN de la capa leucocitaria del padre, el ADN de la capa leucocitaria de la sangre del cordón umbilical y el ADN de la placenta. Para cada muestra de ADN genómico, se trató con ultrasonidos un microgramo de ADN hasta obtener fragmentos de 200 pb (Covaris) para la construcción de bibliotecas. Las concentraciones de la biblioteca variaron de 34 a 58 nM en 20 µl de biblioteca. Para la muestra de ADN plasmático materno de 4,5 ml de plasma ($9{,}28 \times 10^{10}$ fragmentos), el rendimiento de la biblioteca fue de 2995 pM en 20 µl de biblioteca, que equivalía a 59.910 moles, es decir, $3{,}61 \times 10^{10}$ fragmentos de ADN plasmático de 166 pb. La conversión de ADN a biblioteca fue del 38,9 %.

5. Secuenciación de bibliotecas de ADN

Todas las bibliotecas de ADN se secuenciaron en las plataformas de secuenciación HiSeq 1500, HiSeq 2000 o HiSeq 2500 (Illumina) para 75 pb x 2 (extremo emparejado). Se secuenciaron múltiples carriles para cada biblioteca de ADN genómico. Las profundidades de secuenciación de las bibliotecas de AND materno, paterno, del cordón umbilical y de la placenta fueron 40x, 45x, 50x y 30x, respectivamente. Toda la biblioteca de ADN plasmático materno se utilizó para la secuenciación. Se agotó la biblioteca con 45 carriles y se obtuvieron aproximadamente 5.740 millones de lecturas de extremos emparejados mapeadas no duplicadas. La profundidad de secuenciación fue de ~255x.

Para calcular la recuperación de la biblioteca de ADN plasmático, se utilizaron 16 µl de biblioteca de ADN a 2.995 nM como entrada (se usaron 4 µl de la biblioteca de ADN de 20 µl para la validación y cuantificación de la biblioteca). El número total de fragmentos de entrada fue de 2.995 x 16 x 6,02 x 10^{23} / 10^{9} = 2,89 x 10^{10} fragmentos. Después de la secuenciación, se obtuvieron 5,74 x 10^{9} lecturas (fragmentos). La recuperación de la biblioteca de ADN después de la secuenciación fue del 19,9 %. El 80 % de la biblioteca de entrada se perdió durante la generación y/o secuenciación de grupos. Se sospechaba que se requeriría un exceso de biblioteca de 5 veces como entrada para conseguir una alta eficacia de generación de grupos en la celda de flujo de secuenciación. El exceso de fragmentos de la biblioteca se lavaría y solo se secuenciarían los que formaran un grupo.

Siguiendo la estimación anterior, la tasa de conversión de ADN a biblioteca fue del 38,9 % y la recuperación de la biblioteca de ADN después de la secuenciación fue del 19,9 %. Se estimó que desde los fragmentos de ADN plasmático hasta los fragmentos de salida de secuenciación, la recuperación fue del 7,7 %.

B. Análisis

Se identificaron 298.364 sitios con SNP informativos donde el padre y la madre eran ambos homocigotos, pero con un alelo diferente. Por lo tanto, el feto era un heterocigoto obligado en estos sitios. Se confirmó que el 99,8 % de estos sitios con SNP eran heterocigotos en el tejido de la placenta. A continuación se determinó la fracción de ADN fetal en el plasma materno. Combinando los recuentos de los alelos paternos y expresando esto como una proporción de los recuentos combinados de los alelos maternos en estos 298.364 sitios con SNP informativos, la fracción de ADN fetal se estimó en un 31,8 %. A continuación, se determinó la fracción fetal en cada uno de estos sitios con SNP informativos.

La figura 32 muestra la distribución de frecuencias de las fracciones fetales medidas en dichos sitios con SNP individuales. El 95 % de los sitios presentan una fracción de ADN fetal superior al 20 %.

La figura 33A muestra una distribución de tamaño de ADN específico fetal y ADN compartido en plasma materno. La figura 33B muestra un gráfico de frecuencias acumulativas para el tamaño del ADN plasmático para el fragmento de ADN compartido y específico fetal. La figura 33C muestra la diferencia en frecuencias acumulativas, denominada ΔF. De manera similar a las observaciones informadas previamente (Lo *et al.* Sci Transl Med 2010; 2: 61ra91), las moléculas de ADN fetal en el plasma materno muestran un tamaño más pequeño que las moléculas de ADN plasmático no específico del feto.

Para determinar las mutaciones *de novo* presentes en el genoma de este feto, se buscaron variantes de ADN, en su mayoría mutaciones puntuales o variantes de un solo nucleótido, que estuvieran presentes tanto en el ADN de la placenta como en el ADN de la sangre del cordón umbilical, pero no en el ADN genómico materno ni en el ADN genómico paterno. Se identificaron cuarenta y siete tales de dichos sitios mutantes *de novo*. Después se buscaron moléculas de ADN que presentaran el alelo mutante *de novo* en plasma materno. Después se estudió la distribución de tamaño de las moléculas de ADN en el plasma materno.

La figura 34A muestra la distribución de tamaño de los fragmentos de ADN plasmático con el alelo mutante. La figura 34B muestra un gráfico de frecuencias acumulativas para el tamaño del ADN plasmático para el alelo mutante y el alelo de tipo silvestre. La figura 34C muestra la diferencia en frecuencias acumulativas, denominada ΔF. Los perfiles de tamaño y los valores de ΔF de los alelos mutantes mostraron un gran parecido con los valores derivados de los alelos específicos del feto (figuras 33A-33C). Su tamaño relativamente corto en el plasma materno brinda evidencia de que aquellas moléculas de ADN con el alelo mutante son de origen fetal.

A continuación, se estudió la eficacia de la estrategia de los presentes inventores para identificar mutaciones *de novo* a partir de los datos de ADN plasmático materno. En esta estrategia, se necesitaría obtener la información de la secuencia genómica materna y paterna. Después se buscaron variantes presentes entre las moléculas de ADN plasmático materno pero no en las secuencias de ADN genómico materno y paterno.

La figura 35 muestra un proceso 3500 de filtrado (que utiliza valor de corte dinámico, realineación, fracción de mutación y valor de corte por tamaño) y los datos resultantes para mutaciones *de novo* identificadas a partir de plasma según las realizaciones de la presente invención. El proceso de filtrado 3500 se puede utilizar para identificar las mutaciones *de novo* a partir de datos de ADN plasmático materno sin células. En este estudio, se utilizaron datos de secuenciación de ADN plasmático de genoma completo generados utilizando un protocolo de preparación de bibliotecas sin PCR.

En primer lugar, se utilizó un valor de corte dinámico para cribar las supuestas mutaciones en plasma. Los valores de corte dinámicos se utilizaron para controlar las apariciones teóricas de falsos positivos en el genoma humano por debajo de cierto nivel, por ejemplo, una vez por genoma. En este modelo de valores de corte dinámicos se pueden tener en cuenta dos tipos de fuentes atribuidas a falsos positivos. Una fuente serían los errores de secuenciación que, por casualidad, harían que algunos sitios mostraran el mismo cambio de nucleótido en la misma posición. La probabilidad de este tipo de falso positivo se puede estimar según la regla de probabilidad de la multiplicación para una tasa de error de secuenciación dada. El error de secuenciación se puede deducir de los sitios en los que tanto la

madre como el padre eran homocigóticos y poseían la misma información alélica. En este caso, el error de secuenciación se estimó en un 0,3 %. Otra fuente serían los SNP heterocigotos en la madre o el padre, que se denominaron erróneamente homocigotos debido al muestreo insuficiente de alelos alternativos.

En segundo lugar, para minimizar aún más los errores de secuenciación y alineación en los datos de secuenciación reales, se aplicó un algoritmo de filtrado adicional. Las lecturas de secuenciación portadoras de las mutaciones se realinearían (mapearían) con genoma humano de referencia mediante el uso de un alineador independiente, por ejemplo, Bowtie2 (Langmead *et al.* Nat Methods 2012; 9: 357-9). En algunas realizaciones, los siguientes criterios de realineación se pueden usar para identificar una lectura mapeada como una lectura de secuencia de baja calidad: (1) la lectura de secuencia portadora de la mutación no se puede recuperar mediante un alineador independiente; (2) la lectura de secuencia portadora de la mutación muestra resultados de mapeo inconsistentes cuando se utiliza un alineador independiente para verificar la alineación original (por ejemplo, una lectura mapeada se coloca en un cromosoma diferente en comparación con el resultado de la alineación original). (3) la lectura de la secuencia portadora de la mutación alineada con la misma coordenada genómica muestra una calidad de mapeo ≤ Q20 (es decir, probabilidad de alineación errónea <1 %); (4) la lectura de secuencia tiene la mutación ubicada dentro de las 5 pb de cualquiera de los extremos de lectura (es decir, extremos 5' o 3'). Esta última regla de filtrado puede ser importante porque los errores de secuenciación se producen con más frecuencia en ambos extremos de una lectura de secuencia. Si la proporción de lecturas de secuencias de baja calidad entre las lecturas de secuencias portadoras de la mutación es mayor que determinado umbral, por ejemplo, 40 %, se descartarán los sitios mutantes candidatos. Esta etapa de realineación de las lecturas de secuenciación portadoras de la mutación se denomina criterio de filtrado de nivel A.

En tercer lugar, solamente la fracción mutante (M%) que exceda un cierto umbral se consideraría como una mutación verdadera más probable, por ejemplo, 20 % (criterios de filtrado de nivel B) y 30 % (criterios de filtrado de nivel C). La fracción de ADN fetal estimada a partir de SNP informativos se puede utilizar como referencia para establecer un umbral adecuado de fracción mutante.

En cuarto lugar, debido a que las moléculas de ADN derivadas del feto son más cortas que las moléculas de ADN derivadas de la madre, los presentes inventores han creado adicionalmente un parámetro de filtrado asociado al tamaño en los criterios de filtrado de nivel D. Se requiere una diferencia mínima en los tamaños medianos entre los fragmentos de ADN portadores de alelos mutantes y los alelos de tipo silvestre para tener al menos determinados pares de bases, denominada ΔS, por ejemplo, ΔS ≥ 10 pb. También se pueden utilizar otras pruebas estadísticas, por ejemplo, la prueba de la t, prueba de la U de Mann-Whitney, prueba de Kolmogorov-Smirnov, etc. Se determinaron las tasas de recuperación y los valores predictivos positivos (PPV) aplicando cada nivel sucesivo de filtrado. La tasa de recuperación se basa en la proporción de los 47 mutantes *de novo* conocidos detectados después del filtrado. Los PPV se refieren al número de mutantes *de novo* verdaderos detectados como una proporción de todas las variantes no maternas y no paternas detectadas en los datos de secuenciación de ADN plasmático materno sin células. Cuantos menor sean las variantes *de novo* falsas positivas, mayor será el PPV. Los falsos positivos podrían producirse como resultado de, y sin limitación, errores de secuenciación y errores de alineación. Los PPV conseguidos con esta estrategia son sustancialmente mejores que los informados previamente por Kitzman et al (Sci Transl Med 2012; 137: 137ra76). La secuenciación de una biblioteca de ADN plasmático materno preparada utilizando un protocolo sin PCR para una cobertura de 78x ha dado lugar a la identificación de 2,5 x 10^7 falsos positivos mientras que las verdaderas mutaciones *de novo* fueron solamente 44. El PPV de este estudio fue solamente del 0,000176 %.

Como prueba corroborativa para demostrar que las variantes o mutantes *de novo* presuntos detectados son de origen fetal, se compararon los perfiles de tamaño de las variantes o los mutantes *de novo* identificados utilizando los diferentes niveles de filtrado.

La figura 36A muestra perfiles de tamaño de fragmentos de ADN con las supuestas mutaciones identificadas en plasma utilizando criterio de filtrado de nivel A en comparación con el alelo de tipo silvestre. La figura 36B muestra perfiles de tamaño de fragmentos de ADN con las supuestas mutaciones identificadas en plasma usando criterios de filtrado de nivel B. La figura 36C muestra perfiles de tamaño de fragmentos de ADN con las supuestas mutaciones identificadas en plasma usando criterios de filtrado de nivel C. La figura 36D muestra perfiles de tamaño de fragmentos de ADN con las supuestas mutaciones identificadas en plasma usando criterios de filtrado de nivel D. Como se observa en las figuras 36A-36D, las variantes identificadas por el algoritmo de nivel D muestran la distribución de tamaño más corta.

La figura 37 muestra los perfiles de los valores de ΔF correspondientes a supuestas mutaciones identificadas utilizando diferentes niveles de criterios de filtrado, concretamente, A, B, C y D. Los valores de ΔF derivados de 298.364 SNP informativos en los que tanto la madre como el padre eran homocigotos pero con diferentes alelos se usaron como referencia que representa la diferencia en las frecuencias acumulativas entre fragmentos de ADN derivados del feto y derivados de la madre. El perfil de tamaño deducido de los criterios de filtrado de nivel D resultó ser el más parecido a los valores de ΔF deducidos de los sitios con SNP informativos, sugiriendo que las supuestas mutaciones *de novo* identificadas en el criterio D se habían enriquecido con más mutaciones verdaderas que se presentaban en la placenta/feto.

La figura 38 muestra un recuento de frecuencias de varios tipos de mutaciones en una muestra de plasma materno y

sangre del cordón umbilical. En la figura 38, las mutaciones identificadas en el plasma son similares a las mutaciones extraídas en la sangre del cordón umbilical. Estos datos sugieren que las mutaciones detectadas en el plasma materno están presentes en el genoma fetal como lo muestran los datos de la sangre del cordón umbilical.

La figura 39A muestra un gráfico del % de PPV y tasas de recuperación para filtros de diferentes tamaños según las realizaciones de la presente invención. La figura 39A muestra cómo la variación del parámetro de filtrado por tamaño afecta significativamente el % de PPV y la tasa de recuperación cuando no se aplicó filtrado de fracción extra mutante (M%). La figura 39B muestra un gráfico del % de PPV y las tasas de recuperación para diferentes cortes dinámicos de fracción mutante. La figura 39B muestra que la variación del parámetro de fracción mutante afecta significativamente el % de PPV y la tasa de recuperación cuando no se realiza un filtrado de ΔS adicional.

Las figuras 40A-40D muestran gráficos del % de PPV y tasas de recuperación para filtros de varios tamaños en diferentes cortes dinámicos de fracción mutante. Variar el parámetro de filtrado por tamaño ΔS en diferentes criterios de M% afecta sinérgicamente el % de PPV y las tasas de recuperación.

La figura 41 es un gráfico que muestra las curvas de las tasas de recuperación y el % de PPV en diferentes cortes dinámicos de fracción mutante en función de los cortes dinámicos de tamaño. Gráfico sistemático que revela las interacciones entre ΔS, M% y PPV%, tasa de recuperación.

C. *Confirmación de las supuestas mutaciones de novo*

El objetivo fue confirmar y validar las 47 mutaciones *de novo*. Los cebadores se diseñaron para amplificar específicamente cada una de las supuestas mutaciones *de novo* seguida por la secuenciación de Sanger del ADN genómico paterno, materno, de placenta y de sangre de cordón umbilical. Los resultados se muestran en la figura I, que muestra la secuenciación de última generación (NGS, por sus siglas en inglés) y el análisis de secuenciación de Sanger de las 48 supuestas mutaciones *de novo*. NGS se refiere a la secuenciación masiva en paralelo mencionada anteriormente, y "Sanger seq" se refiere a la secuenciación de Sanger. Los recuentos alélicos se muestran entre paréntesis para mayor claridad. Una de estas mutaciones (TP5) se detectó en la sangre del cordón umbilical pero no en la placenta. Debido a que las moléculas de ADN fetal en el plasma materno en su mayoría se originan en la placenta, la mutación específica de la sangre del cordón umbilical no sería detectable en el plasma materno. Por lo tanto, solamente las 47 mutaciones restantes derivadas de la placenta son pertinentes para la validación.

Las figuras 40 y 41 muestran una tabla de las 47 mutaciones *de novo*. En las figuras 40 y 41, las ubicaciones cromosómicas de la mutación diana se muestran en la columna 2. En la columna 3, se muestran los genotipos detectados en plasma materno. El alelo principal se coloca antes del alelo secundario. En la columna 4, se muestran las proporciones de lecturas que muestran el alelo principal con respecto al alelo secundario en cada uno de los sitios de mutación. En las columnas siguientes, los resultados basados en secuenciación masiva en paralelo o secuenciación de última generación (NGS) se muestran junto con los resultados de la secuenciación de Sanger. 43 de las 47 mutaciones se detectaron solamente en el ADN de la placenta pero no en el ADN paterno y materno. Esto significó que el 91 % de las mutaciones identificadas por la secuenciación del ADN plasmático materno eran verdaderas mutaciones *de novo* y, de este modo, la secuenciación de Sanger confirmó los datos de NGS para el ADN plasmático, ADN materno, ADN paterno, ADN placentario. Las reacciones de secuenciación de Sanger para la detección de la mutación TP45 fallaron. Los ensayos para las mutaciones TP21, TP30 y TP44 mostraron resultados inconsistentes entre la secuenciación NGS y de Sanger.

VIII. ANÁLISIS DE SIMULACIÓN PARA LA DETECCIÓN DE MUTACIONES DE CÁNCER A PARTIR DE ADN SIN CÉLULAS EN PLASMA HUMANO

Utilizando los datos de secuenciación generados a partir del caso de la gestante, se seleccionaron 3.000 variantes de un solo nucleótido que el feto había heredado de su padre y se asumió que eran mutaciones somáticas desarrolladas por un cáncer en un paciente con cáncer. En otras palabras, se analizaron los datos de secuenciación del ADN plasmático materno como si se hubiera secuenciado el ADN sin células de una muestra de plasma de un paciente con cáncer. Después se determinaron cuántas de las variantes y falsos positivos se detectarían si las muestras de plasma solamente se secuenciaran a una cobertura del genoma de humano de 25x, 50x y 100x cuando se aplicara el algoritmo de filtrado de nivel D. Se seleccionaron al azar 25x, 50x y 100x, respectivamente, de los datos de secuenciación entre los 255x de datos de secuenciación de ADN plasmático.

La figura 44 muestra las tasas de recuperación y los PPV para la detección de las 47 mutaciones *de novo* y las 3.000 supuestas mutaciones somáticas. Los algoritmos de filtrado de nivel D para los números de la Tabla 1 incluyen: valores de corte dinámicos, realineación, fracción mutante >20 % y filtro de tamaño 10 pb.

A continuación se realizó un análisis más extenso mediante simulación por ordenador.

Las figuras 45A-45C y 46A-46C muestran simulaciones con cantidades variables de mutaciones para diversas profundidades de secuenciación y fracciones tumorales. En este conjunto de análisis, se simularon las situaciones en

las que se tenía una profundidad de secuenciación del ADN plasmático que variaba de 25x a 800x, con concentraciones de fracción tumoral que variaban entre el 1 % y el 40 % y cuando el número de mutaciones somáticas desarrolladas por el tumor variaba entre 3.000 y 30.000. Todos los análisis se basan en el algoritmo de filtrado de nivel D.

Para cada una de estas simulaciones, el número de mutaciones somáticas detectadas así como el número de falsos positivos se muestran en las figuras 45A-45C y 46A-46C. Como se muestra en las figuras 45A-45C y 46A-46C, muchas condiciones permitirían detectar más mutaciones somáticas que falsos positivos. Estas condiciones serían clínicamente útiles como una "prueba de carga de mutación" para evaluar la carga de mutaciones presentes entre las moléculas de ADN plasmático. Cuando este nivel es mayor que un intervalo de referencia, por ejemplo, en comparación con controles de la misma edad y/o sexo, o en comparación con el propio ADN de las células sanguíneas, se sospecharía cáncer. Este estrategia se estaría utilizando como una herramienta de cribado para la detección del cáncer.

IX. MÉTODOS PARA DETERMINAR CÁNCER

Como se describe anteriormente, las realizaciones pueden proporcionar métodos para identificar con precisión mutaciones somáticas en un sujeto que se somete a pruebas. Varias realizaciones pueden usar secuenciación sin amplificación, secuenciación con amplificación mínima (por ejemplo, menos del 2 % de duplicación) y varios criterios de filtrado. Las identificación de mutaciones se puede utilizar para determinar un nivel de cáncer, así como otros fines.

A. Identificación de mutaciones

La figura 47 es un diagrama de flujo que ilustra un método 4700 para identificar mutaciones somáticas en un sujeto humano mediante el análisis de una muestra biológica del sujeto humano según las realizaciones de la presente invención. La muestra biológica incluye fragmentos de ADN que se originan a partir de células normales y potencialmente de células tumorales o asociadas con cáncer y la muestra biológica incluye fragmentos de ADN sin células. El método 4700 puede realizarse al menos parcialmente mediante un sistema informático, al igual que otros métodos descritos en el presente documento.

En el bloque 4710, los fragmentos de ADN plantilla se obtienen de la muestra biológica que se va a analizar. Los fragmentos de ADN plantilla incluyen fragmentos de ADN sin células. En diversas realizaciones, los fragmentos de ADN sin células de células tumorales o células asociadas con cáncer comprenden menos del 50 %, 40 %, 30 %, 20 %, 15 %, 10 %, 5 % o 1 % de los fragmentos de ADN sin células en la muestra biológica. La muestra biológica puede ser plasma o suero, u otros tipos de muestras mencionadas en el presente documento o que incluyan ADN sin células.

En el bloque 4720, se prepara una biblioteca de secuenciación de moléculas de ADN analizables utilizando los fragmentos de ADN plantilla. En una realización, la preparación de la biblioteca de secuenciación de moléculas de ADN analizables no incluye una etapa de amplificación del ADN de las moléculas de ADN plantilla. En otra realización, se puede realizar alguna amplificación de modo que se produzca dicho algún nivel de duplicación. Pero, el nivel de duplicación puede ser mínimo. En diversas implementaciones, una tasa de duplicación de la biblioteca de secuenciación de los fragmentos de ADN plantilla es inferior al 5 %, inferior al 2 % o inferior al 1 %. El número de moléculas de ADN analizables en la biblioteca de secuenciación puede ser menor que el número de fragmentos de ADN plantilla presentes originalmente en la muestra biológica antes de la preparación de la biblioteca.

En el bloque 4730, la biblioteca de secuenciación de moléculas de ADN analizables se secuencia para obtener una pluralidad de lecturas de secuencia. Se pueden utilizar varios tipos de procedimientos de secuenciación, como se describe en el presente documento. Se pueden utilizar varias profundidades y amplitudes. Como otro ejemplo, se puede realizar la secuenciación de una sola molécula. Y, la secuenciación puede ser una secuenciación con reconocimiento de la metilación.

En el bloque 4740, la pluralidad de lecturas de secuencia se recibe en un sistema informático. Las lecturas de secuencia se pueden recibir de cualquier manera o formato adecuados, por ejemplo, a través de una red desde una máquina de secuenciación o en un dispositivo de almacenamiento. Los datos recibidos de la máquina de secuenciación pueden ser valores de intensidad sin procesar que se usan para determinar asignaciones de bases.

En el bloque 4750, el ordenador puede alinear la pluralidad de lecturas de secuencia con un genoma humano de referencia para determinar posiciones genómicas para la pluralidad de lecturas de secuencia. En diversas realizaciones, se pueden utilizar profundidades de secuenciación de al menos 30x, 35x, 40x, 50x, 75x, 100x, 150x o 200x. Las lecturas de secuencia alineadas pueden comprender varias porciones del genoma humano de referencia, tal como al menos un 0,1 %, 1 %, 5 %, 10 % y 15 % del genoma humano de referencia.

En el bloque 4760, el sistema informático puede obtener información sobre un genoma constitutivo correspondiente al sujeto humano. El genoma constitutivo puede ser el del sujeto humano o un genoma de referencia que corresponda al sujeto humano. Por ejemplo, el genoma constitutivo puede ser un genoma de referencia para una población específica de sujetos humanos.

En el bloque 4770, el sistema informático puede comparar las lecturas de secuencia con el genoma constitutivo para identificar un conjunto de locus filtrado como portadores de mutaciones somáticas en algún tejido del sujeto humano. En un aspecto, en cada locus del conjunto filtrado, un número de lecturas de secuencia que tienen una variante de secuencia con respecto al genoma constitutivo está por encima de un valor de corte, donde el valor de corte es mayor que uno. El valor de corte dinámico puede ser un valor de corte dinámico como se describe en el presente documento. El valor de corte puede ser un criterio de filtro y se pueden aplicar otros. El conjunto filtrado puede ser el resultado final después de todos las etapas de filtrado, potencialmente usando varios criterios de filtrado.

En el bloque 4780, se pueden usar otros criterios de filtrado para identificar el conjunto de locus filtrado como portadores de mutaciones somáticas en algún tejido del sujeto humano. Dichos criterios de filtrado se describen en otra parte y a continuación.

En el bloque 4790, las mutaciones somáticas identificadas pueden utilizarse para diversos fines. A continuación se proporcionan varios ejemplos de fines. Por ejemplo, se puede determinar una carga mutacional y usarla para determinar un nivel de cáncer. Las mutaciones se pueden utilizar para diseñar pruebas adicionales, potencialmente para una evaluación adicional de un paciente, y para determinar el tratamiento de un paciente.

A continuación se describen ejemplos de la aplicación de otros criterios de filtrado, así como en otras secciones del presente documento. Se pueden usar los otros criterios de filtrado para identificar el conjunto de locus filtrado como portador de mutaciones somáticas en algún tejido del sujeto humano. Para algunos de los criterios de filtrado, se puede analizar un conjunto de locus candidatos identificados como potencialmente portadores de una mutación somática. Los locus candidatos se pueden haber identificado utilizando cualquier criterio adecuado, por ejemplo, un valor de corte fijo, un valor de corte dinámico u otros criterios de filtrado utilizados anteriormente. Por lo tanto, el conjunto resultante de locus candidatos puede ser el resultado de aplicar otro criterio de filtrado.

1. Realineación

Para la realineación, se puede analizar cada uno de un primer conjunto de locus candidatos identificados como potencialmente portadores de una mutación somática. Cada una de las lecturas de secuencia que se alinean con el locus candidato usando un primer procedimiento de alineación y teniendo la variante de secuencia se pueden analizar adicionalmente en un procedimiento de realineación. Se puede determinar si la lectura de secuencia se alinea con el locus candidato utilizando un segundo procedimiento de alineación que utiliza un algoritmo de coincidencia diferente al utilizado para el primer procedimiento de alineación, por ejemplo, como se describe en la sección V.B. Cuando la lectura de secuencia se realinea con el locus candidato utilizando el segundo procedimiento de alineación, se puede determinar una calidad de mapeo de la realineación para el segundo procedimiento de alineación.

Una vez que se determina la calidad del mapeo para la segunda alineación, la calidad del mapeo se puede comparar con un umbral de calidad, para determinar si la lectura de secuencia es de baja calidad. Entonces se puede determinar si se descarta la lectura de secuencia basándose en la comparación de la calidad del mapeo con el umbral de calidad. La determinación puede ser que las lecturas por debajo del umbral se pueden descartar. En otras realizaciones, se puede determinar una puntuación (por ejemplo, un peso) en función de la comparación, donde se pueden realizar comparaciones con múltiples umbrales de calidad para determinar la puntuación, por ejemplo, cada umbral corresponde a una puntuación de realineación diferente. Después, la puntuación se puede usar de manera colectiva con las puntuaciones de uno o más criterios de filtrado para determinar si se descarta la lectura. Con independencia de la manera específica (e inclusive de los ejemplos proporcionados anteriormente), el hecho de que la calidad del mapeo sea menor que el umbral de calidad proporciona una mayor probabilidad de descartar la lectura secuencia que el hecho de que la calidad del mapeo sea mayor que el umbral de calidad.

Como parte de este proceso de filtrado, se obtienen varias lecturas de secuencia restantes. El número de lecturas de secuencia restantes se puede comparar con un umbral candidato, que puede ser el mismo valor umbral utilizado originalmente para identificar los locus candidatos. En un análisis de probabilidad similar al de la lectura de secuencia, se puede determinar si se descarta el locus candidato basándose en la comparación del número de lecturas de secuencia restantes con el umbral candidato. El análisis puede ser estricto en función de la comparación con el umbral, o utilizar un sistema de puntuación (ponderación) como se menciona anteriormente. Con independencia, el número de lecturas de secuencia restantes que es menor que el umbral candidato proporciona una mayor probabilidad de descartar el locus candidato que el número de lecturas de secuencia restantes que es mayor que el umbral candidato. El conjunto de locus filtrado puede identificarse como portador de mutaciones somáticas utilizando los locus candidatos restantes.

2. Tamaño

Para un análisis de tamaño, puede analizarse cada uno de un conjunto de locus candidatos. Puede determinarse una diferencia de tamaño entre un primer grupo de fragmentos de ADN que tienen la variante de secuencia y un segundo grupo de fragmentos de ADN que tienen un alelo de tipo silvestre. Dichos análisis de tamaño se han descrito en el presente documento. La diferencia de tamaño puede estar entre cualquier valor estadístico de distribuciones de

tamaño para los dos grupos. Por ejemplo, se puede utilizar una diferencia en la mediana del tamaño del primer grupo de fragmentos de ADN y el segundo grupo de fragmentos de ADN. Como otro ejemplo, un máximo en una frecuencia acumulativa por tamaño entre el primer grupo y el segundo grupo. Cualquier valor de tamaño descrito en las publicaciones de patentes de Estados Unidos 2011/0276277 y 2013/0237431.

La diferencia de tamaño se puede comparar con un umbral de tamaño, que se puede determinar a partir de muestras que se sabe que tienen cáncer u otro estado que se está clasificando. A continuación, se puede determinar si se descarta el locus candidato como una posible mutación en función de la comparación. En cuanto a otros criterios de filtrado, la comparación se puede utilizar de manera estricta o como una puntuación. Con independencia, la diferencia de tamaño que es menor que el umbral de tamaño proporciona una mayor probabilidad de descartar el locus candidato que la diferencia de tamaño que es mayor que el umbral de tamaño. El conjunto de locus filtrado puede identificarse como portador de mutaciones somáticas en el sujeto humano utilizando los locus candidatos restantes.

3. Modificaciones de histonas

Para la modificación de histonas, se puede identificar un grupo de regiones que se sabe que están asociadas con modificaciones de histonas que están asociadas con cáncer. Cada uno de un conjunto de locus candidatos puede analizarse determinando si descartar el locus candidato basándose en si el locus candidato está en uno del grupo de regiones. En cuanto a otros criterios de filtrado, la comparación se puede utilizar de manera estricta o como una puntuación. Con independencia, el locus candidato que no está en uno de los grupos de regiones proporciona una mayor probabilidad de descartar el locus candidato que cuando el locus candidato está en uno de los grupos de regiones. El conjunto de locus filtrado puede identificarse como portador de mutaciones somáticas en el sujeto humano utilizando los locus candidatos restantes.

4. Fracción mutante

Para la fracción mutante, puede analizarse cada uno de un conjunto de locus candidatos. Se puede determinar una fracción de lecturas de secuencia que tienen la variante de secuencia y después compararla con el umbral de fracción. A continuación, se puede determinar si se descarta el locus candidato como una posible mutación en función de la comparación, por ejemplo, utilizando puntuaciones o valores de corte estrictos. De cualquier manera, la fracción que es menor que el umbral de fracción proporciona una mayor probabilidad de descartar el locus candidato que la fracción que es mayor que el umbral de fracción (por ejemplo, 5 %, 10 %, 20 % o 30 %). El conjunto de locus filtrado puede identificarse como portador de mutaciones somáticas en el sujeto humano utilizando los locus candidatos restantes.

En algunas realizaciones, el umbral de fracción se puede determinar basándose en una concentración fraccionaria medida de ADN tumoral en la muestra biológica. La concentración fraccionaria de ADN tumoral en la muestra biológica puede medirse para cada una de una pluralidad de regiones (por ejemplo, usando técnicas similares pero con datos específicos para uno o más locus en las regiones). El umbral de fracción utilizado para un locus candidato puede ser la concentración fraccionaria medida para la región en la que reside el locus candidato.

En otra realización, pueden usarse regiones aberrantes para determinar un umbral de fracción. Se pueden identificar una o más regiones aberrantes que tienen una aberración en el número de copias. El umbral de fracción utilizado para un locus candidato en una región aberrante puede depender de si la región aberrante presenta una ganancia de número de copias o una pérdida de número de copias. Se puede usar un umbral más alto para una ganancia y un umbral más bajo para una pérdida.

También se pueden usar una o más regiones aberrantes que tienen una aberración en el número de copias como parte para determinar si se descarta lecturas de secuencia para determinar el número de las lecturas de secuencia que tienen una variante de secuencia en relación con el genoma constitutivo para cada conjunto de locus filtrado. Es más probable que una primera lectura de secuencia de una primera región aberrante que muestre una ganancia de número de copias tenga una mutación somática que una segunda lectura de secuencia de una segunda región aberrante que muestre una pérdida de número de copias.

Se pueden identificar una o más regiones aberrantes analizando un conjunto de locus candidatos. Se puede calcular una fracción mutante aparente de una variante de secuencia en relación con genoma constitutivo. Puede determinarse una varianza en las fracciones mutantes aparentes de los locus candidatos en la región aberrante para cada una de una pluralidad de regiones. La varianza se puede comparar con un umbral de varianza, donde una región aberrante que muestra una ganancia de número de copias tiene una varianza mayor que el umbral.

5. Estado de metilación

Para el estado de metilación, la secuenciación es una secuenciación con reconocimiento de la metilación. Puede analizarse cada uno de un conjunto de locus candidatos, alineándose cada una de las lecturas de secuencia con el locus candidato y que tiene la variante de secuencia a analizar. Para una lectura de secuencia, se puede determinar un estado de metilación de la molécula de ADN analizable correspondiente en uno o más sitios (por ejemplo, sitios CpG). Se puede determinar si se descarta la lectura de secuencia en función del estado de metilación. En cuanto a

otros criterios de filtrado, la comparación se puede utilizar de manera estricta o como una puntuación. Con independencia, el estado de metilación que no está metilado proporciona una mayor probabilidad de descartar la lectura de secuencia que el estado de metilación que está metilado.

El número de lecturas de secuencia restantes se puede comparar con un umbral candidato, que puede ser el mismo que se usa para identificar los locus candidatos (como también ocurre con otros usos de un umbral candidato para otros criterios de filtrado). En un análisis de probabilidad similar al de la lectura de secuencia, se puede determinar si se descarta el locus candidato basándose en la comparación del número de lecturas de secuencia restantes con el umbral candidato. El análisis puede ser estricto en función de la comparación con el umbral, o utilizar un sistema de puntuación (ponderación) como se menciona anteriormente. Con independencia, el número de lecturas de secuencia restantes que es menor que el umbral candidato proporciona una mayor probabilidad de descartar el locus candidato que el número de lecturas de secuencia restantes que es mayor que el umbral candidato. El conjunto de locus filtrado puede identificarse como portador de mutaciones somáticas utilizando los locus candidatos restantes.

6. Ubicaciones finales del ADN plasmático

Para las ubicaciones finales del ADN plasmático, puede analizarse cada uno de un conjunto de locus candidatos, alineándose cada una de las lecturas de secuencia con el locus candidato y que tiene la variante de secuencia a analizar. Para una lectura de secuencia, se puede determinar una ubicación final correspondiente a donde se alinea un final de la lectura de secuencia. La ubicación final se puede comparar con una pluralidad de ubicaciones terminales paraneoplásicas o específicas de cáncer. La decisión de descartar la lectura de secuencia se determina en función de la comparación. La ubicación final que no es una ubicación terminal específica de cáncer o paraneoplásica proporciona una mayor probabilidad de descartar la lectura de secuencia que la ubicación final que es una ubicación terminal específica de cáncer o paraneoplásica. El número restante de lecturas de secuencias se puede utilizar para determinar si se descarta el locus candidato.

7. Secuenciación monocatenaria

La secuenciación se puede realizar mediante un proceso de preparación de bibliotecas de secuenciación monocatenaria que proporciona una etapa de secuenciación posterior para producir lecturas de dos cadenas para cada molécula de ADN plantilla. Un ejemplo de un proceso de preparación de una biblioteca de secuenciación monocatenaria se describe en Snyder *et al.* Cell 2016; 164: 57-68. Puede analizarse cada uno de un conjunto de locus candidatos, con cada par de lecturas de cadena alineándose con el locus candidato que se está analizando. Puede determinarse si ambas cadenas tienen la variante de secuencia. A continuación, se puede determinar si se descarta la lectura de secuencia en función de si ambas cadenas tienen la variante de secuencia. El hecho de que ambas cadenas no tengan la variante de secuencia proporciona una mayor probabilidad de descartar las lecturas de la cadena que la lectura de una sola cadena que tenga la variante de secuencia. El número restante de lecturas de secuencias se puede utilizar para determinar si se descarta el locus candidato.

B. Determinación del nivel de cáncer

La figura 48 es un diagrama de flujo que ilustra un método 4800 para usar mutaciones somáticas identificadas para analizar una muestra biológica de un sujeto según las realizaciones de la presente invención.

En el bloque 4810, se identifican las mutaciones somáticas. Las mutaciones somáticas pueden identificarse como se describe para el método 4700 de la figura 47.

En el bloque 4820, se determina una carga mutacional para el sujeto humano usando una cantidad de locus en el conjunto de locus filtrado. En diversas realizaciones, la carga mutacional se puede determinar como un número de mutaciones somáticas sin procesar, una densidad de mutaciones somáticas por número de bases, un porcentaje de locus de una región genómica que se identifican como portadores de mutaciones somáticas, un número de mutaciones somáticas observadas en una cantidad particular de muestra o un aumento en comparación con una carga de referencia.

En el bloque 4830, la carga mutacional se compara con un umbral de cáncer para determinar un nivel de cáncer. El umbral de cáncer se puede determinar basándose en una discriminación entre pacientes con cáncer y sujetos sin cáncer. Un experto en la materia apreciará que pueden usarse diferentes umbrales, dependiendo de la sensibilidad y especificidad deseadas. Como se muestra en el presente documento, las realizaciones se pueden usar para determinar una carga mutacional que puede discriminar entre un sujeto sano y uno con cáncer, por ejemplo, HCC.

En el bloque 4840, cuando el nivel de cáncer indica la existencia de un tumor, se puede determinar el tejido de origen del cáncer. Como ejemplos, dicha determinación se puede realizar utilizando firmas de metilación o modificaciones de histonas o distribución de las ubicaciones finales de los fragmentos de ADN analizados.

En una realización usando modificaciones de histonas, se determina una primera cantidad de modificaciones de histonas para cada uno de una primera pluralidad de segmentos del genoma humano de referencia. Esta primera

cantidad se puede determinar a partir de la información de referencia disponible sobre qué locus están asociados con las modificaciones de histonas pertinentes. Puede determinarse una segunda cantidad del conjunto de locus filtrado para cada uno de una segunda pluralidad de segmentos del genoma humano de referencia. Los segmentos de diferencia se pueden correlacionar entre sí después. En consecuencia, se puede determinar un primer conjunto de segmentos que tienen la primera cantidad de modificaciones de histonas por encima de un primer umbral y que tienen la segunda cantidad del conjunto de locus filtrados por encima de un segundo umbral. Los dos umbrales pueden ser iguales. Los umbrales pueden asegurar que los segmentos del genoma sean aquellos con altas modificaciones de histonas y un alto número de mutaciones somáticas. Las cantidades y los umbrales pueden ser números o densidades sin procesar (por ejemplo, por megabase).

En el bloque 4850, el tratamiento puede proporcionarse según el nivel de cáncer determinado, las mutaciones identificadas y/o el tejido de origen. Por ejemplo, las mutaciones identificadas pueden ser diana de un fármaco o quimioterapia en particular. El tejido de origen se puede utilizar para guiar una cirugía. Y, el nivel de cáncer puede utilizarse para determinar la agresividad de cualquier tipo de tratamiento, que también puede determinarse en función del nivel de cáncer.

C. *Otros usos para mutaciones identificadas*

Tal como se menciona anteriormente, el número de mutaciones se puede utilizar como indicación de que el sujeto analizado tiene cáncer. En una realización, un individuo puede clasificarse como que tiene alta probabilidad de tener cáncer si el número de mutaciones detectadas es mayor que el detectado en sujetos sin cáncer.

El conjunto de mutaciones, una vez identificadas, podría usarse para informar el diseño de ensayos más dirigidos (basados en las mutaciones representadas en la carga mutacional) para la monitorización futura del cáncer del paciente, a efectos de confirmación, a efectos de medición más precisa, o a efectos de medición en serie (que sería más económico que repetir la secuenciación exhaustiva varias veces). Dichas mediciones en serie serían útiles a efectos de seguimiento, por ejemplo, para ver si la concentración de la firma mutacional en plasma está aumentando (potencialmente un signo de mal pronóstico) o disminuyendo (potencialmente un signo de buen pronóstico o si el cáncer responde al tratamiento elegido).

Las mutaciones específicas detectadas en la carga mutacional proporcionarían información para que los profesionales médicos elijan la terapia o el fármaco pertinente, por ejemplo, terapia dirigida. A modo de ejemplo, se pueden usar inhibidores de tirosina cinasa para tratar cánceres con mutaciones específicas en el gen del receptor del factor de crecimiento epidérmico.

El espectro de mutaciones identificadas se puede usar para ayudar a identificar el sitio del tumor porque se ha encontrado que los tumores desarrollados a partir de diferentes órganos/tejidos tienen diferentes perfiles mutacionales (Polak *et al.* Nature 2015; 518: 360-364). También podría proporcionar información sobre la exposición ambiental y los carcinógenos que están causalmente relacionados con el conjunto de mutaciones detectadas (Alexandrov *et al.* Nature 2013; 500: 415-421). El espectro de mutaciones identificadas se puede utilizar para ayudar en el pronóstico. Por ejemplo, algunas mutaciones pueden ser marcadores de cánceres que son particularmente agresivos o indolentes.

En el contexto de las pruebas prenatales, el conjunto de mutaciones identificadas podría usarse para informar el diseño de ensayos más específicos (basándose en las mutaciones representadas en la carga mutacional) para la detección específica de dichas mutaciones en el plasma materno. Además, en el contexto de las pruebas prenatales, el conjunto de mutaciones identificadas podría utilizarse para informar a los profesionales médicos de la necesidad de un control clínico especial del caso. Como ejemplo, la detección de una mutación de hemofilia esporádica en un feto masculino podría indicar la necesidad de precaución durante el procedimiento de parto (por ejemplo, evitar el parto con fórceps) si la mujer gestante decide continuar con el embarazo a término. Como otro ejemplo, la detección de un feto femenino que es homocigoto o heterocigoto compuesto para mutaciones de hiperplasia suprarrenal congénita (CAH, por sus siglas en inglés) en una familia sin antecedentes familiares de CAH alertaría al médico sobre la necesidad de una terapia temprana con dexametasona para la mujer gestante, para reducir el riesgo de virilización de los genitales fetales.

X. MÉTODOS PARA EL ANÁLISIS FETAL

La figura 49 es un diagrama de flujo que ilustra un método 4900 para identificar mutaciones *de novo* de un feto mediante el análisis de una muestra biológica de una mujer gestante del feto según las realizaciones de la presente invención. La muestra biológica incluye fragmentos de ADN sin células del feto y del sujeto femenino.

En el bloque 4910, los fragmentos de ADN plantilla se obtienen de la muestra biológica que se va a analizar. Los fragmentos de ADN plantilla incluyen fragmentos de ADN sin células. El bloque 4910 se puede realizar de manera similar al bloque 4710 de la figura 47.

En el bloque 4920, se prepara una biblioteca de secuenciación de moléculas de ADN analizables utilizando los fragmentos de ADN plantilla. El bloque 4920 se puede realizar de manera similar al bloque 4720 de la figura 47.

En el bloque 4930, la biblioteca de secuenciación de moléculas de ADN analizables se secuencia para obtener una pluralidad de lecturas de secuencia. El bloque 4930 se puede realizar de manera similar al bloque 4730 de la figura 47.

En el bloque 4940, la pluralidad de lecturas de secuencia se recibe en un sistema informático. El bloque 4940 se puede realizar de manera similar al bloque 4740 de la figura 47.

En el bloque 4950, el ordenador puede alinear la pluralidad de lecturas de secuencia con un genoma humano de referencia para determinar posiciones genómicas para la pluralidad de lecturas de secuencia. El bloque 4950 se puede realizar de manera similar al bloque 4750 de la figura 47.

En el bloque 4960, el sistema informático puede obtener información sobre un genoma materno del sujeto femenino y un genoma paterno del padre del feto. La información puede incluir información de genotipo sobre ambos padres en los locus examinados para determinar la existencia de una mutación. Tal información de genotipo puede obtenerse a través de cualquier técnica adecuada como conocería por un experto en la materia.

En el bloque 4970, el sistema informático puede comparar las lecturas de secuencia con el genoma materno y el genoma paterno para identificar un conjunto de locus filtrado que tiene mutaciones *de novo* en el feto. En un aspecto, en cada locus del conjunto filtrado, un número de lecturas de secuencia que tienen una variante de secuencia que no está en el genoma materno ni en el genoma paterno está por encima de un valor de corte, donde el valor de corte es mayor que uno.

En el bloque 4980, se pueden usar otros criterios de filtrado para identificar el conjunto de locus filtrado que tiene mutaciones *de novo* en el feto. Dichos criterios de filtrado se describen en otra parte, por ejemplo, en la sección IX.

En el bloque 4990, las mutaciones *de novo* identificadas pueden utilizarse para diversos fines. Se pueden encontrar ejemplos de dichos fines en la sección IX.C.

XI. SISTEMA INFORMÁTICO

Cualquiera de los sistemas informáticos mencionados en este documento puede utilizar cualquier número adecuado de subsistemas. En la figura 15 relacionada con el aparato informático 10 se muestran ejemplos dichos subsistemas. En algunas realizaciones, un sistema informático incluye un único aparato informático, donde los subsistemas pueden ser los componentes del aparato informático. En otras realizaciones, un sistema informático puede incluir varios aparatos informáticos, siendo cada uno de ellos un subsistema, con componentes internos. Un sistema informático puede incluir ordenadores de escritorio y portátiles, tabletas, teléfonos móviles y otros dispositivos móviles.

Los subsistemas mostrados en la figura 15 están interconectados a través de un bus de sistema 75. Se muestran subsistemas adicionales tales como una impresora 74, teclado 78, dispositivo(s) de almacenamiento 79, un monitor 76, que está acoplado al adaptador de pantalla 82 y otros. Los dispositivos periféricos y de entrada/salida (I/O, por sus siglas en inglés), que se acoplan al controlador de I/O 71, pueden conectarse al sistema informático mediante cualquier tipo de medio conocido en la técnica, tal como el puerto de entrada/salida (I/O) 77 (por ejemplo, USB, FireWire®). Por ejemplo, el puerto de I/O 77 o la interfaz externa 81 (por ejemplo, Ethernet, Wi-Fi, etc.) pueden utilizarse para conectar el sistema informático 10 a una red de área amplia, como Internet, un dispositivo de entrada de ratón o un escáner. La interconexión mediante el bus de sistema 75 permite que el procesador central 73 se comunique con cada subsistema y controle la ejecución de instrucciones procedentes de la memoria del sistema 72 o de los dispositivos de almacenamiento 79 (por ejemplo, un disco fijo, tal como un disco duro o un disco óptico), así como el intercambio de información entre subsistemas. La memoria del sistema 72 y/o los dispositivos de almacenamiento 79 pueden ser realizaciones de un medio legible por ordenador. Otro subsistema es un dispositivo de recogida de datos 85, tal como una cámara, micrófono, acelerómetro y similares. Cualquiera de los datos mencionados en el presente documento puede ser la salida de un componente a otro componente y puede ser la salida al usuario.

Un sistema informático puede incluir una pluralidad de los mismos componentes o subsistemas, por ejemplo, conectados entre sí por la interfaz externa 81 o por una interfaz interna. En algunas realizaciones, los sistemas, subsistema o aparatos informáticos pueden comunicarse a través de una red. En dichos casos, un ordenador puede considerarse como cliente y otro ordenador como servidor, donde cada uno puede formar parte de un mismo sistema informático. Un cliente y un servidor pueden incluir cada uno múltiples sistemas, subsistemas o componentes.

Debe entenderse que cualquiera de las realizaciones de la presente invención puede implementarse en forma de lógica de control utilizando un hardware (por ejemplo, un circuito integrado específico de la aplicación o una matriz de selección programable por campo) y/o utilizando un software de ordenador con un procesador generalmente programable de manera modular o integrada. Tal como se usa en el presente documento, un procesador incluye un procesador de un solo núcleo, un procesador de varios núcleos en un mismo chip integrado, o múltiples unidades de procesamiento en una sola placa de circuito o en red. Basándose en la divulgación y las enseñanzas proporcionadas en el presente documento, un experto habitual en la materia conocerá y percibirá otras formas y/o métodos para

implementar las realizaciones de la presente invención utilizando hardware y una combinación de hardware y software.

Puede implementarse cualquiera de los componentes o las funciones de software descritos en la presente solicitud como código de software para su ejecución mediante un procesador utilizando cualquier lenguaje informático adecuado, tal como, por ejemplo, Java, C, C++, C#, Objective-C, Swift, o lenguaje de scripting como Perl o Python utilizando, por ejemplo, técnicas convencionales u orientadas a objetivos. El código de software puede almacenarse como una serie de instrucciones o comandos en un medio legible por ordenador para su almacenamiento y/o transmisión, los medios adecuados incluyen la memoria de acceso aleatorio (RAM), una memoria de solo lectura (ROM), un medio magnético como un disco duro o disquete o un medio óptico tal como un disco compacto (CD) o un DVD (disco digital versátil), memoria flash y similares. El medio legible por ordenador puede ser cualquier combinación de tales dispositivos de almacenamiento o transmisión.

Dichos programas también pueden codificarse y transmitirse utilizando señales portadoras adaptadas para la transmisión a través de redes cableadas, ópticas y/o inalámbricas que se ajusten a una variedad de protocolos, incluida Internet. De este modo, se puede crear un medio legible por ordenador según una realización de la presente invención utilizando una señal de datos codificada con dichos programas. Los medios legibles por ordenador codificados con el código del programa pueden empaquetarse con un dispositivo compatible o proporcionarse por separado de otros dispositivos (por ejemplo, a través de descarga de Internet). Cualquier medio legible por ordenador puede residir en o dentro de un solo producto informático (por ejemplo, un disco duro, un CD o un sistema informático completo) y puede estar presente en o dentro de diferentes productos informáticos dentro de un sistema o red. Un sistema informático puede incluir un monitor, una impresora u otro dispositivo de presentación adecuado para proporcionar a un usuario cualquiera de los resultados mencionados en el presente documento.

Cualquiera de los métodos descritos en el presente documento puede realizarse total o parcialmente con un sistema informático que incluya uno o más procesadores, que pueden configurarse para realizar las etapas. Por lo tanto, las realizaciones pueden dirigirse a sistemas informáticos configurados para realizar las etapas de cualquiera de los métodos descritos en el presente documento, potencialmente con diferentes componentes que realizan una etapa o un grupo de etapas respectivas. Aunque se presentan como etapas numeradas, las etapas de los métodos descritos en el presente documento pueden realizarse al mismo tiempo o en un orden diferente. Además, pueden usarse partes de estas etapas con partes de otras etapas de otros métodos. Además, la totalidad de una etapa o una parte de la misma puede ser opcional. Además, cualquiera de las etapas de cualquiera de los métodos puede realizarse con módulos, circuitos, u otros medios para realizar estas etapas.

La descripción anterior de las realizaciones ilustrativas de la invención se ha presentado con fines ilustrativos y descriptivos. No pretenden ser exhaustivos ni limitar la invención a la forma precisa descrita y son posibles muchas modificaciones y variaciones a la luz de las enseñanzas anteriores.

Una cita de "un", "una" o "el/la" se entiende como "uno o más" a menos que se indique específicamente lo contrario. El uso de "o" pretende hacer referencia a un "o inclusivo", y no a un "o exclusivo", a menos que se indique específicamente lo contrario.

Una cita de "un", "una" o "el/la" se entiende como "uno o más" a menos que se indique específicamente lo contrario. El uso de "o" pretende hacer referencia a un "o inclusivo", y no a un "o exclusivo", a menos que se indique específicamente lo contrario.

REIVINDICACIONES

1. Un método para identificar mutaciones somáticas en un sujeto humano mediante el análisis de una muestra biológica del sujeto humano, incluyendo la muestra biológica fragmentos de ácido desoxirribonucleico (ADN) procedentes de células normales y potencialmente de células tumorales o células asociadas con cáncer, incluyendo la muestra biológica fragmentos de ADN sin células, comprendiendo el método, realizar, mediante un sistema informático:

la recepción de una o más lecturas de secuencias para cada uno de una pluralidad de fragmentos de ADN en la muestra biológica;
la alineación de la pluralidad de lecturas de secuencia con un genoma humano de referencia utilizando un primer procedimiento de alineación para determinar las posiciones genómicas para la pluralidad de lecturas de secuencia;
la comparación de las lecturas de secuencias con un genoma constitutivo, correspondiente al sujeto humano, para identificar un conjunto de locus filtrado como portadores de mutaciones somáticas en algún tejido del sujeto humano, en donde:
en cada locus del conjunto filtrado, un número de lecturas de secuencia que tienen una variante de secuencia con respecto al genoma constitutivo está por encima de un valor de corte, siendo el valor de corte mayor que uno;
para cada uno de un primer conjunto de locus candidatos identificados como potencialmente portadores de una mutación somática:

determinar una diferencia de tamaño entre un primer grupo de fragmentos de ADN que tienen la variante de secuencia y un segundo grupo de fragmentos de ADN que tienen un alelo de tipo silvestre;
comparar la diferencia de tamaño con un umbral de tamaño;
cuando la diferencia de tamaño es menor que el umbral de tamaño, descartar el locus candidato como una posible mutación; e

identificar el conjunto de locus filtrado como portadores de mutaciones somáticas en el sujeto humano utilizando los locus candidatos restantes.

2. El método de la reivindicación 1, en donde identificar el conjunto de locus filtrado como portadores de mutaciones somáticas en algún tejido del sujeto humano incluye además:

identificar un grupo de regiones que se sabe que están asociadas con modificaciones de histonas que están asociadas con cáncer;
para cada uno de un segundo conjunto de locus candidatos identificados como potencialmente portadores de una mutación somática:

determinar si el locus candidato está en uno de los grupos de regiones;
determinar si se descarta el locus candidato en función de si el locus candidato se encuentra en uno de los grupos de regiones, en donde el locus candidato que no está en uno de los grupo de regiones proporciona una mayor probabilidad de descartar el locus candidato que cuando el locus candidato está en uno de los grupos de regiones;

identificar el conjunto de locus filtrado como portadores de mutaciones somáticas utilizando los locus candidatos restantes.

3. El método de la reivindicación 1 o 2, que comprende además:
determinar una carga mutacional para el sujeto humano usando una cantidad de locus en el conjunto de locus filtrado.

4. El método de la reivindicación 3, en donde la carga mutacional se determina como un número sin procesar de mutaciones somáticas, una densidad de mutaciones somáticas por número de bases, un porcentaje de locus de una región genómica que se identifican como portadores de mutaciones somáticas, un número de mutaciones somáticas observadas en una cantidad particular de muestra o un aumento en comparación con una carga de referencia.

5. El método de la reivindicación 3 o 4, que comprende además:
comparar la carga mutacional con un umbral de cáncer para determinar un nivel de cáncer.

6. El método de la reivindicación 5, en donde el nivel de cáncer indica un tumor, que comprende además:

determinar una primera cantidad de modificaciones de histonas para cada uno de una primera pluralidad de segmentos del genoma humano de referencia;
determinar una segunda cantidad del conjunto de locus filtrado para cada uno de una segunda pluralidad de segmentos del genoma humano de referencia;
determinar un primer conjunto de segmentos que tienen la primera cantidad de modificaciones de histonas por encima de un primer umbral y que tienen la segunda cantidad del conjunto de locus filtrado por encima de un segundo umbral; e
identificar un tejido de origen del tumor basándose en el primer conjunto de segmentos.

7. El método de la reivindicación 1, en donde identificar el conjunto de locus filtrado como portadores de mutaciones somáticas en algún tejido del sujeto humano incluye además:

para cada uno de un segundo conjunto de locus candidatos identificados como potencialmente portadores de una mutación somática:

determinar una fracción de lecturas de secuencia que tienen la variante de secuencia;
comparar la fracción con un umbral de fracción;
determinar si se descarta el locus candidato como una posible mutación en función de la comparación,
en donde la fracción que es menor que el umbral de fracción proporciona una mayor probabilidad de descartar el locus candidato que la fracción que es mayor que el umbral de fracción; e

identificar el conjunto de locus filtrado como portadores de mutaciones somáticas en el sujeto humano utilizando los locus candidatos restantes.

8. El método de la reivindicación 7, que comprende además:
medir una concentración fraccionaria de ADN tumoral en la muestra biológica, en donde el umbral de fracción se determina en función de la concentración fraccionaria.

9. El método de la reivindicación 7 u 8, que comprende además:
identificar una o más regiones aberrantes que tienen una aberración en el número de copias, en donde el umbral de fracción utilizado para un locus candidato en una región aberrante depende de si la región aberrante presenta una ganancia de número de copias o una pérdida de número de copias.

10. El método de la reivindicación 7 u 8, que comprende además:

identificar una o más regiones aberrantes que tienen una aberración en el número de copias; e
identificar una primera lectura de secuencia de una primera región aberrante que muestra una ganancia de número de copias para que sea más probable que tenga una mutación somática que una segunda lectura de secuencia de una segunda región aberrante que muestra una pérdida de número de copias como parte para determinar si se descartan las lecturas de secuencia para determinar el número de lecturas de secuencia que tienen una variante de secuencia con respecto al genoma constitutivo para cada conjunto de locus filtrado.

11. El método de la reivindicación 10, en donde la una o más regiones aberrantes se identifican:

para cada uno del segundo conjunto de locus candidatos identificados como potencialmente portadores de una mutación somática:
calculando una fracción mutante aparente de una variante de secuencia en relación con el genoma constitutivo;
para cada una de una pluralidad de regiones:

determinando una varianza en las fracciones mutantes aparentes de los locus candidatos en la región aberrante;
comparando la varianza con un umbral de varianza, donde una región aberrante que muestra una ganancia de número de copias tiene una varianza mayor que el umbral de varianza.

12. El método de la reivindicación 1, en donde la secuenciación es una secuenciación con reconocimiento de la metilación, y en donde la identificación del conjunto de locus filtrado como portadores de mutaciones somáticas en algún tejido del sujeto humano incluye además:

para cada uno de un segundo conjunto de locus candidatos identificados como potencialmente portadores de una mutación somática:

para cada una de las lecturas de secuencia que se alinean con el locus candidato y tienen la variante de secuencia:

determinar un estado de metilación de la correspondiente molécula de ADN analizable en uno o más sitios;
determinar si se descarta la lectura de secuencia en función del estado de metilación, en donde el estado de metilación que no está metilado proporciona una mayor probabilidad de descartar la lectura de secuencia que el estado de metilación que está metilado, obteniendo así un número de lecturas de secuencia restantes;

comparar el número de lecturas de secuencias restantes con un umbral candidato; y
determinar si se descarta el locus candidato basándose en la comparación del número de lecturas de secuencia restantes con el umbral candidato, en donde el número de lecturas de secuencia restantes que es menor que

el umbral candidato proporciona una mayor probabilidad de descartar el locus candidato que el número de lecturas de secuencia restantes que es mayor que el umbral candidato; e

identificar el conjunto de locus filtrado como portadores de mutaciones somáticas utilizando los locus candidatos restantes.

13. El método de la reivindicación 1, en donde identificar el conjunto de locus filtrado como portadores de mutaciones somáticas en algún tejido del sujeto humano incluye además:

para cada uno de un segundo conjunto de locus candidatos identificados como potencialmente portadores de una mutación somática:

para cada una de las lecturas de secuencia que se alinean con el locus candidato y tienen la variante de secuencia:

determinar una ubicación final correspondiente a donde se alinea un extremo de la lectura de secuencia;
comparar la ubicación final con una pluralidad de ubicaciones terminales específicas del cáncer o asociadas con cáncer;
determinar si se descarta la lectura de secuencia en función de la comparación, en donde la ubicación final que no es una ubicación terminal específica de cáncer o paraneoplásica proporciona una mayor probabilidad de descartar la lectura de secuencia que la ubicación final que es una ubicación terminal específica de cáncer o paraneoplásica, obteniendo así un número de lecturas de secuencia restantes;

comparar el número de lecturas de secuencias restantes con un umbral candidato; y
determinar si se descarta el locus candidato basándose en la comparación del número de lecturas de secuencia restantes con el umbral candidato, en donde el número de lecturas de secuencia restantes que es menor que el umbral candidato proporciona una mayor probabilidad de descartar el locus candidato que el número de lecturas de secuencia restantes que es mayor que el umbral candidato; e

identificar el conjunto de locus filtrado como portadores de mutaciones somáticas utilizando los locus candidatos restantes.

14. El método de la reivindicación 1, en donde la secuenciación se realiza utilizando un proceso de preparación de bibliotecas de secuenciación monocatenarias que proporciona una etapa de secuenciación posterior para producir lecturas de dos cadenas para cada molécula de ADN plantilla, en donde identificar el conjunto de locus filtrado como portadores de mutaciones somáticas en algún tejido del sujeto humano incluye además:

para cada uno de un segundo conjunto de locus candidatos identificados como potencialmente portadores de una mutación somática:

para cada par de cadenas que se alinean con el locus candidato:

determinar si ambas cadenas tienen la variante de secuencia;
determinar si se descarta la lectura de secuencia en función de si ambas cadenas tienen la variante de secuencia, en donde el hecho de que ninguna de las cadenas tenga la variante de secuencia proporciona una mayor probabilidad de descartar las lecturas de las cadenas que la lectura de una sola cadena que tenga la variante de secuencia, obteniendo así un número de lecturas de secuencia restantes;

comparar el número de lecturas de secuencias restantes con un umbral candidato; y
determinar si se descarta el locus candidato basándose en la comparación del número de lecturas de secuencia restantes con el umbral candidato, en donde el número de lecturas de secuencia restantes que es menor que el umbral candidato proporciona una mayor probabilidad de descartar el locus candidato que el número de lecturas de secuencia restantes que es mayor que el umbral candidato; e

identificar el conjunto de locus filtrado como portadores de mutaciones somáticas utilizando los locus candidatos restantes.

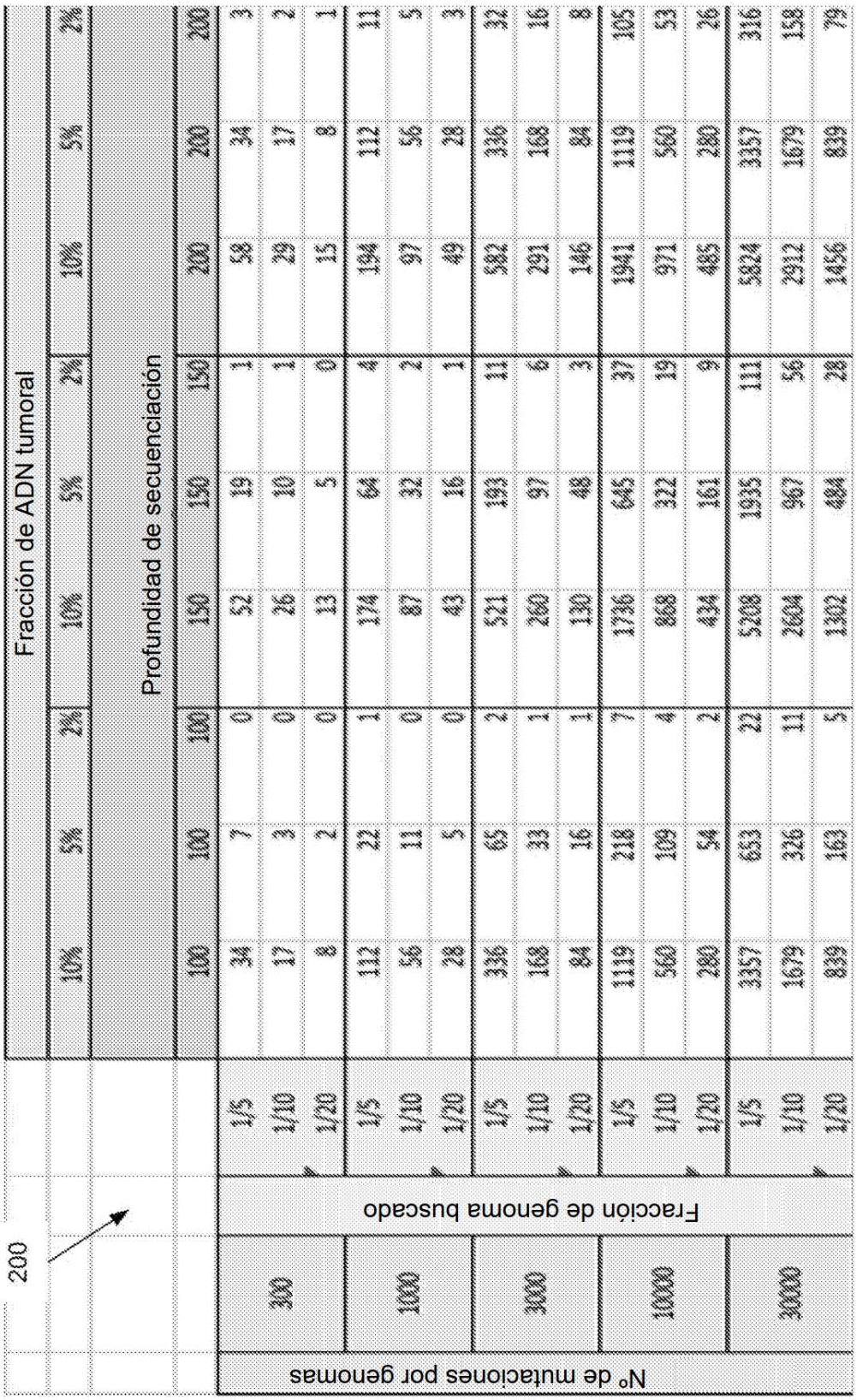
15. El método de una cualquiera de las reivindicaciones anteriores, en donde los fragmentos de ADN sin células de células tumorales o células asociadas con cáncer comprenden menos del 50 % de los fragmentos de ADN sin células en la muestra biológica.

16. Un programa informático que comprende una pluralidad de instrucciones capaces de ejecutarse mediante un sistema informático que, cuando se ejecutan, controlan el sistema informático para realizar una operación del método de una cualquiera de las reivindicaciones anteriores.

100

	Mama	Riñón	Colon	Hígado	Pulmón	Próstata	Cuello uterino	Estómago	Piel	Endometrio	Tiroides	Ovario
TP53	23%	6%	44%	27%	34%	13%	5%	33%	24%	18%	6%	46%
APC	2%	1%	45%	3%	2%	2%	2%	8%	6%	5%	3%	2%
BRAF	1%	2%	12%	1%	2%	1%	1%	1%	42%	2%	41%	7%
PTEN	4%	3%	5%	2%	2%	7%	4%	4%	8%	38%	2%	3%
VHL	0%	36%	1%	0%	0%	0%	0%	0%	1%	1%	0%	0%
KRAS	2%	0%	33%	2%	16%	4%	6%	6%	2%	15%	2%	12%
EGFR	1%	1%	2%	1%	29%	3%	1%	4%	3%	4%	1%	2%
PIK3CA	27%	2%	13%	3%	4%	2%	14%	10%	7%	23%	3%	9%
RET	1%	1%	4%	1%	2%	0%	1%	2%	3%	2%	26%	1%
ARID1A	4%	2%	11%	8%	4%	1%	4%	15%	6%	24%	0%	9%
PBRM1	1%	24%	4%	1%	2%	0%	1%	3%	4%	2%	0%	0%
CTNNB1	1%	7%	5%	23%	2%	3%	3%	8%	6%	18%	3%	6%
FOXL2	0%	0%	1%	0%	1%	0%	0%	1%	0%	0%	0%	21%
TSHR	1%	0%	2%	0%	2%	0%	0%	2%	4%	1%	21%	0%
TERT	3%	2%	3%	15%	1%	0%	1%	1%	20%	2%	11%	4%
CDKN2A	2%	2%	1%	5%	8%	1%	5%	4%	18%	2%	3%	5%
GRIN2A	2%	1%	8%	2%	5%	1%	2%	6%	17%	4%	1%	1%
PIK3R1	2%	0%	4%	1%	1%	0%	2%	1%	2%	16%	0%	1%
PTCH1	1%	1%	6%	1%	2%	0%	1%	5%	16%	3%	1%	1%
NRAS	1%	0%	4%	0%	1%	1%	1%	1%	16%	2%	7%	1%
ROS1	1%	1%	5%	1%	4%	1%	2%	5%	14%	4%	1%	2%
CDH1	11%	0%	3%	1%	1%	1%	1%	13%	2%	3%	0%	0%
FGFR3	0%	0%	2%	0%	0%	1%	1%	1%	13%	1%	0%	0%
KMT2C	9%	3%	11%	4%	8%	4%	8%	11%	12%	8%	3%	3%
NF1	3%	1%	7%	2%	6%	0%	3%	6%	12%	5%	1%	4%
HRAS	0%	0%	0%	0%	0%	2%	4%	1%	11%	0%	4%	0%
GATA3	11%	0%	3%	0%	2%	0%	0%	3%	2%	1%	1%	1%
KMT2D	3%	3%	10%	3%	6%	2%	8%	9%	11%	7%	0%	1%

FIG. 1



Nº de mutaciones por genomas	Fracción de genoma buscado	Fracción de ADN tumoral 10% Profundidad de secuenciación 100	Fracción de ADN tumoral 10% Profundidad de secuenciación 150	Fracción de ADN tumoral 10% Profundidad de secuenciación 200	Fracción de ADN tumoral 5% Profundidad de secuenciación 100	Fracción de ADN tumoral 5% Profundidad de secuenciación 150	Fracción de ADN tumoral 5% Profundidad de secuenciación 200	Fracción de ADN tumoral 2% Profundidad de secuenciación 100	Fracción de ADN tumoral 2% Profundidad de secuenciación 150	Fracción de ADN tumoral 2% Profundidad de secuenciación 200
300	1/5	34	52	34	7	19	34	0	1	3
	1/10	17	26	17	3	10	17	0	1	2
	1/20	8	13	8	2	5	8	0	0	1
1000	1/5	112	174	112	22	64	112	1	4	11
	1/10	56	87	56	11	32	56	0	2	5
	1/20	28	43	28	5	16	28	0	1	3
3000	1/5	336	521	336	65	193	336	2	11	32
	1/10	168	260	168	33	97	168	1	6	16
	1/20	84	130	84	16	48	84	1	3	8
10000	1/5	1119	1736	1119	218	645	1119	7	37	105
	1/10	560	868	560	109	322	560	4	19	53
	1/20	280	434	280	54	161	280	2	9	26
30000	1/5	3357	5208	3357	653	1935	3357	22	111	316
	1/10	1679	2604	1679	326	967	1679	11	56	158
	1/20	839	1302	839	163	484	839	5	28	79

FIG. 2

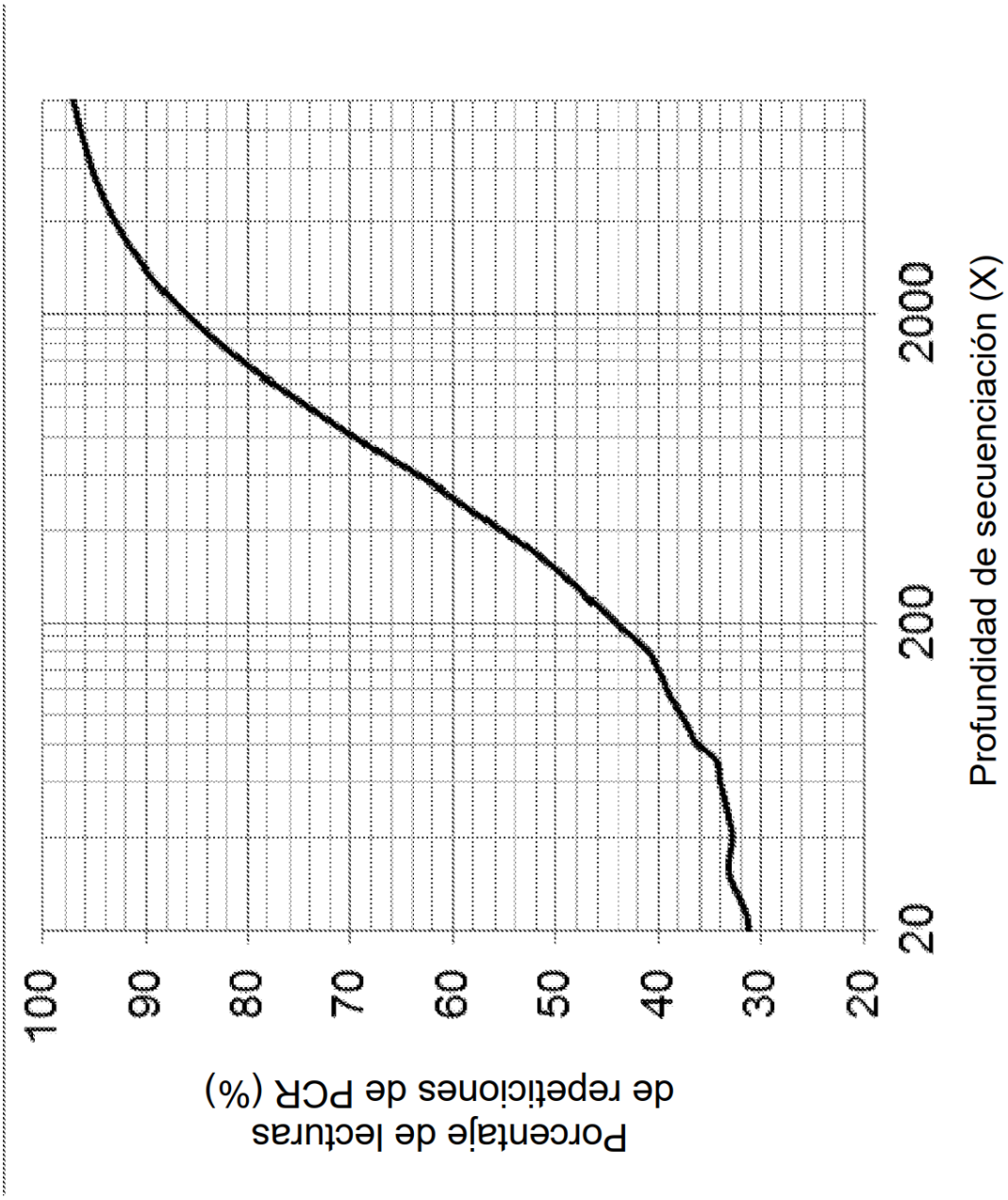

300
Porcentaje de lecturas de repeticiones de PCR (%)
100
90
80
70
60
50
40
30
20
20
200
2000
Profundidad de secuenciación (X)

FIG. 3

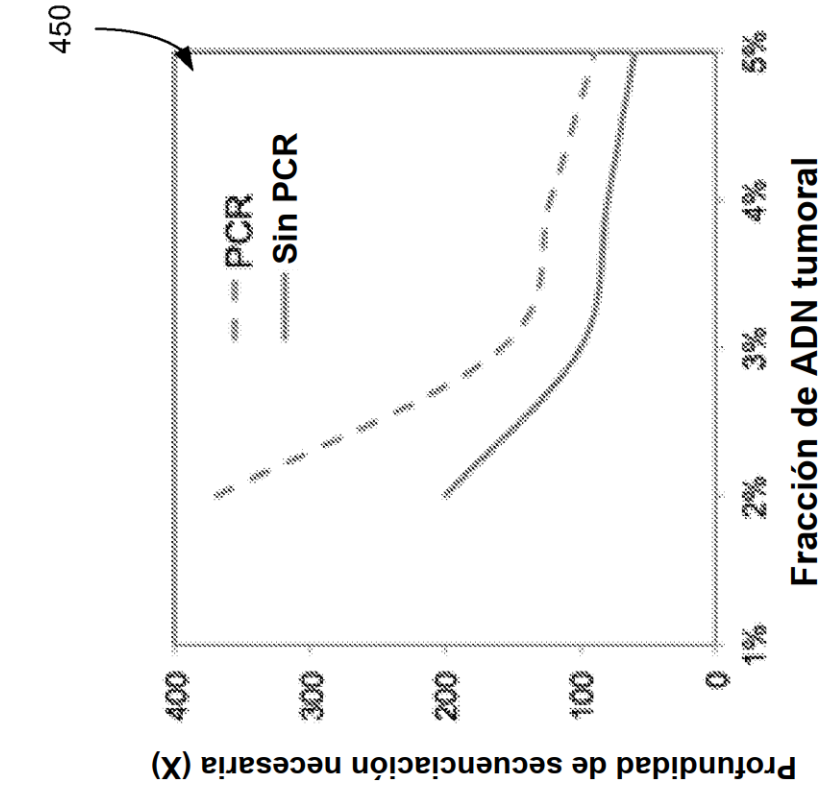

450
PCR
Sin PCR
400
300
200
100
0
1%
2%
3%
4%
5%
Fracción de ADN tumoral
Profundidad de secuenciación necesaria (X)

FIG. 4B

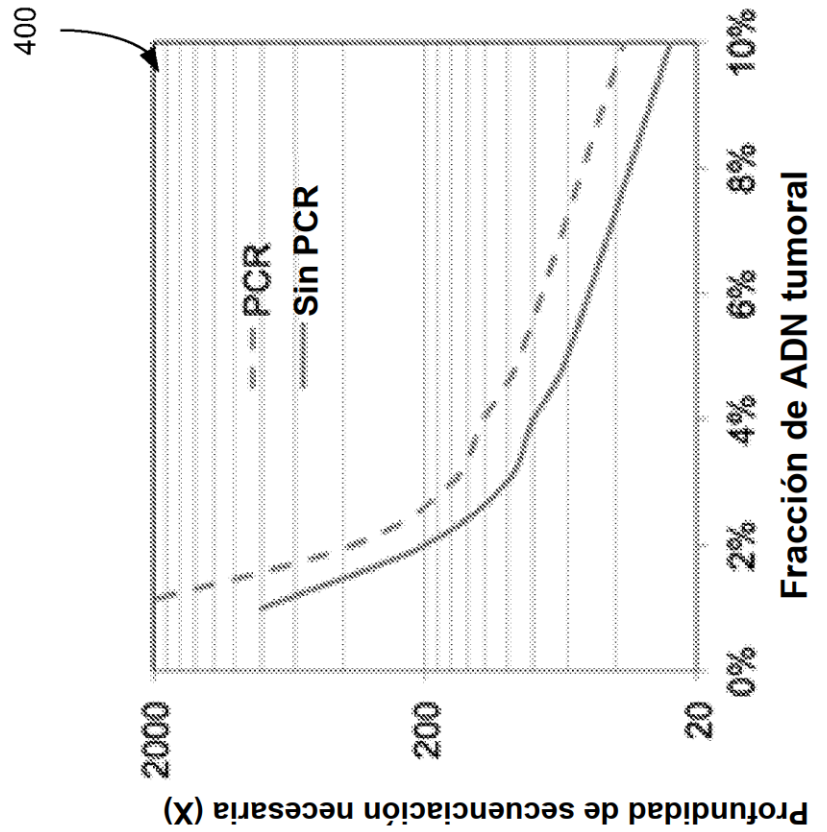

400
PCR
Sin PCR
2000
200
20
0%
2%
4%
6%
8%
10%
Fracción de ADN tumoral
Profundidad de secuenciación necesaria (X)

FIG. 4A

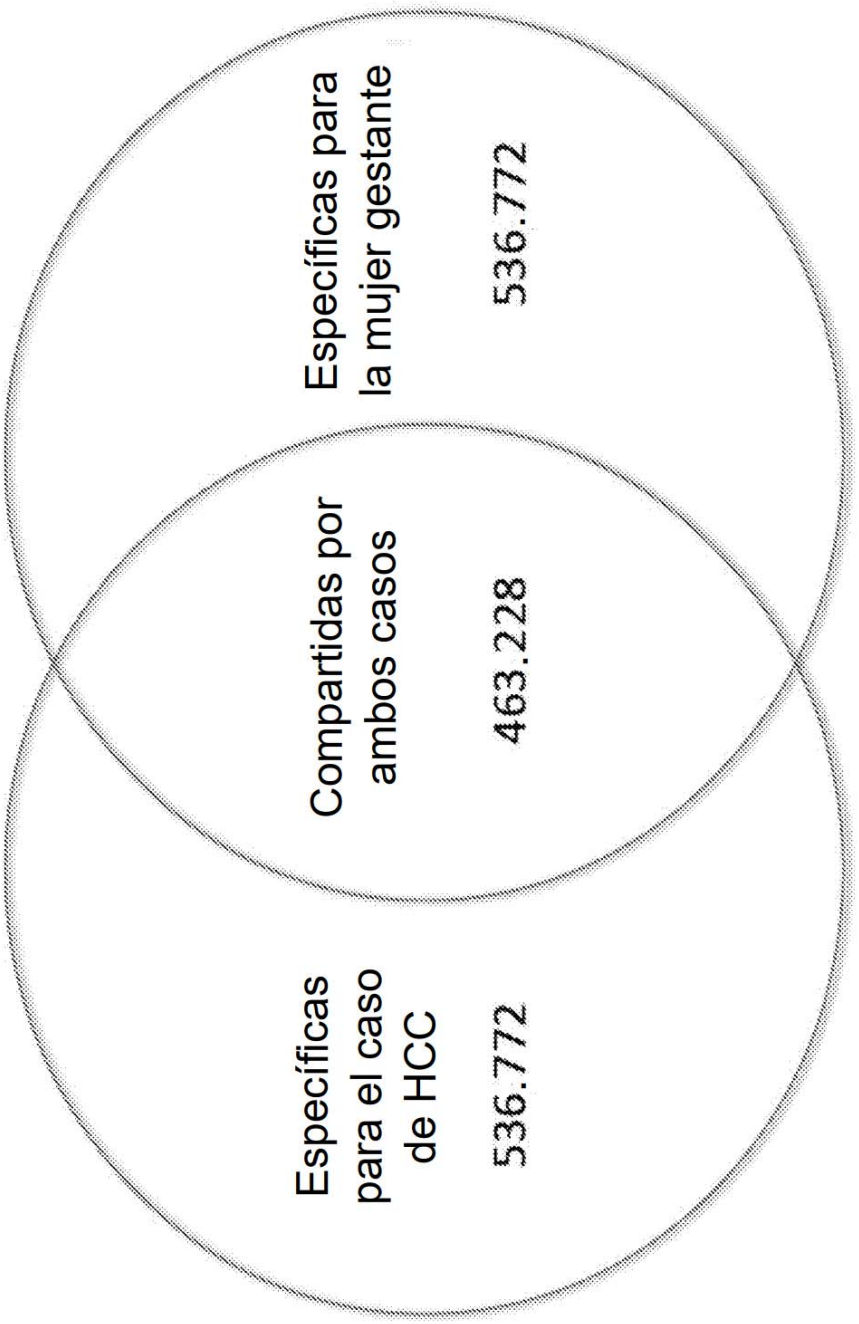

Específicas para la mujer gestante
536.772
Compartidas por ambos casos
463.228
Específicas para el caso de HCC
536.772

FIG. 5

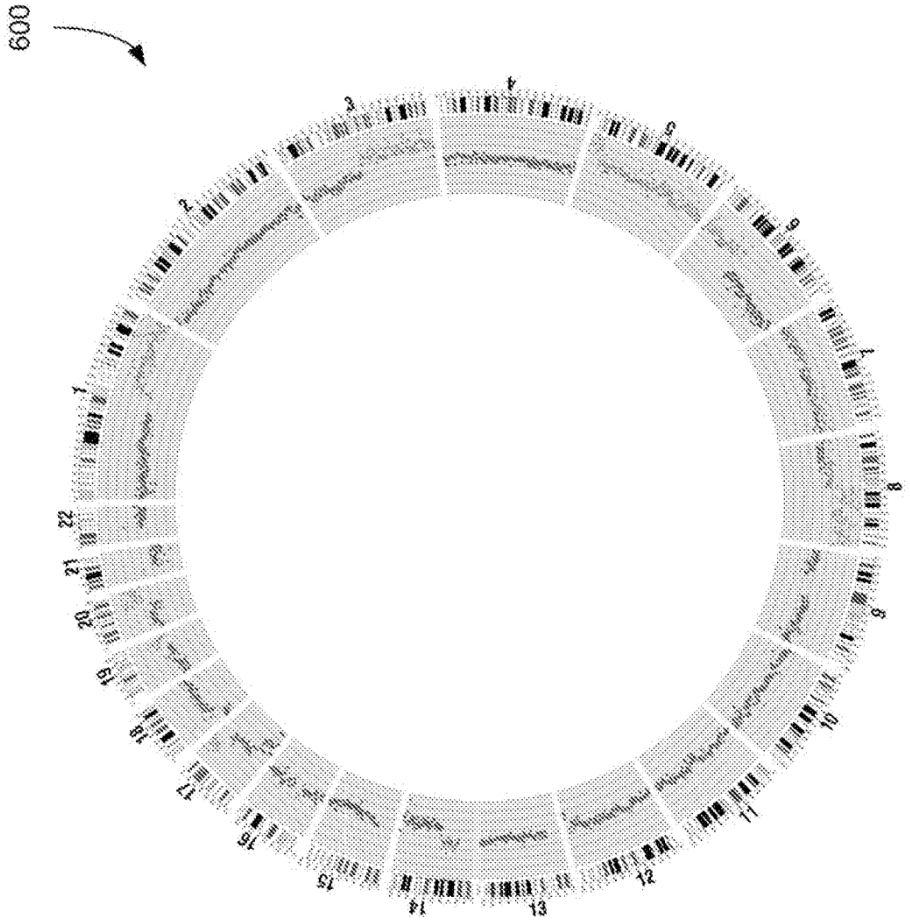


FIG. 6

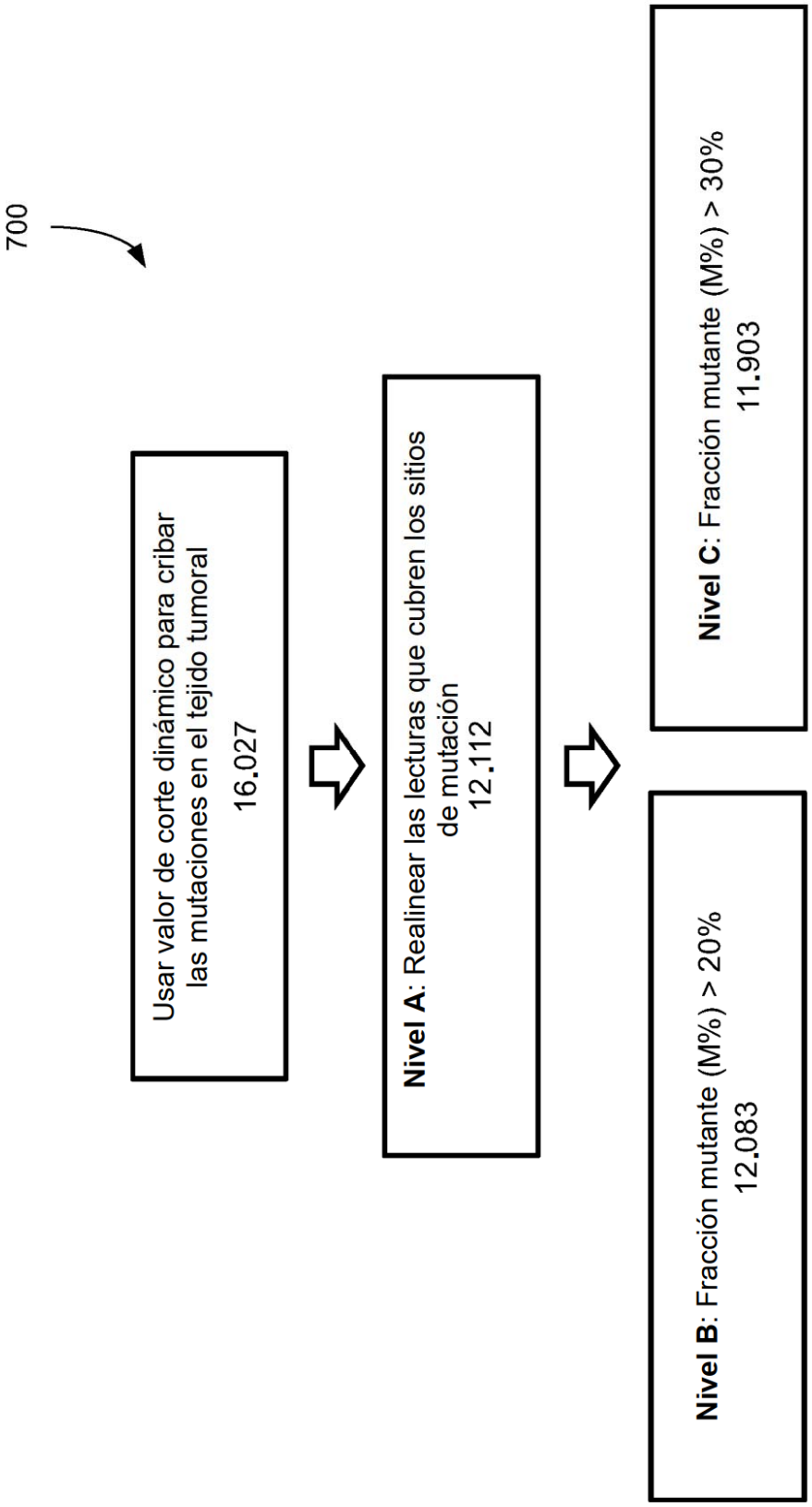

700
Usar valor de corte dinámico para cribar las mutaciones en el tejido tumoral
16.027
Nivel A: Realinear las lecturas que cubren los sitios de mutación
12.112
Nivel B: Fracción mutante (M%) > 20%
12.083
Nivel C: Fracción mutante (M%) > 30%
11.903

FIG. 7

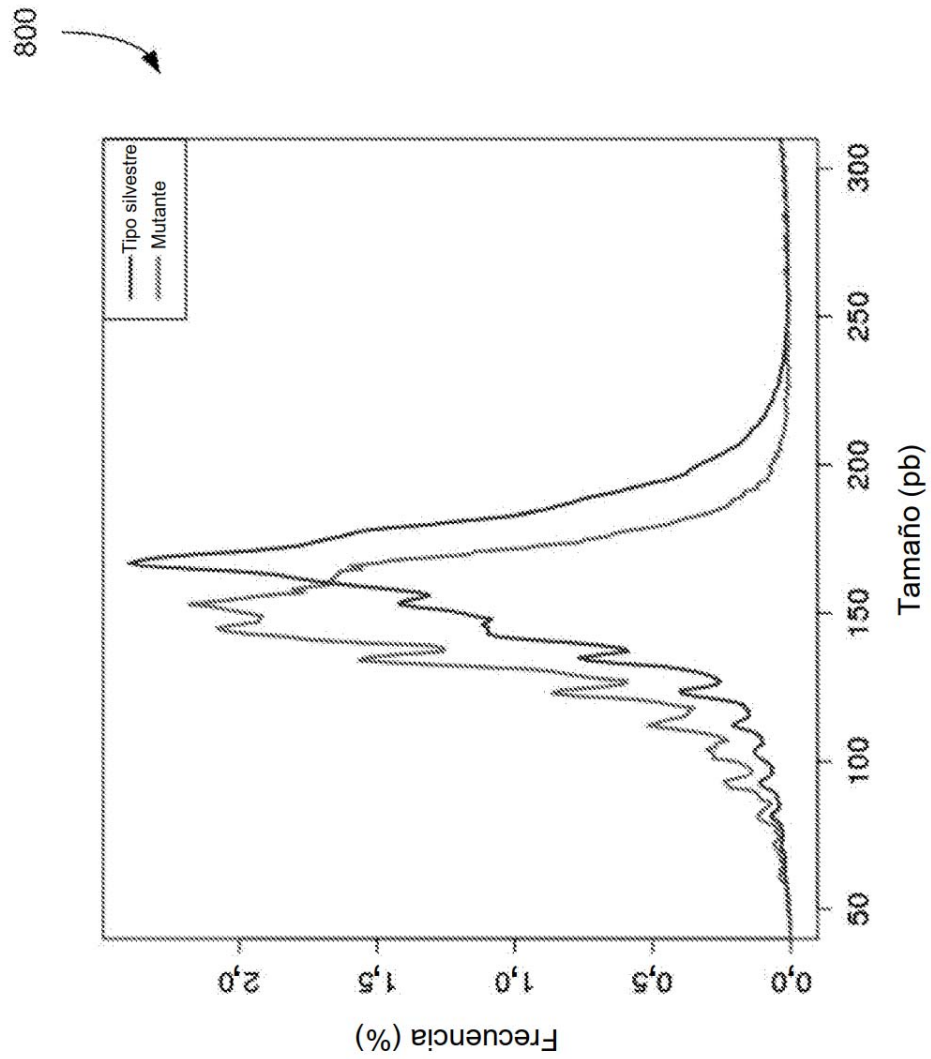

800
Tipo silvestre
Mutante
Frecuencia (%)
0,0
0,5
1,0
1,5
2,0
50
100
150
200
250
300
Tamaño (pb)

FIG. 8

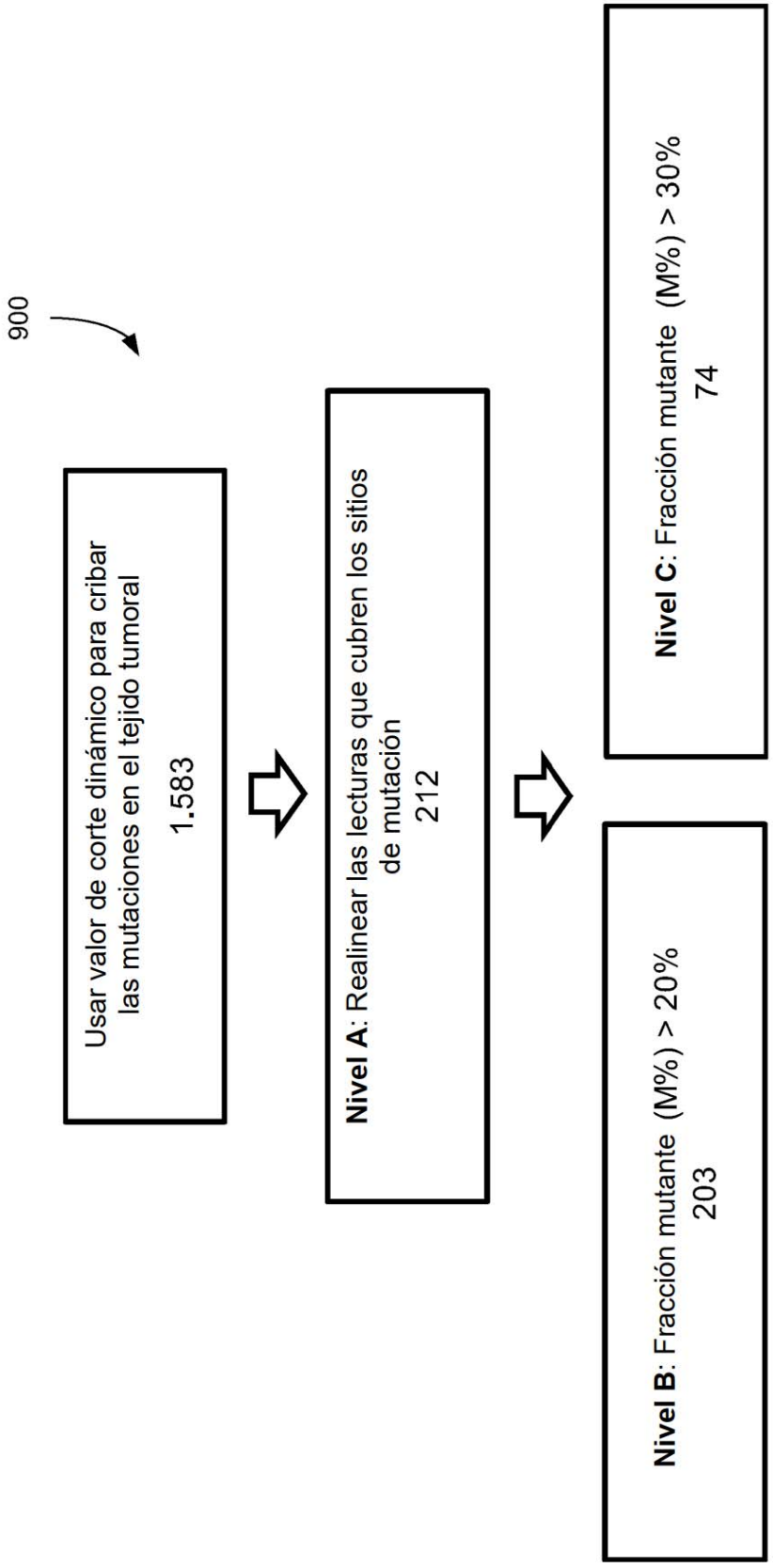

900
Usar valor de corte dinámico para cribar las mutaciones en el tejido tumoral
1.583
Nivel A: Realinear las lecturas que cubren los sitios de mutación
212
Nivel B: Fracción mutante (M%) > 20%
203
Nivel C: Fracción mutante (M%) > 30%
74

FIG. 9

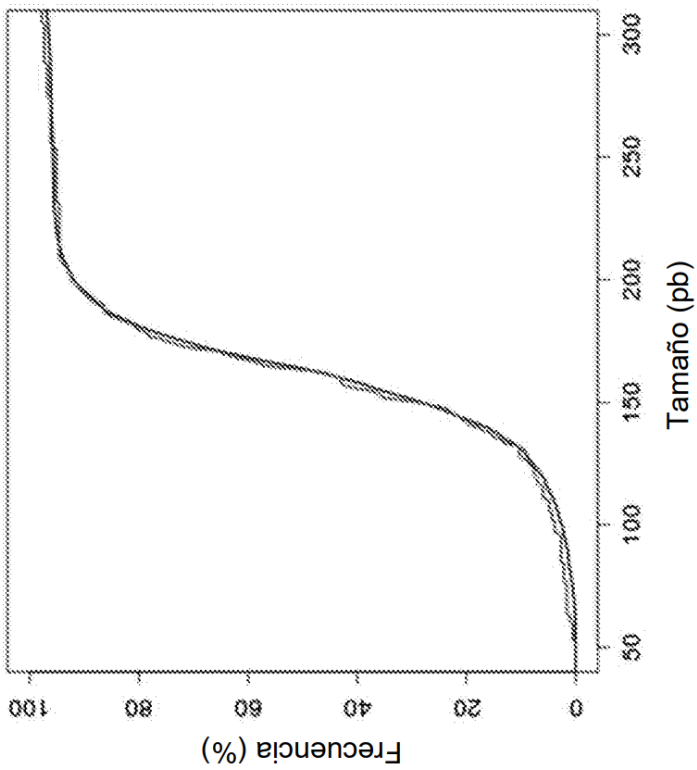


FIG. 10B

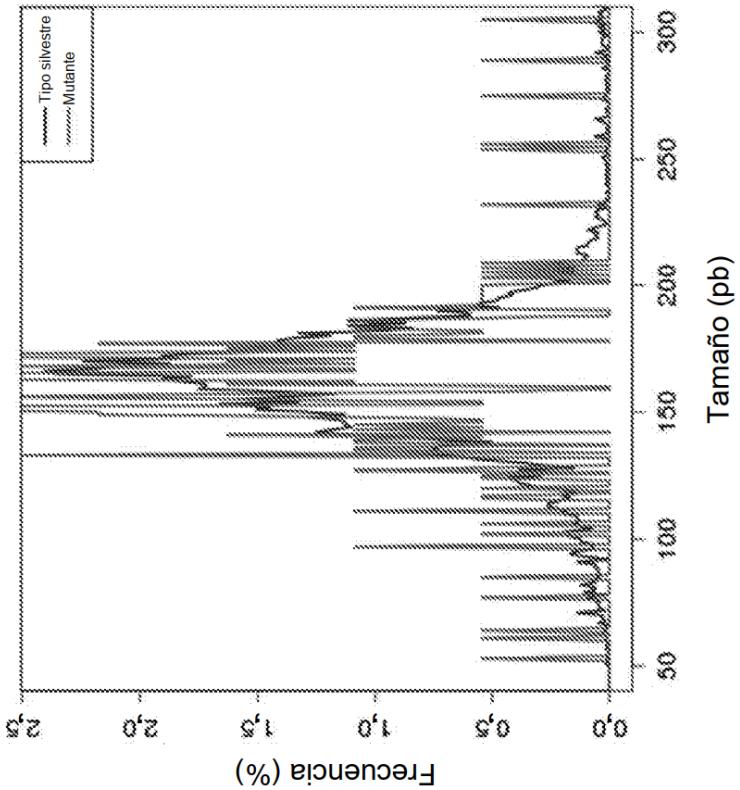


FIG. 10A

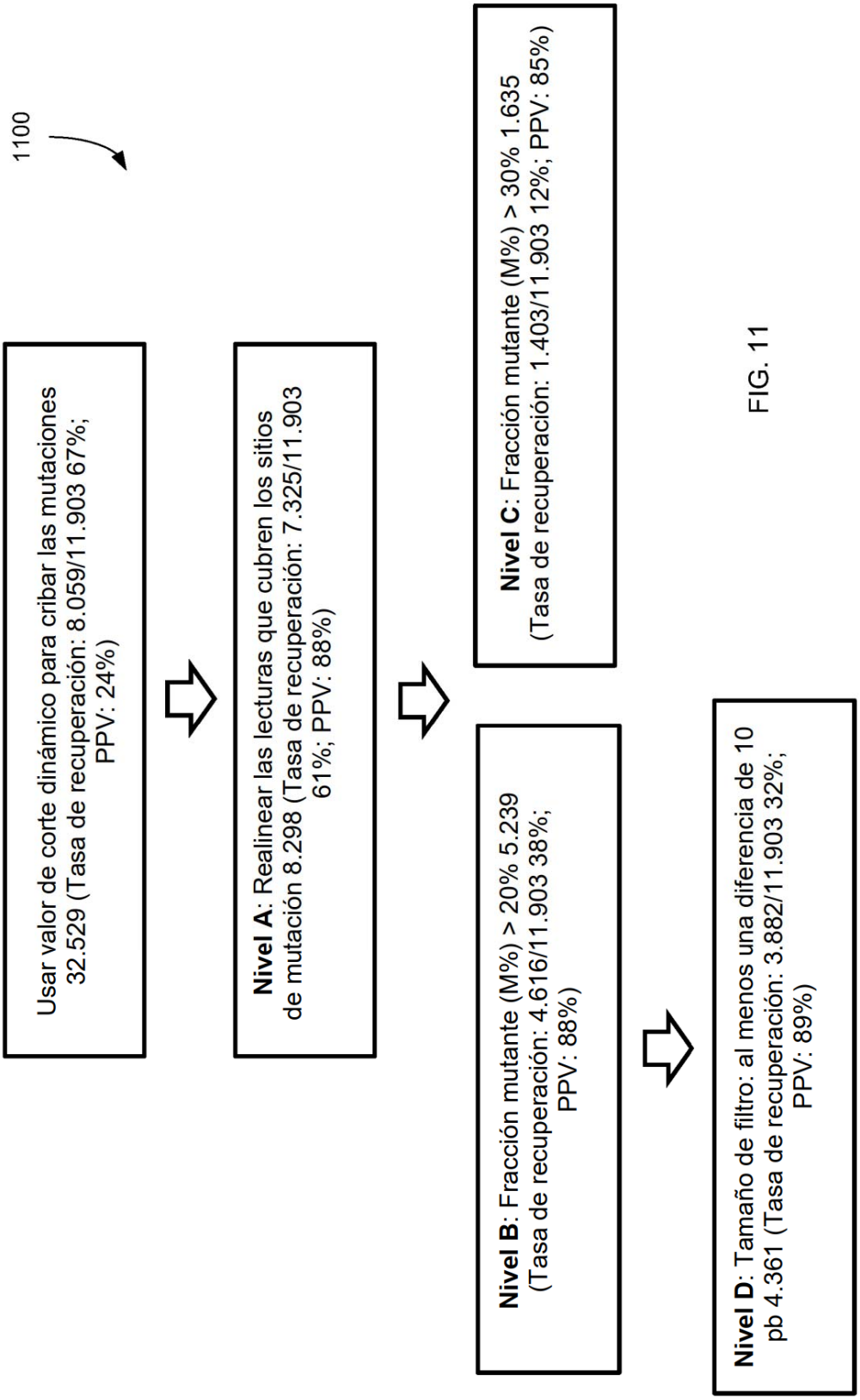

1100
Usar valor de corte dinámico para cribar las mutaciones 32.529 (Tasa de recuperación: 8.059/11.903 67%; PPV: 24%)
Nivel A: Realinear las lecturas que cubren los sitios de mutación 8.298 (Tasa de recuperación: 7.325/11.903 61%; PPV: 88%)
Nivel C: Fracción mutante (M%) > 30% 1.635 (Tasa de recuperación: 1.403/11.903 12%; PPV: 85%)
Nivel B: Fracción mutante (M%) > 20% 5.239 (Tasa de recuperación: 4.616/11.903 38%; PPV: 88%)
Nivel D: Tamaño de filtro: al menos una diferencia de 10 pb 4.361 (Tasa de recuperación: 3.882/11.903 32%; PPV: 89%)

FIG. 11

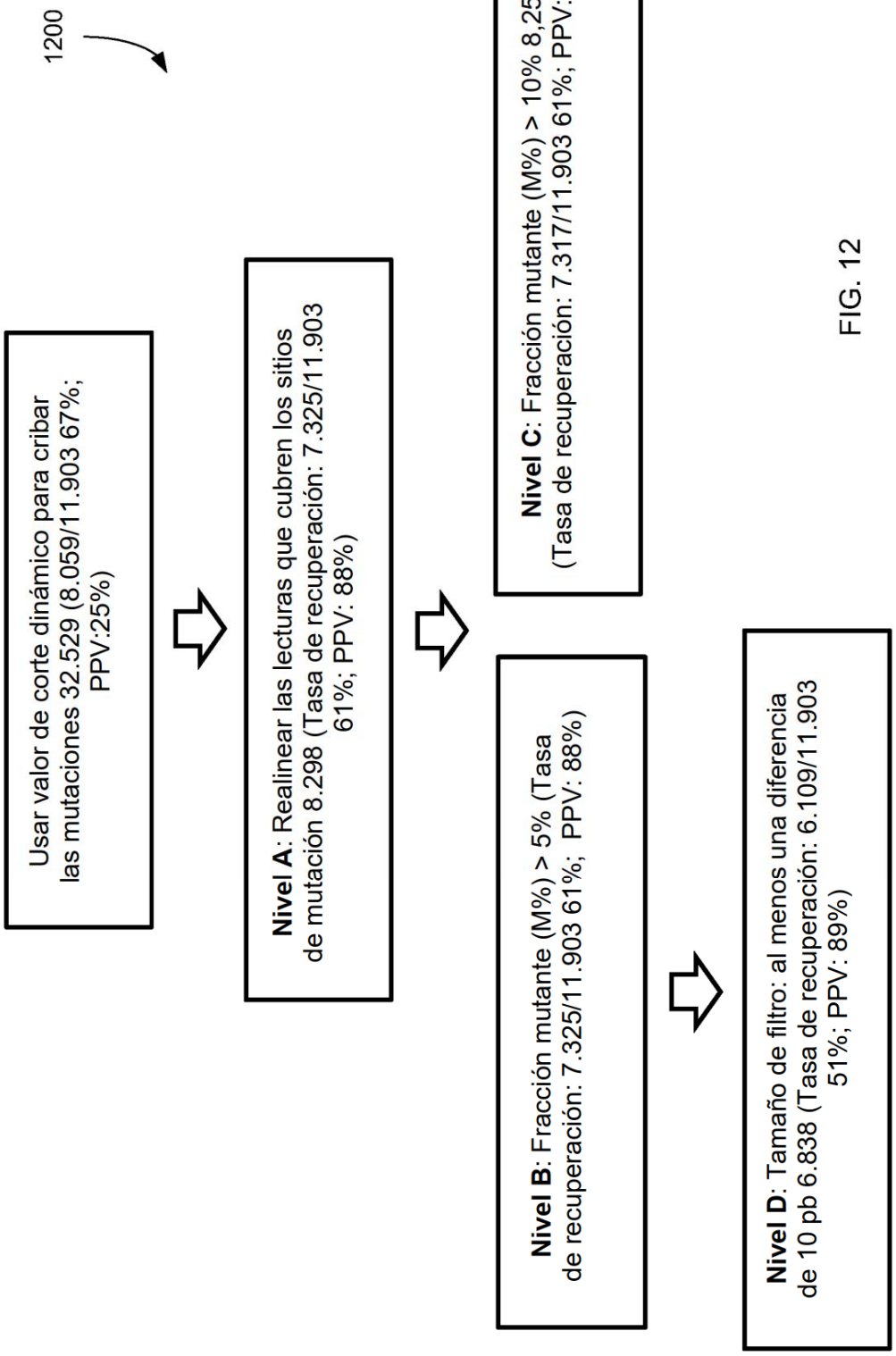

1200
Usar valor de corte dinámico para cribar las mutaciones 32.529 (8.059/11.903 67%; PPV:25%)
Nivel A: Realinear las lecturas que cubren los sitios de mutación 8.298 (Tasa de recuperación: 7.325/11.903 61%; PPV: 88%)
Nivel C: Fracción mutante (M%) > 10% 8,257 (Tasa de recuperación: 7.317/11.903 61%; PPV: 88%)
Nivel B: Fracción mutante (M%) > 5% (Tasa de recuperación: 7.325/11.903 61%; PPV: 88%)
Nivel D: Tamaño de filtro: al menos una diferencia de 10 pb 6.838 (Tasa de recuperación: 6.109/11.903 51%; PPV: 89%)

FIG. 12

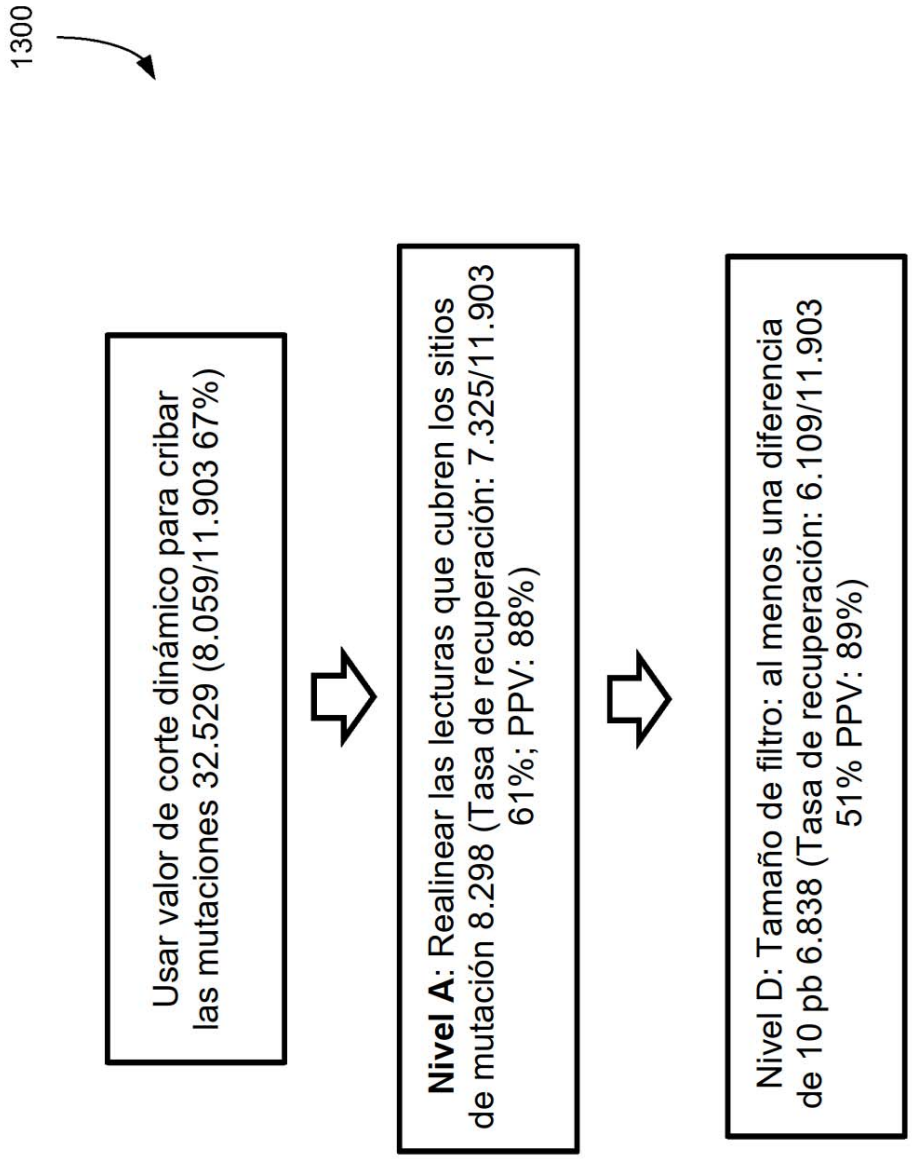

1300
Usar valor de corte dinámico para cribar las mutaciones 32.529 (8.059/11.903 67%)
Nivel A: Realinear las lecturas que cubren los sitios de mutación 8.298 (Tasa de recuperación: 7.325/11.903 61%; PPV: 88%)
Nivel D: Tamaño de filtro: al menos una diferencia de 10 pb 6.838 (Tasa de recuperación: 6.109/11.903 51% PPV: 89%)

FIG. 13

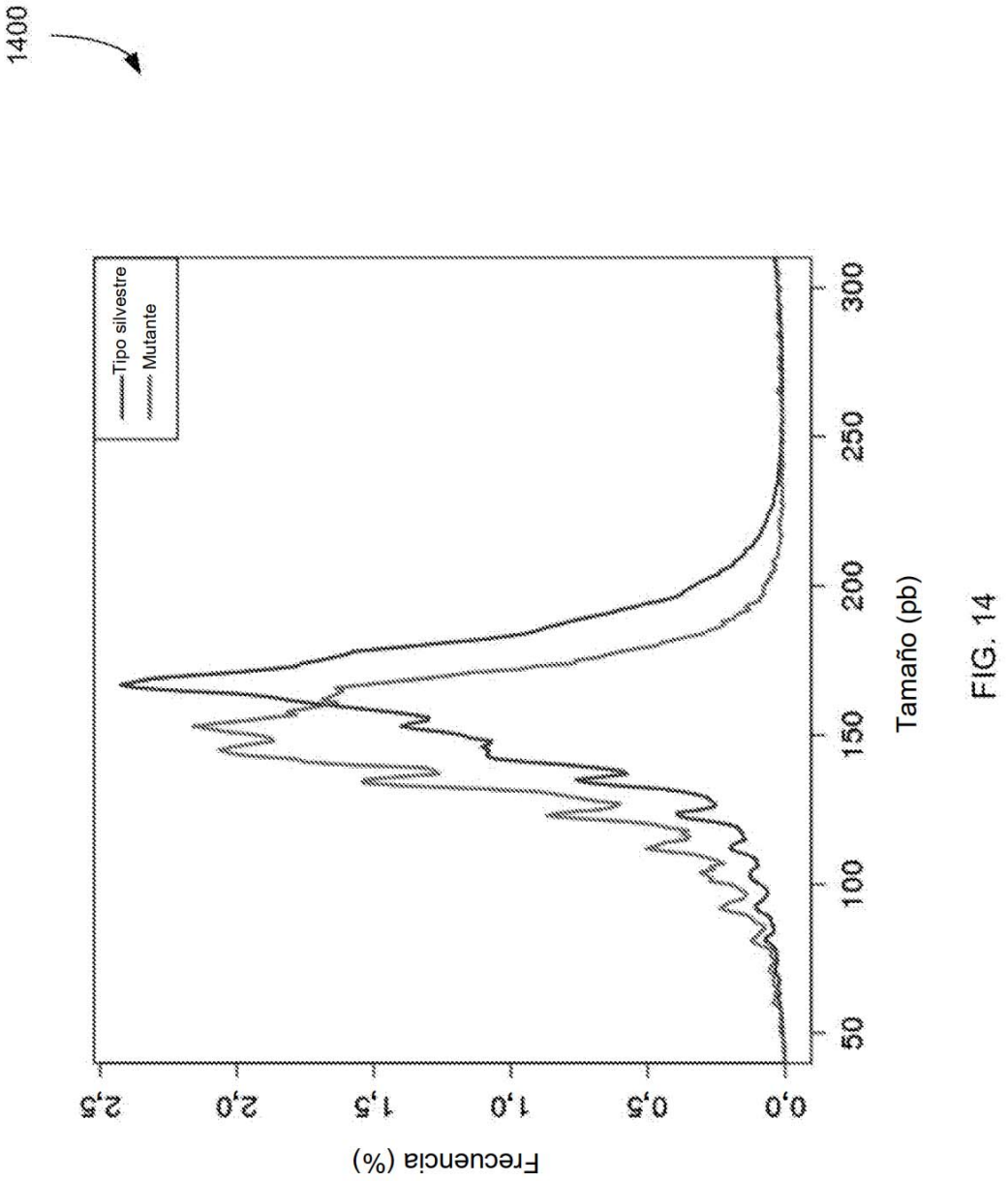

1400
Tipo silvestre
Mutante
Frecuencia (%)
0,0
0,5
1,0
1,5
2,0
2,5
50
100
150
200
250
300
Tamaño (pb)

FIG. 14

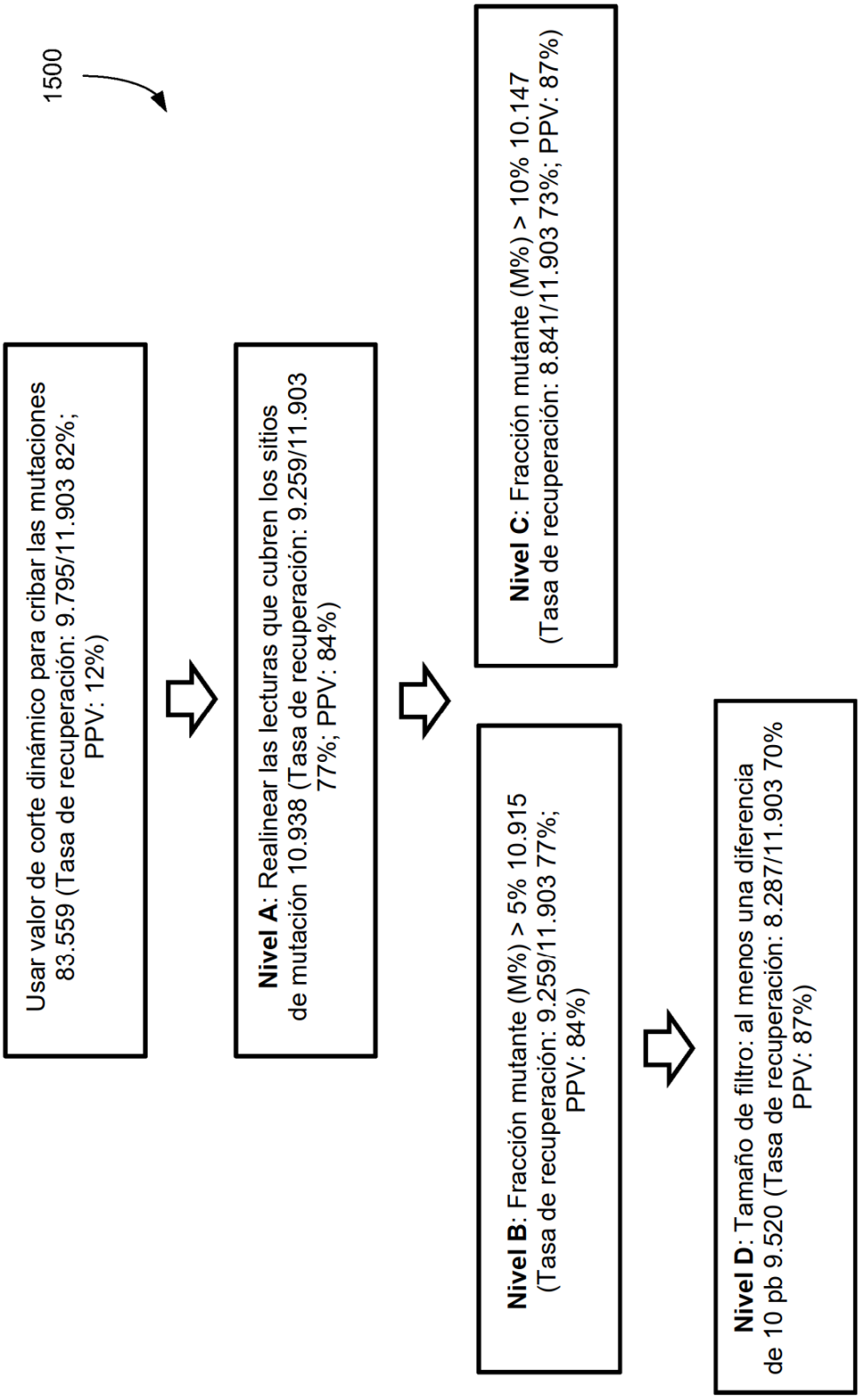

1500
Usar valor de corte dinámico para cribar las mutaciones 83.559 (Tasa de recuperación: 9.795/11.903 82%; PPV: 12%)
Nivel A: Realinear las lecturas que cubren los sitios de mutación 10.938 (Tasa de recuperación: 9.259/11.903 77%; PPV: 84%)
Nivel B: Fracción mutante (M%) > 5% 10.915 (Tasa de recuperación: 9.259/11.903 77%; PPV: 84%)
Nivel C: Fracción mutante (M%) > 10% 10.147 (Tasa de recuperación: 8.841/11.903 73%; PPV: 87%)
Nivel D: Tamaño de filtro: al menos una diferencia de 10 pb 9.520 (Tasa de recuperación: 8.287/11.903 70% PPV: 87%)

FIG. 15

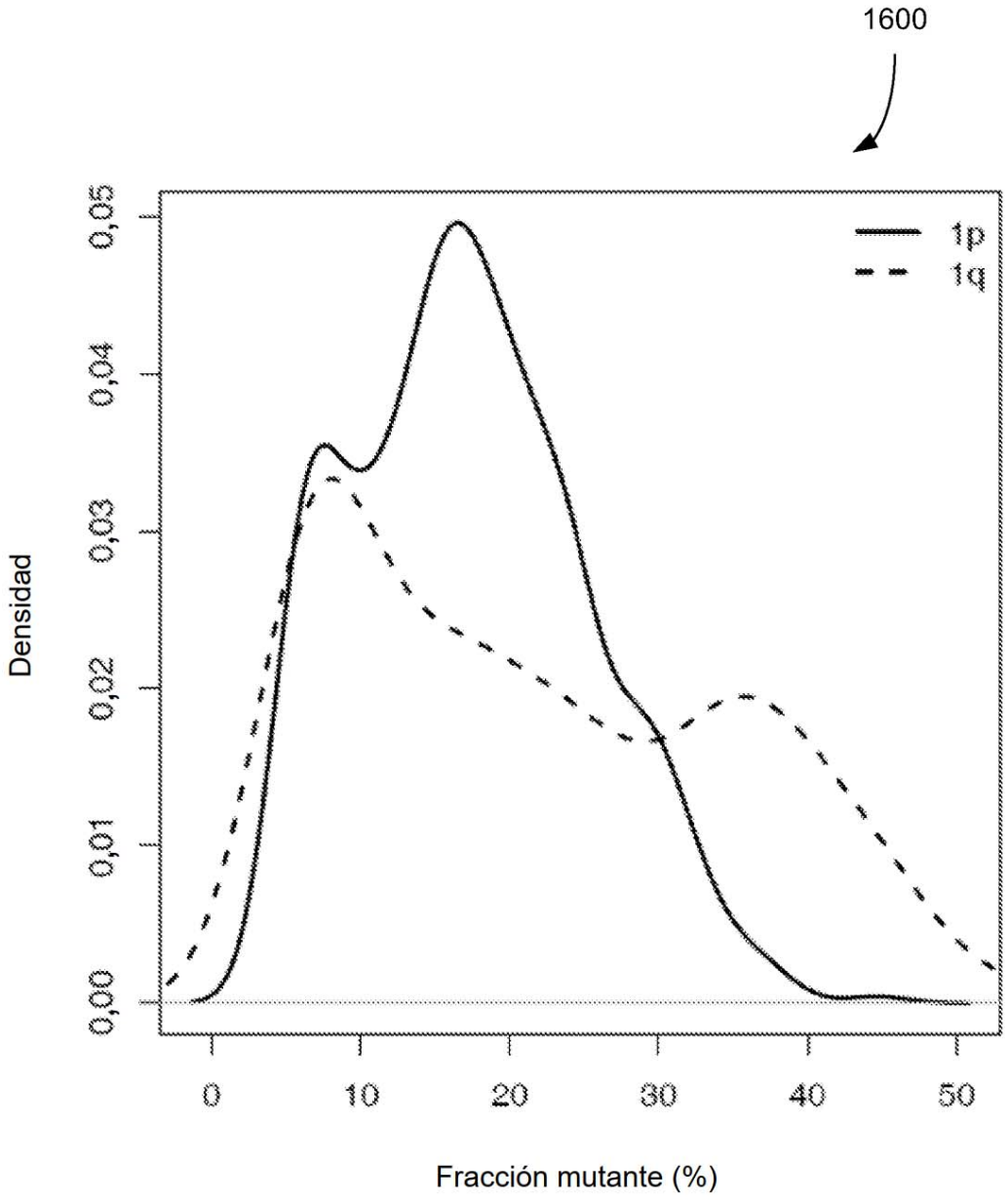

1600
1p
1q
Densidad
0,00
0,01
0,02
0,03
0,04
0,05
0
10
20
30
40
50
Fracción mutante (%)

FIG. 16

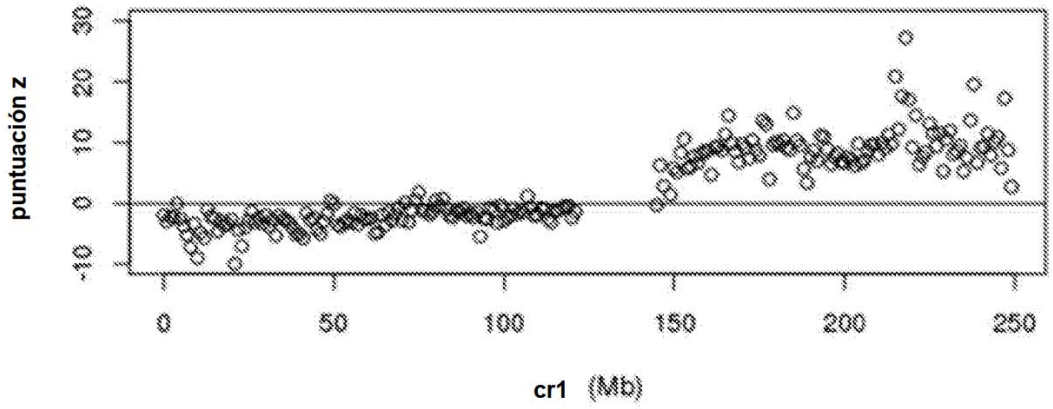


FIG. 17A

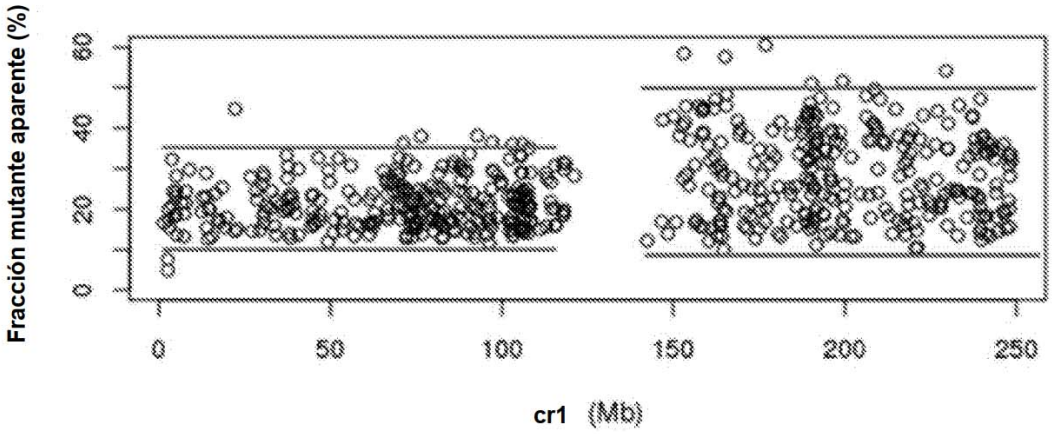


FIG. 17B

1800

FRP<1e-10 y puntuación1> 0,01			Sensibilidad prevista de la detección de mutaciones en plasma (%)					
Profundidad de secuenciación	Valor de corte para alelos mutantes		1%	2%	5%	10%	15%	20%
	Capa leucocitaria	Plasma						
50	<=1	>=6	0,0001	0,0	1,4	23,8	62,2	87,0
100	<=1	>=7	0,0010	0,1	13,3	78,0	98,2	99,9
150	<=2	>=7	0,0170	1,2	47,5	98,2	100	100
200	<=2	>=8	0,0237	2,1	66,7	99,8	100	100
250	<=2	>=9	0,0277	3,2	79,9	100	100	100
500	<=3	>=10	1,3695	41,7	99,9	100	100	100
600	<=4	>=11	2,0092	53,8	100	100	100	100
1000	<=6	>=12	41,6960	98,9	100	100	100	100

FIG. 18

1900

FRP<0,1% y puntuación1>0,01			Sensibilidad prevista de la detección de mutaciones en plasma (%)					
	Valor de corte para alelos mutantes		1%	2%	5%	10%	15%	20%
Profundidad de secuenciación	Capa leucocitaria	Plasma						
50	<=1	>=3	0,175	1,899	24,242	73,497	94,085	98,966
100	<=1	>=3	1,90	14,29	73,50	98,97	99,98	100
150	<=2	>=4	1,86	18,47	86,79	99,91	100	100
200	<=2	>=4	5,27	37,12	97,07	100	100	100
250	<=2	>=4	10,9	55,95	99,47	100	100	100
500	<=3	>=5	38,40	93,29	100	100	100	100
600	<=4	>=6	39,37	95,42	100	100	100	100
1000	<=6	>=7	77,98	99,92	100	100	100	100

FIG. 19

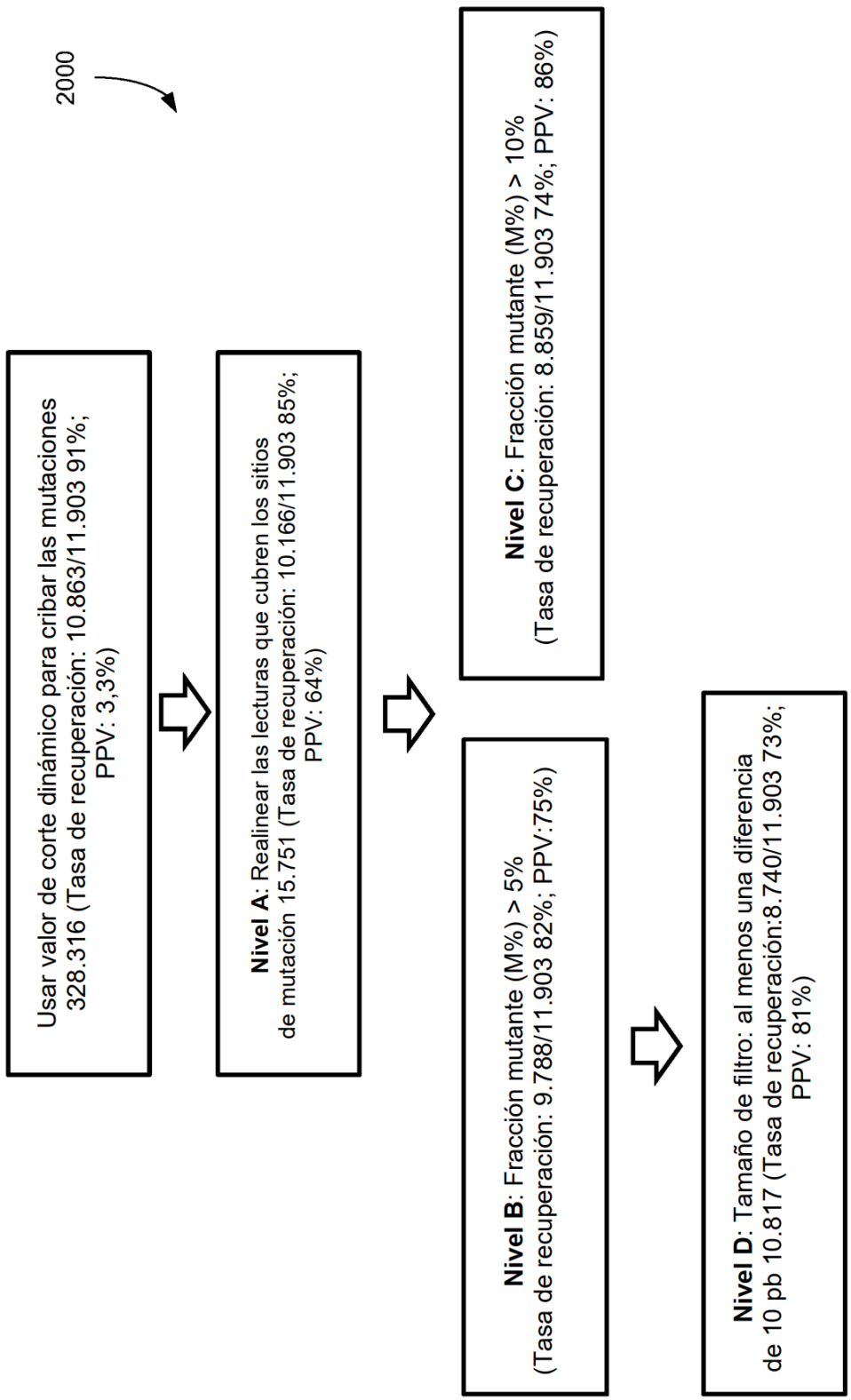

2000
Usar valor de corte dinámico para cribar las mutaciones
328.316 (Tasa de recuperación: 10.863/11.903 91%;
PPV: 3,3%)
Nivel A: Realinear las lecturas que cubren los sitios
de mutación 15.751 (Tasa de recuperación: 10.166/11.903 85%;
PPV: 64%)
Nivel C: Fracción mutante (M%) > 10%
(Tasa de recuperación: 8.859/11.903 74%; PPV: 86%)
Nivel B: Fracción mutante (M%) > 5%
(Tasa de recuperación: 9.788/11.903 82%; PPV:75%)
Nivel D: Tamaño de filtro: al menos una diferencia
de 10 pb 10.817 (Tasa de recuperación:8.740/11.903 73%;
PPV: 81%)

FIG. 20

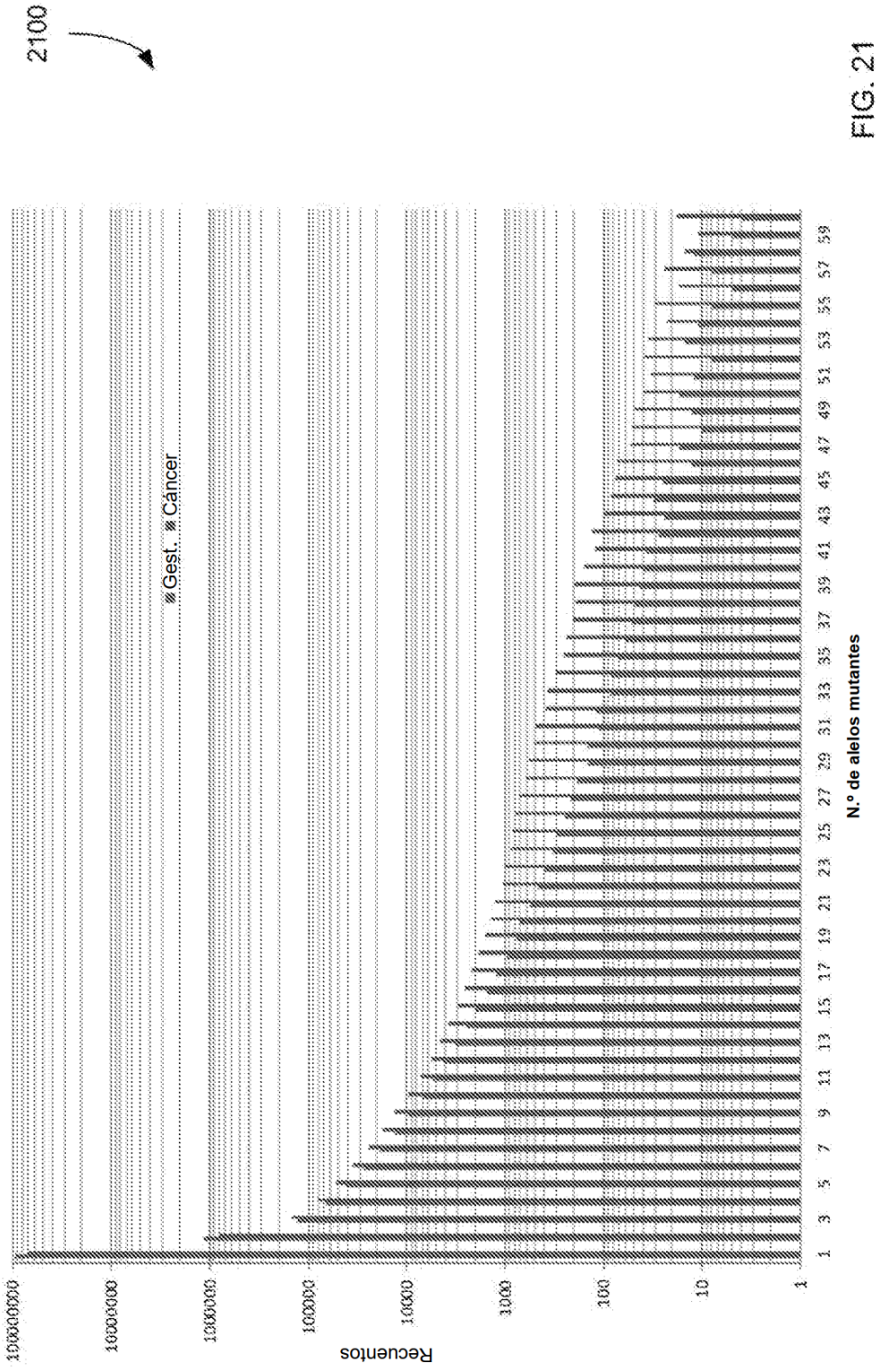


FIG. 21

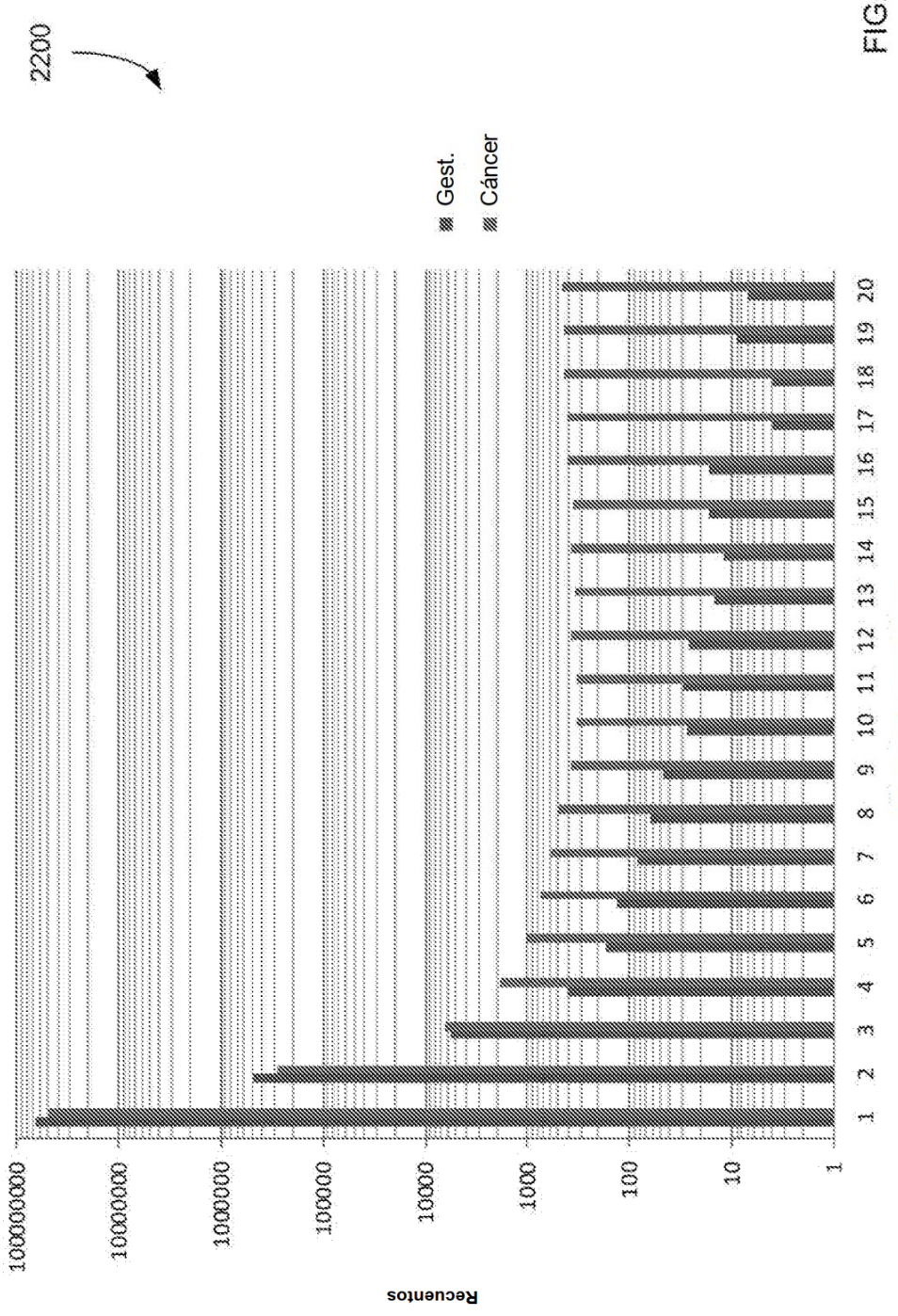

2200
Gest.
Cáncer
Recuentos
100000000
10000000
1000000
100000
10000
1000
100
10
1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
N.º de alelos mutantes

FIG. 22

Valor de corte fijado					Tamaño de filtro				
Valor de corte utilizado para identificación de mutaciones (al menos N alelos mutantes encontrados en plasma)	El número de supuestas mutaciones identificadas en plasma	El número de mutaciones somáticas verificadas en tejido tumoral	PPV%	Tasa de recuperación %	Valor de corte de tamaño de la menos una diferencia de X pb	El número de supuestas mutaciones identificadas en el genoma humano	El número de mutaciones somáticas verificadas en tejido tumoral	PPV%	Tasa de recuperación %
2	1238513	11395	0,92	95,73	5	616199	10822	1,76	90,9
					10	500002	10117	2,02	85,0
					15	393749	8583	2,18	72,1
					20	302989	6176	2,04	51,9
3	423581	11114	2,62	93,37	5	238275	10595	4,45	89,0
					10	186803	9901	5,30	83,2
					15	140332	8404	5,99	70,6
					20	100861	6031	5,98	50,7
4	278504	10847	3,89	91,13	5	149269	10358	6,94	87,0
					10	115078	9680	8,41	81,3
					15	84331	8215	9,74	69,0
					20	58226	5878	10,10	49,4
5	197631	10590	5,36	88,97	5	106435	10133	9,52	85,1
					10	81133	9475	11,68	79,6
					15	58571	8037	13,72	67,5
					20	39388	5730	14,55	48,1
6	145772	10375	7,12	87,16	5	79705	9930	12,46	83,4
					10	60411	9285	15,37	78,0
					15	43200	7872	18,22	66,1
					20	28653	5600	19,54	47,0

FIG. 23

2300

Valor de corte fijado					Tamaño de filtro				
Valor de corte utilizado para identificación de mutaciones (al menos N alelos mutantes encontrados en plasma)	El número de supuestas mutaciones identificadas en plasma	El número de mutaciones somáticas verificadas en tejido tumoral	PPV%	Tasa de recuperación %	Valor de corte de tamaño de la menos una diferencia de X pb	El número de supuestas mutaciones identificadas en el genoma humano	El número de mutaciones somáticas verificadas en tejido tumoral	PPV%	Tasa de recuperación %
2	303265	10587	3,49	88,94	5	156732	10072	6,43	84,6
					10	127277	9417	7,40	79,1
					15	99588	7986	8,02	67,1
					20	75324	5712	7,58	48,0
3	21837	10388	47,57	87,27	5	28643	9921	34,64	83,3
					10	24250	9273	38,24	77,9
					15	19336	7868	40,69	66,1
					20	14012	5619	40,10	47,2
4	15493	10152	65,53	85,29	5	14896	9712	65,20	81,6
					10	13382	9077	67,83	76,3
					15	11115	7702	69,29	64,7
					20	7949	5486	69,01	46,1
5	13629	9936	72,90	83,47	5	13121	9519	72,55	80,0
					10	11896	8904	74,85	74,8
					15	9925	7556	76,13	63,5
					20	7061	5367	76,01	45,1
6	12596	9765	77,52	82,04	5	12277	9359	76,23	78,6
					10	11143	8754	78,56	73,5
					15	9283	7425	79,98	62,4
					20	6574	5266	80,10	44,2

FIG. 24

2400

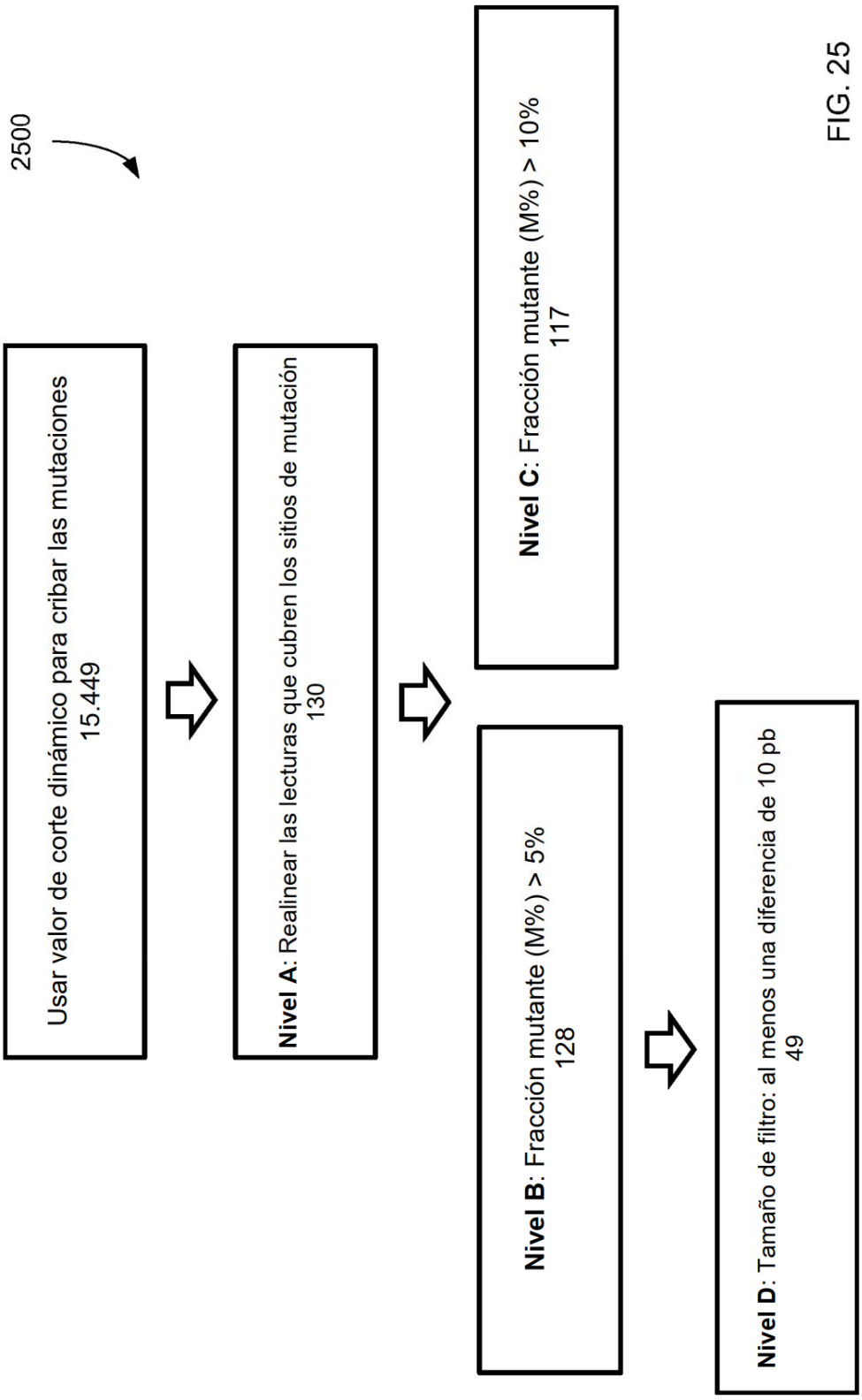

2500
Usar valor de corte dinámico para cribar las mutaciones
15.449
Nivel A: Realinear las lecturas que cubren los sitios de mutación
130
Nivel C: Fracción mutante (M%) > 10%
117
Nivel B: Fracción mutante (M%) > 5%
128
Nivel D: Tamaño de filtro: al menos una diferencia de 10 pb
49

FIG. 25

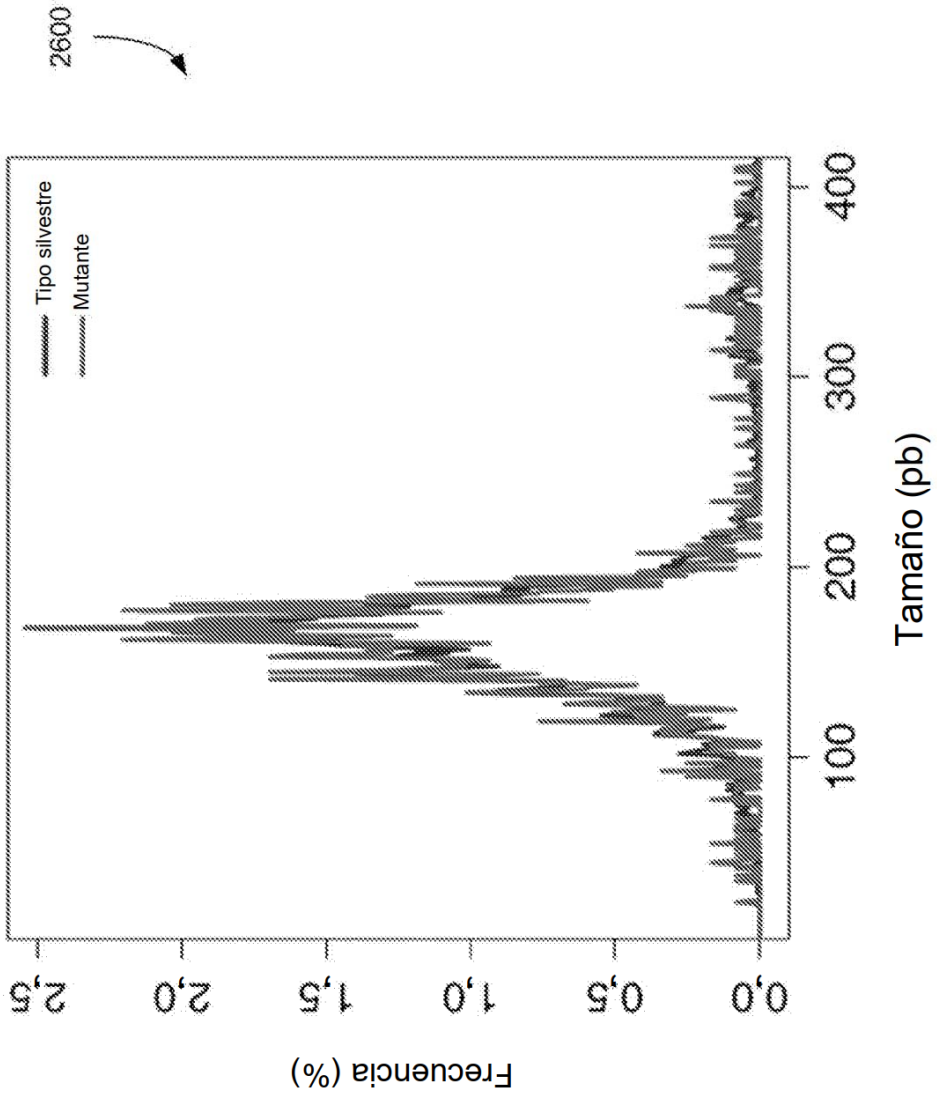

2600
Tipo silvestre
Mutante
0,0
0,5
1,0
1,5
2,0
2,5
Frecuencia (%)
100
200
300
400
Tamaño (pb)

FIG. 26

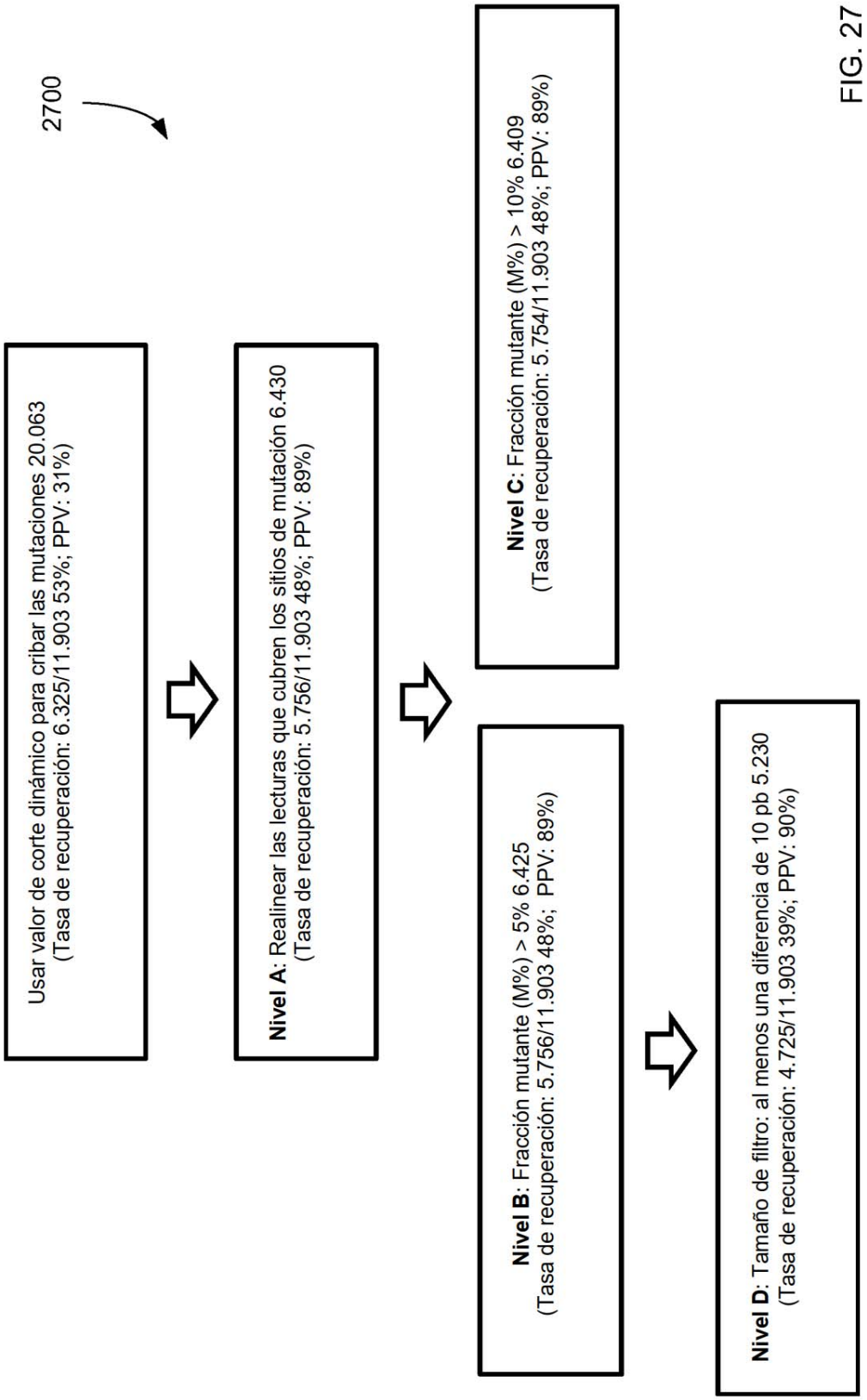

2700
Usar valor de corte dinámico para cribar las mutaciones 20.063
(Tasa de recuperación: 6.325/11.903 53%; PPV: 31%)
Nivel A: Realinear las lecturas que cubren los sitios de mutación 6.430
(Tasa de recuperación: 5.756/11.903 48%; PPV: 89%)
Nivel C: Fracción mutante (M%) > 10% 6.409
(Tasa de recuperación: 5.754/11.903 48%; PPV: 89%)
Nivel B: Fracción mutante (M%) > 5% 6.425
(Tasa de recuperación: 5.756/11.903 48%; PPV: 89%)
Nivel D: Tamaño de filtro: al menos una diferencia de 10 pb 5.230
(Tasa de recuperación: 4.725/11.903 39%; PPV: 90%)

FIG. 27

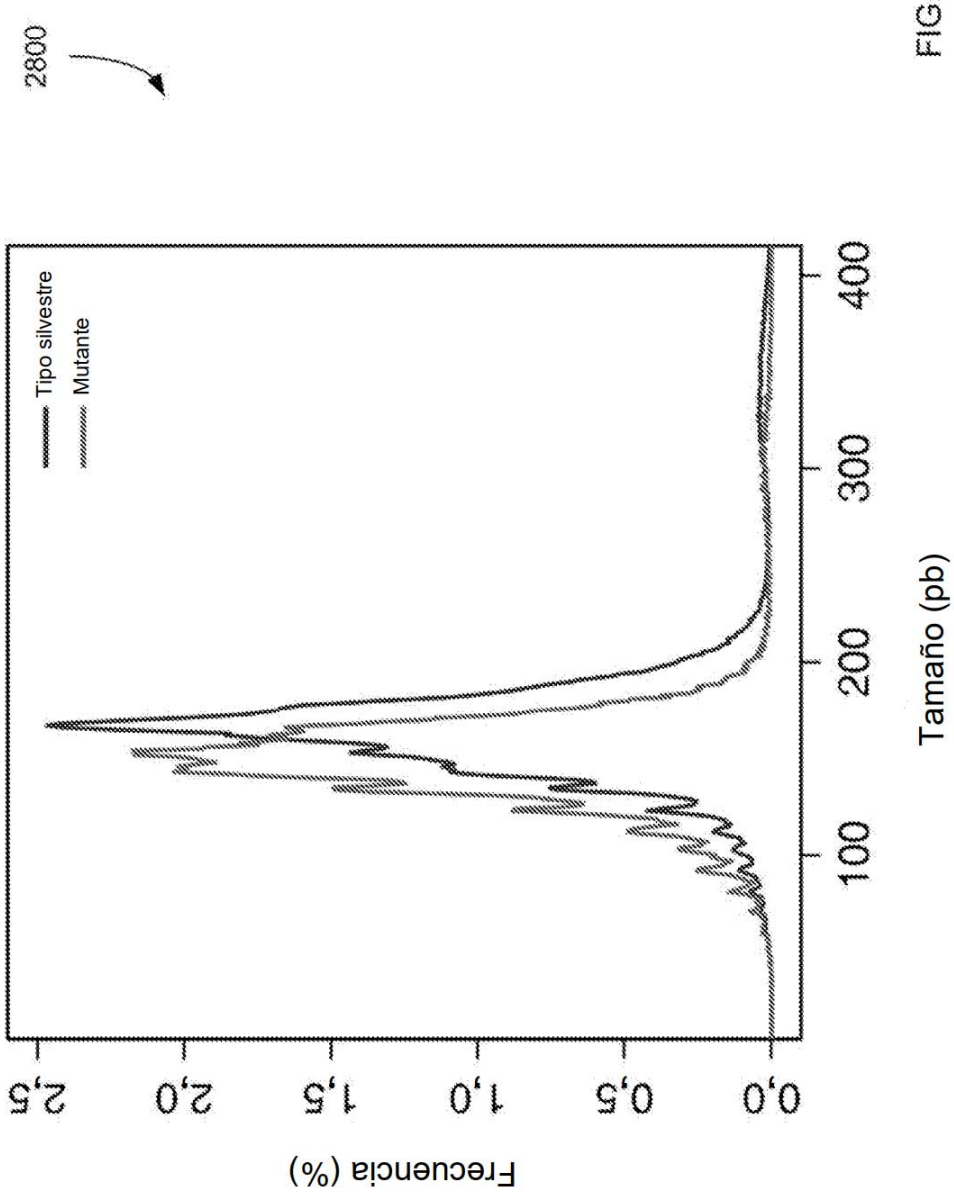

2800
Tipo silvestre
Mutante
0,0
0,5
1,0
1,5
2,0
2,5
Frecuencia (%)
100
200
300
400
Tamaño (pb)

FIG. 28

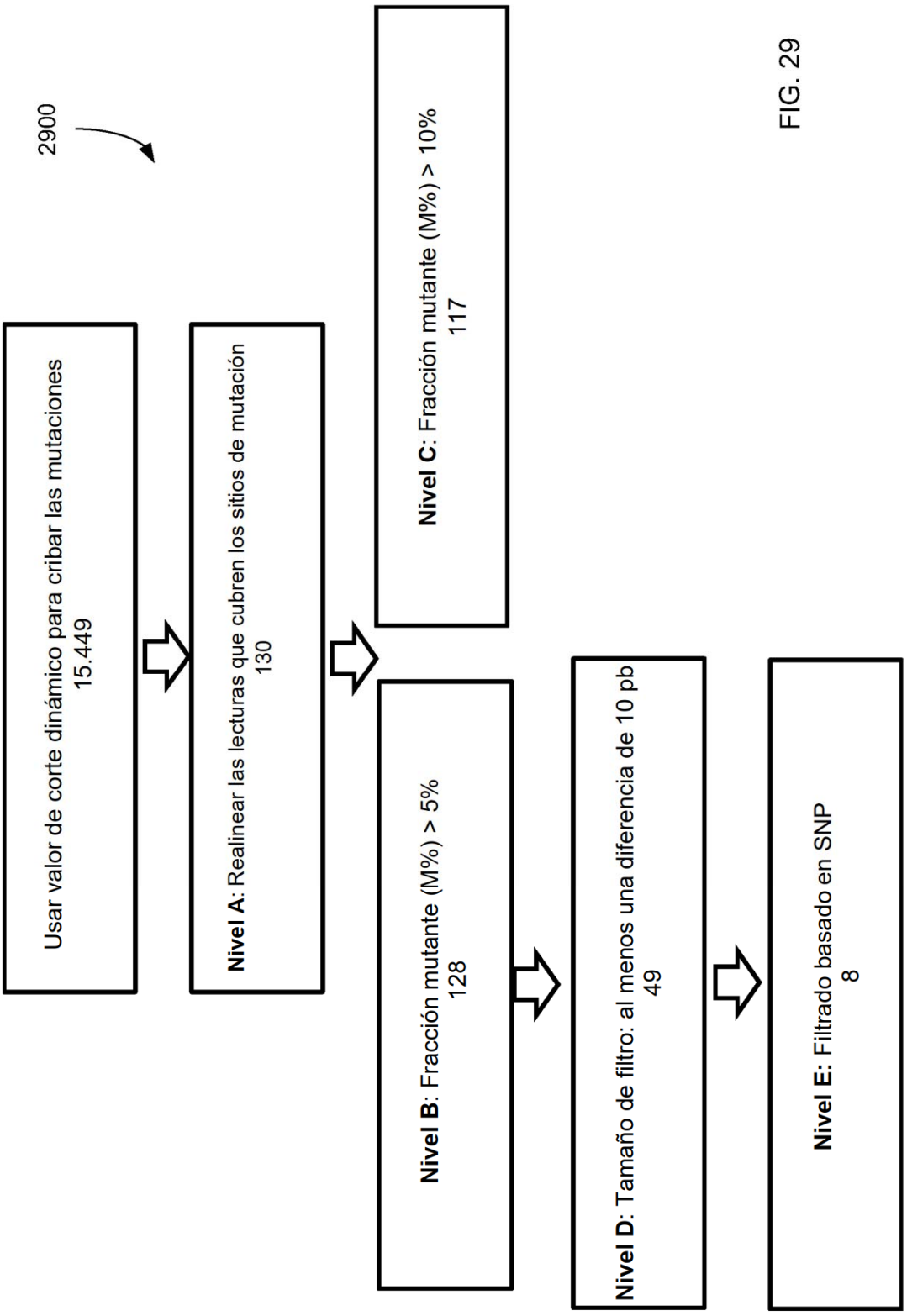

2900
Usar valor de corte dinámico para cribar las mutaciones
15.449
Nivel A: Realinear las lecturas que cubren los sitios de mutación
130
Nivel B: Fracción mutante (M%) > 5%
128
Nivel C: Fracción mutante (M%) > 10%
117
Nivel D: Tamaño de filtro: al menos una diferencia de 10 pb
49
Nivel E: Filtrado basado en SNP
8

FIG. 29

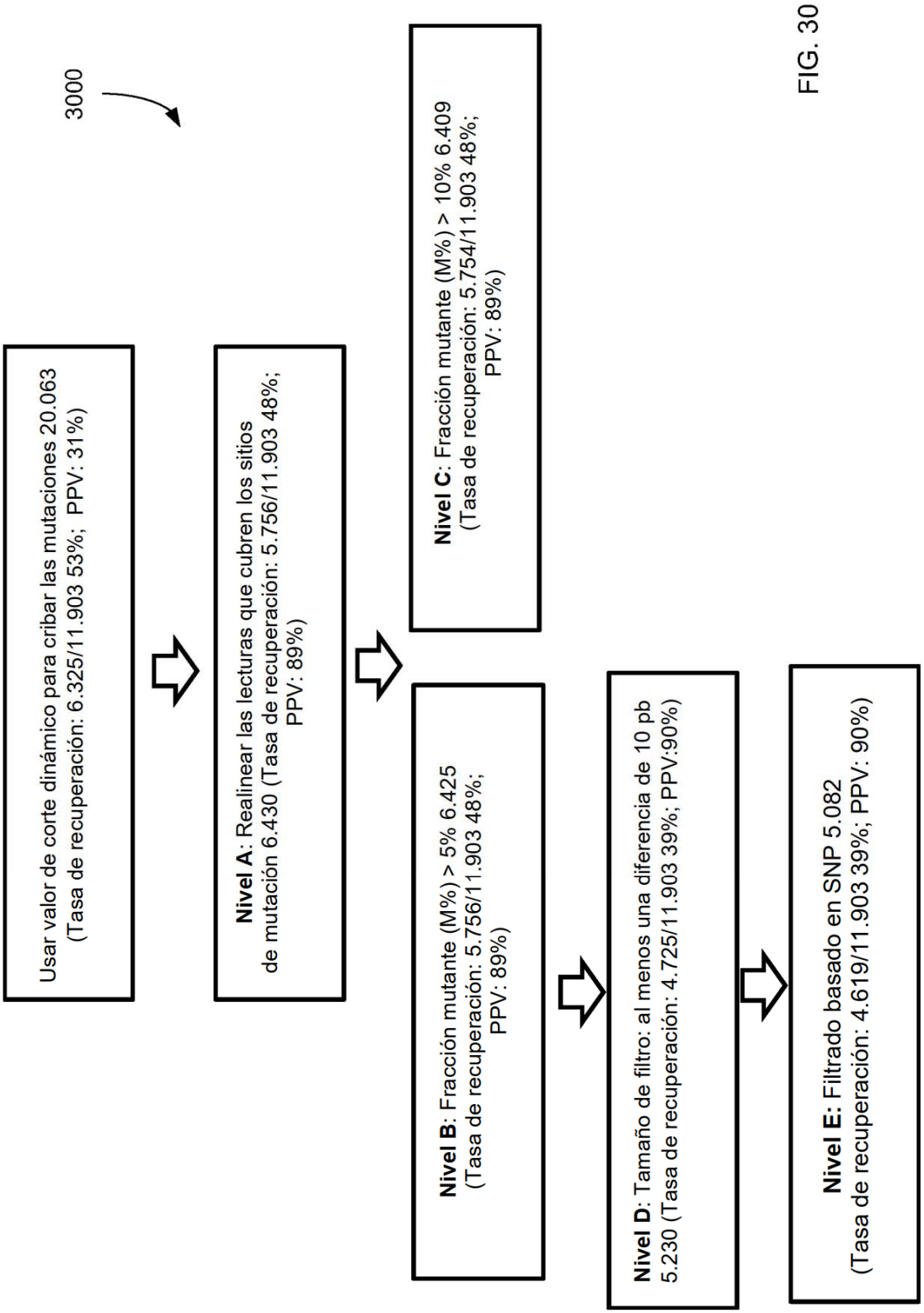

3000
Usar valor de corte dinámico para cribar las mutaciones 20.063 (Tasa de recuperación: 6.325/11.903 53%; PPV: 31%)
Nivel A: Realinear las lecturas que cubren los sitios de mutación 6.430 (Tasa de recuperación: 5.756/11.903 48%; PPV: 89%)
Nivel B: Fracción mutante (M%) > 5% 6.425 (Tasa de recuperación: 5.756/11.903 48%; PPV: 89%)
Nivel C: Fracción mutante (M%) > 10% 6.409 (Tasa de recuperación: 5.754/11.903 48%; PPV: 89%)
Nivel D: Tamaño de filtro: al menos una diferencia de 10 pb 5.230 (Tasa de recuperación: 4.725/11.903 39%; PPV:90%)
Nivel E: Filtrado basado en SNP 5.082 (Tasa de recuperación: 4.619/11.903 39%; PPV: 90%)

FIG. 30

3100

Fuente	ID	Correlación entre mutación en plasma y H3K4me1	Correlación entre mutación en plasma y H3K27ac	Correlación entre mutación en plasma y H3K9me3
Mama	E027	-0,415	NA	0,352
Monocitos	E029	-0,421	-0,38	0,444
Sangre	E030	-0,400	NA	0,482
hígado	E066	**-0,553**	**-0,50**	**0,570**
cerebro	E071	-0,410	-0,42	0,366
colon	E076	-0,401	-0,39	0,134
gástrico	E094	-0,459	-0,48	0,318
pulmón	E096	-0,460	-0,47	0,016
ovario	E097	-0,391	-0,42	0,080
páncreas	E098	-0,463	-0,47	0,301
colon sigmoide	E106	-0,464	-0,44	0,455
intestino delgado	E109	-0,472	-0,42	0,356

FIG. 31

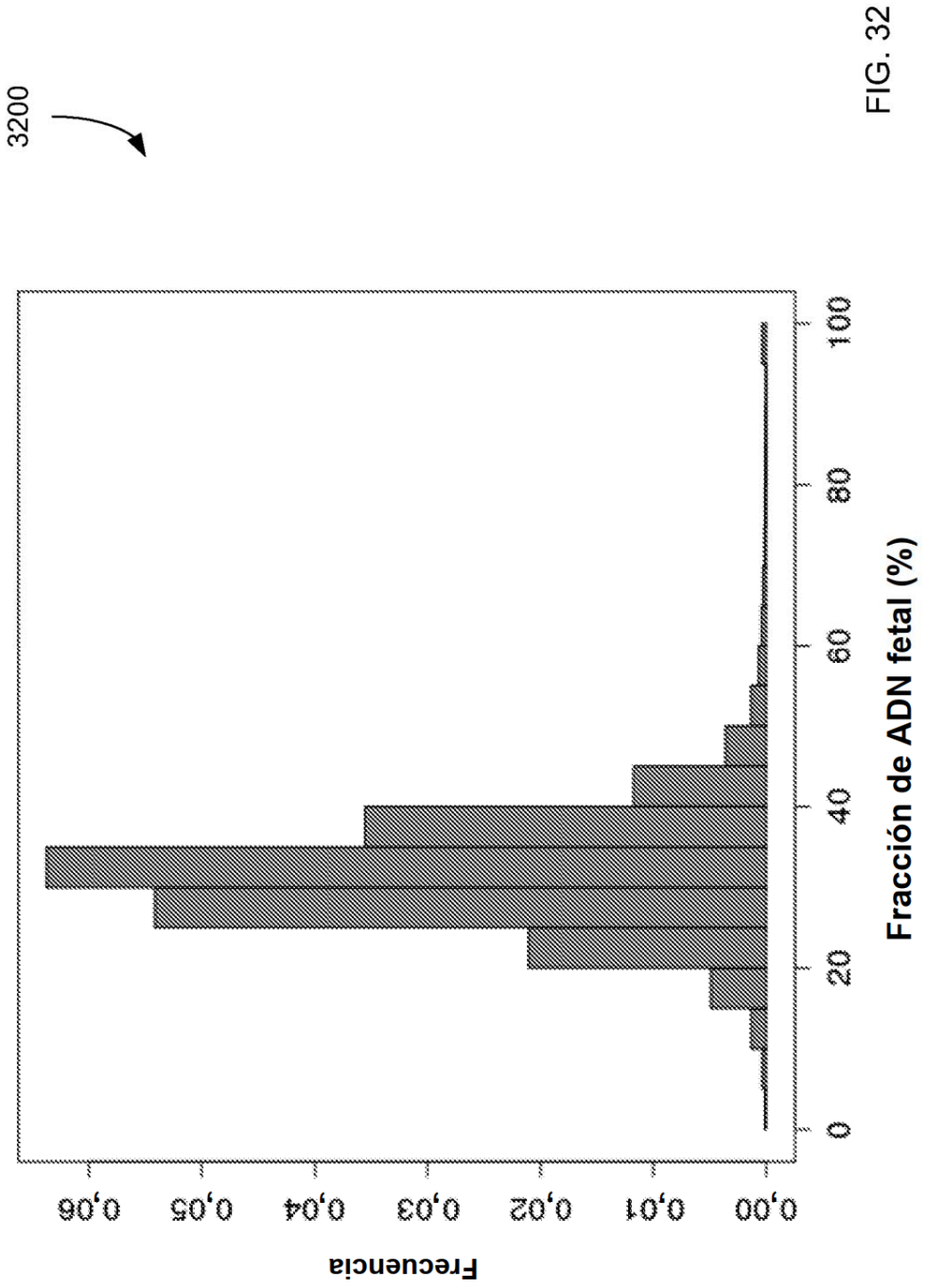

3200
Frecuencia
0,00
0,01
0,02
0,03
0,04
0,05
0,06
0
20
40
60
80
100
Fracción de ADN fetal (%)

FIG. 32

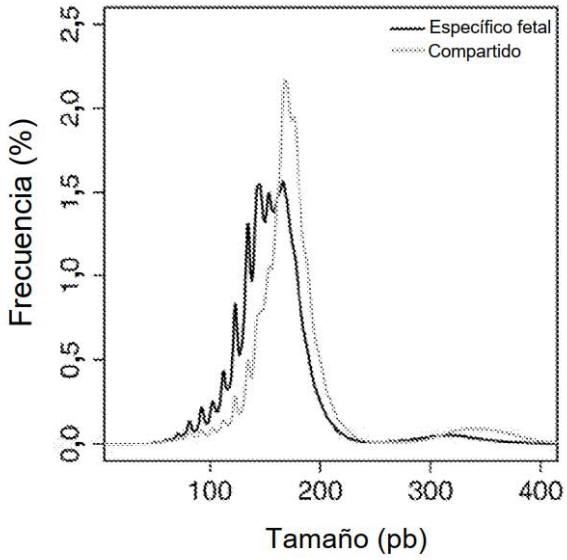


FIG. 33A

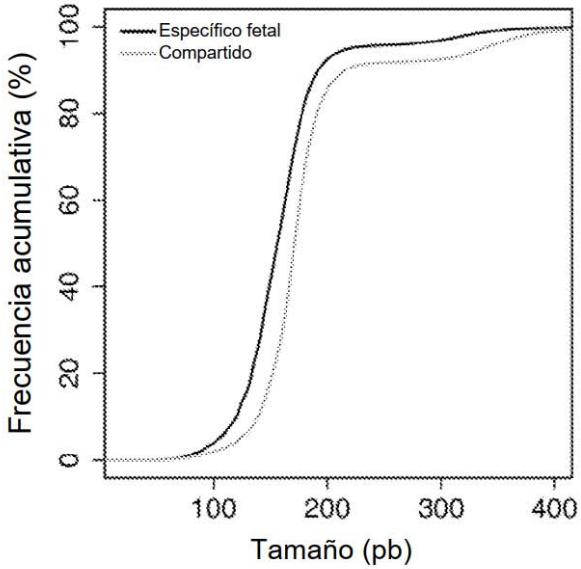


FIG. 33B

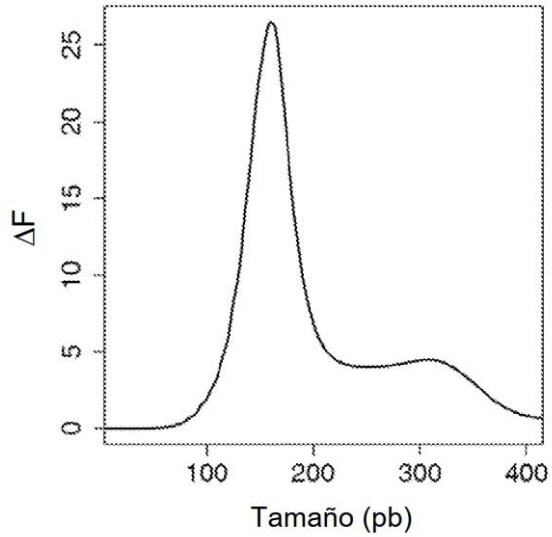


FIG. 33C

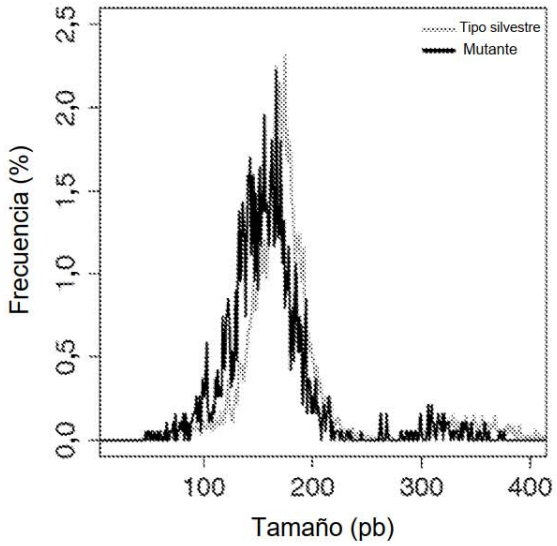


FIG. 34A

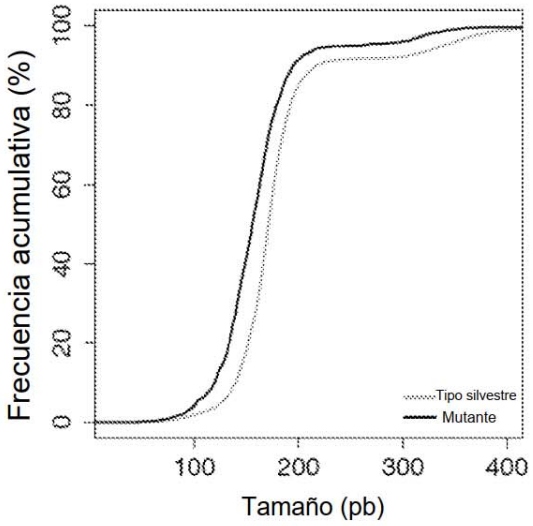


FIG. 34B

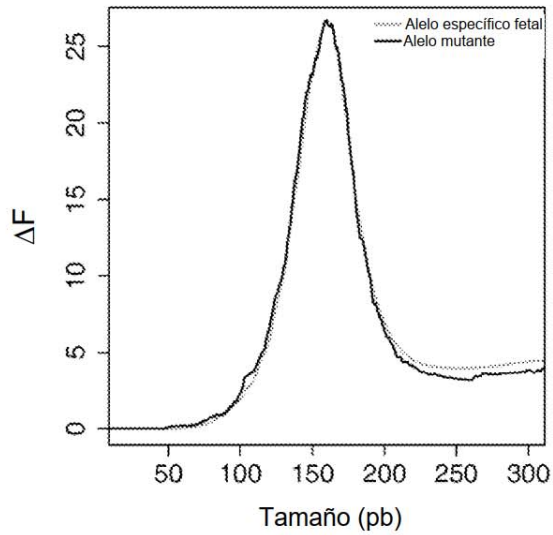


FIG. 34C

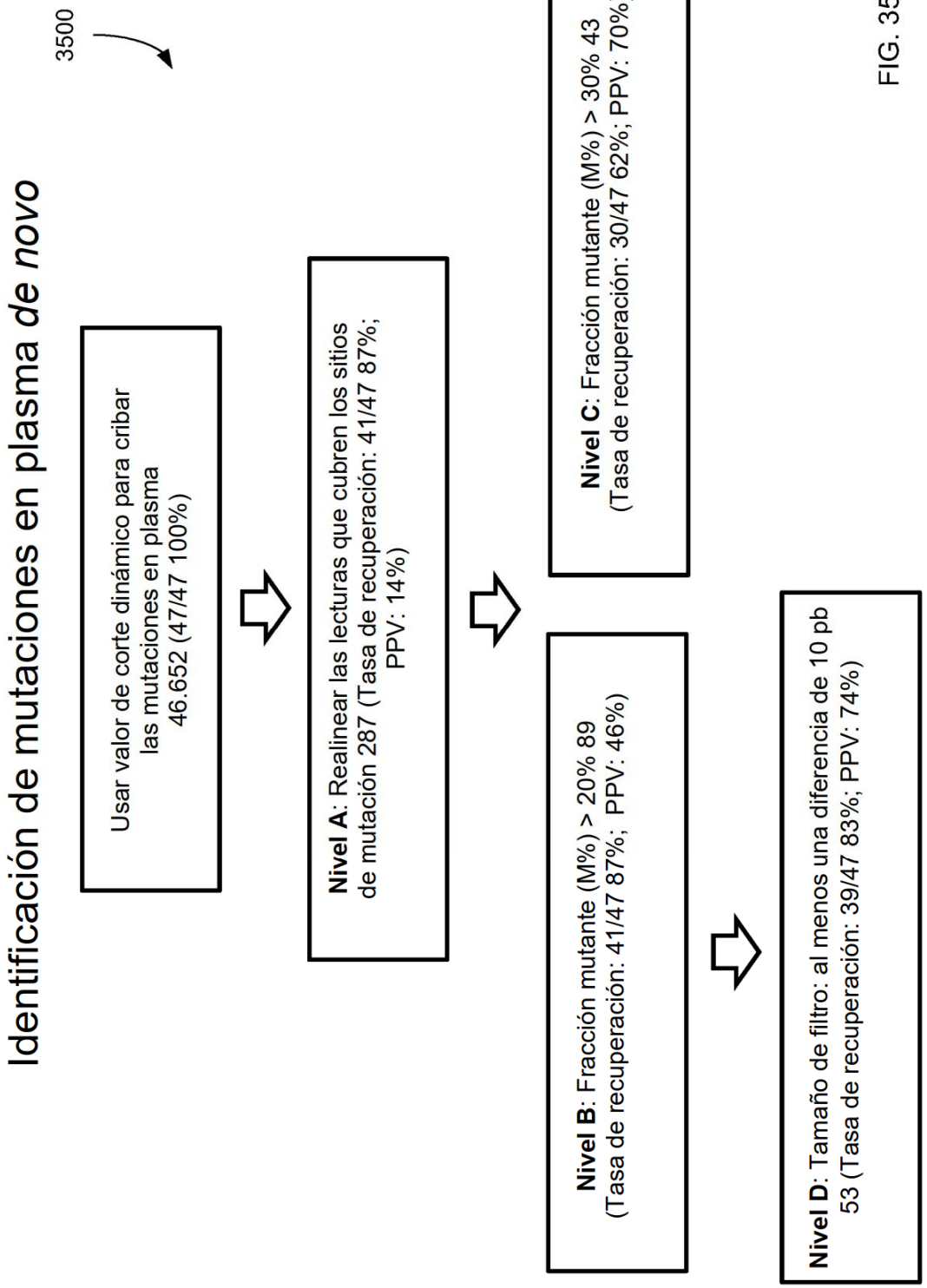

Identificación de mutaciones en plasma *de novo*
3500
Usar valor de corte dinámico para cribar las mutaciones en plasma 46.652 (47/47 100%)
Nivel A: Realinear las lecturas que cubren los sitios de mutación 287 (Tasa de recuperación: 41/47 87%; PPV: 14%)
Nivel B: Fracción mutante (M%) > 20% 89 (Tasa de recuperación: 41/47 87%; PPV: 46%)
Nivel C: Fracción mutante (M%) > 30% 43 (Tasa de recuperación: 30/47 62%; PPV: 70%)
Nivel D: Tamaño de filtro: al menos una diferencia de 10 pb 53 (Tasa de recuperación: 39/47 83%; PPV: 74%)

FIG. 35

FIG. 36B

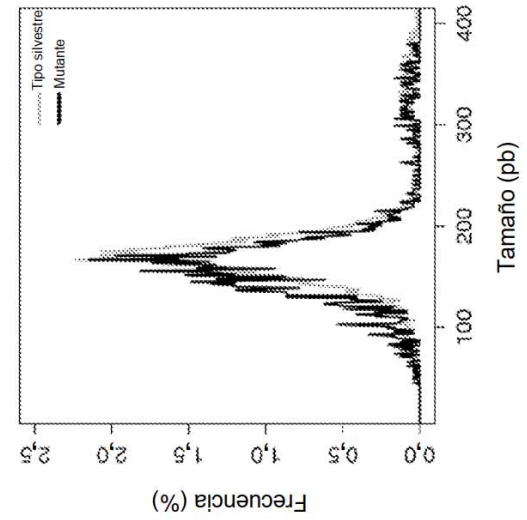


FIG. 36D

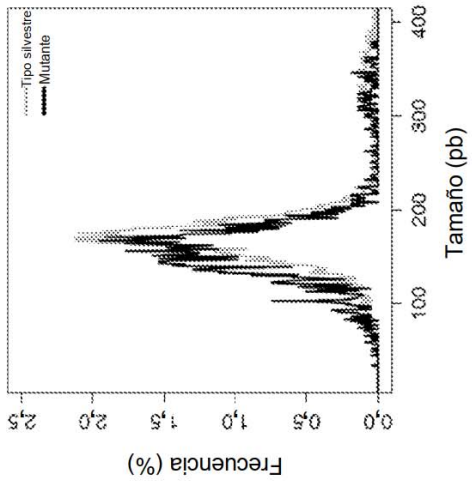


FIG. 36A

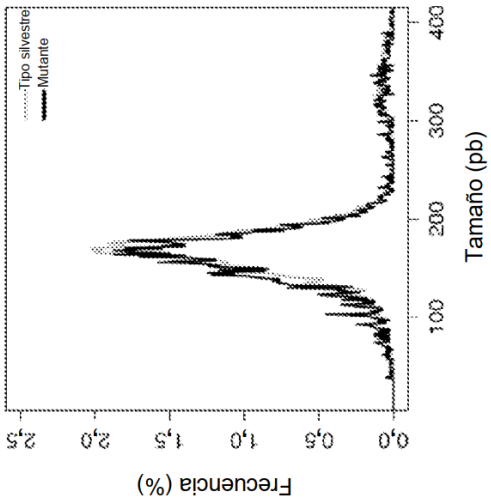
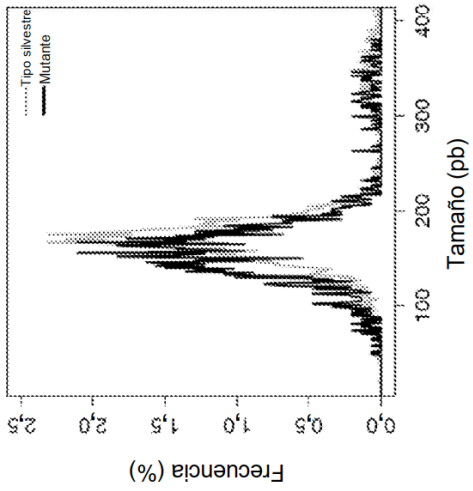


FIG. 36C



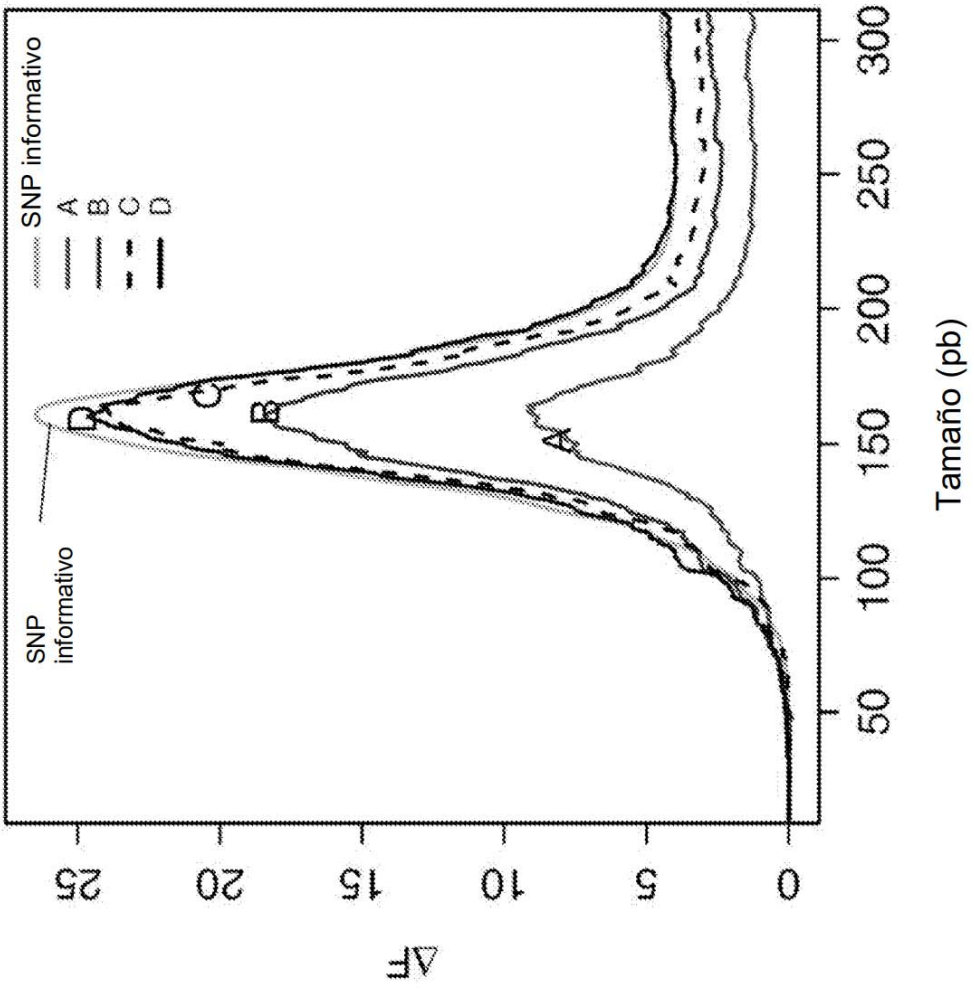

SNP informativo
A
B
C
D
SNP informativo
ΔF
0
5
10
15
20
25
50
100
150
200
250
300
Tamaño (pb)

FIG. 37

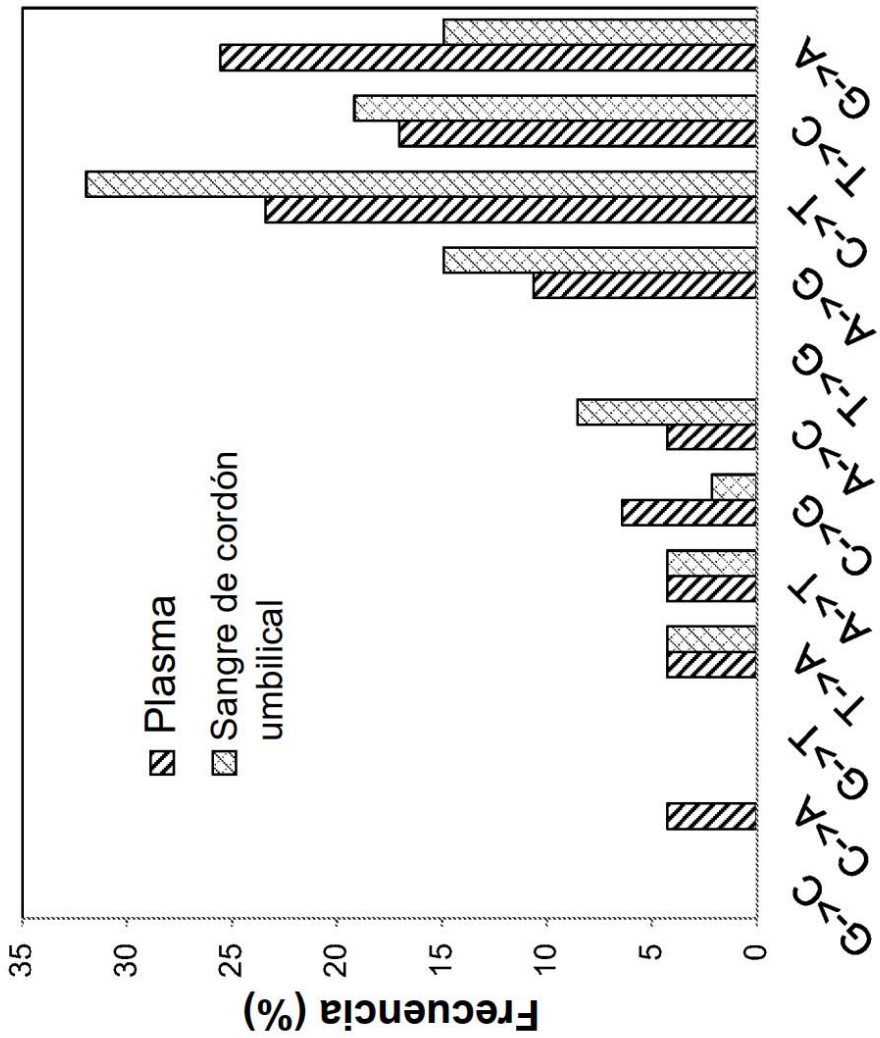

Frecuencia (%)
0
5
10
15
20
25
30
35
Plasma
Sangre de cordón umbilical
G-7-C
C-7-G
G-7-T
A-7-T
T-7-A
C-7-G
A-7-T
G-7-C
T-7-C
C-7-T
A-7-G
Tipos de mutaciones

FIG. 38

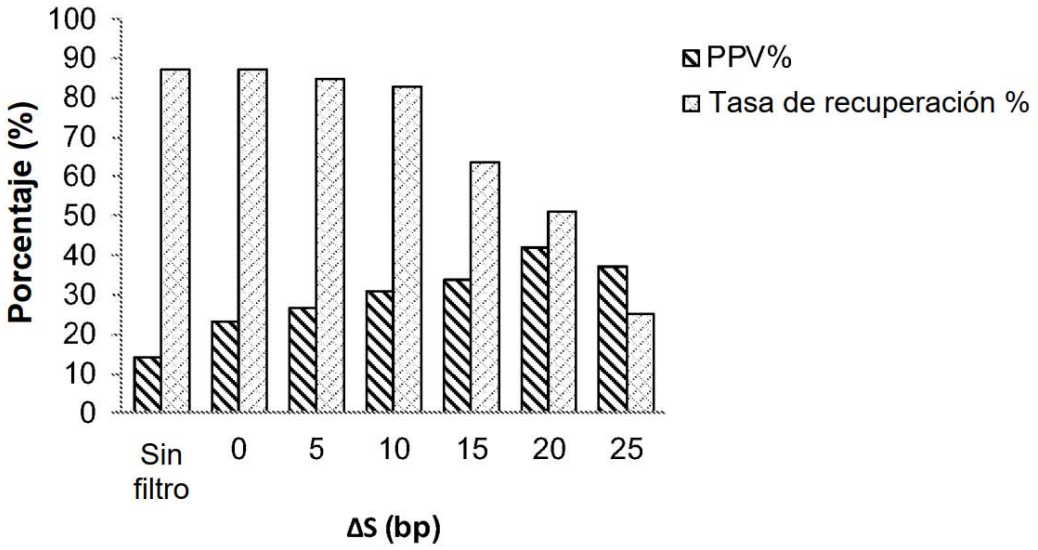


FIG. 39A

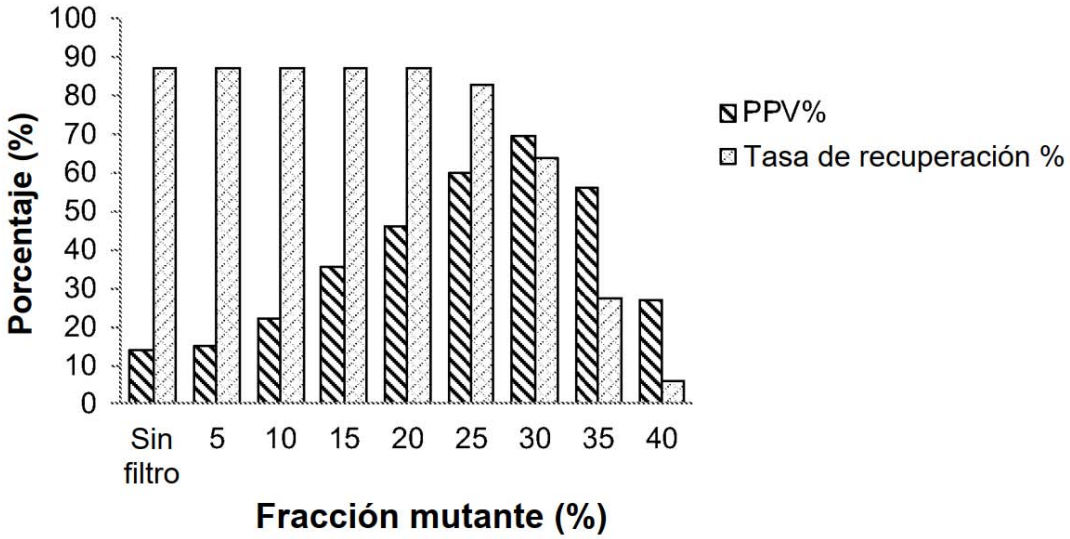


FIG. 39B

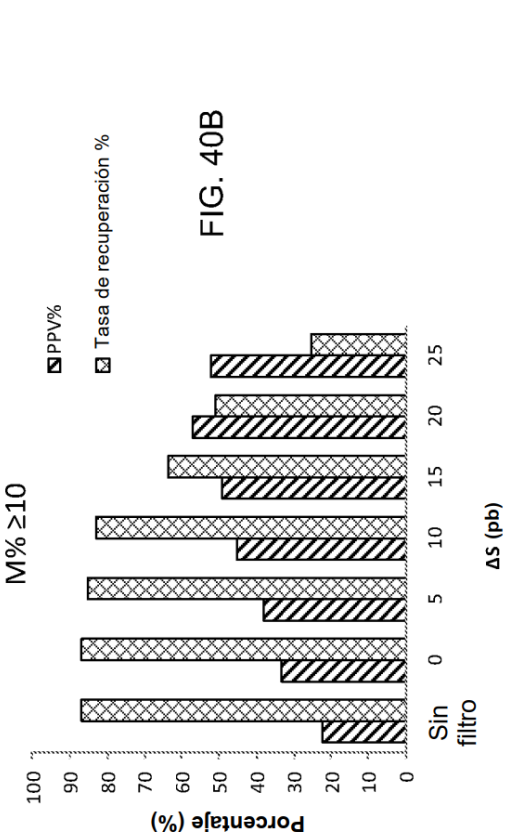


FIG. 40B

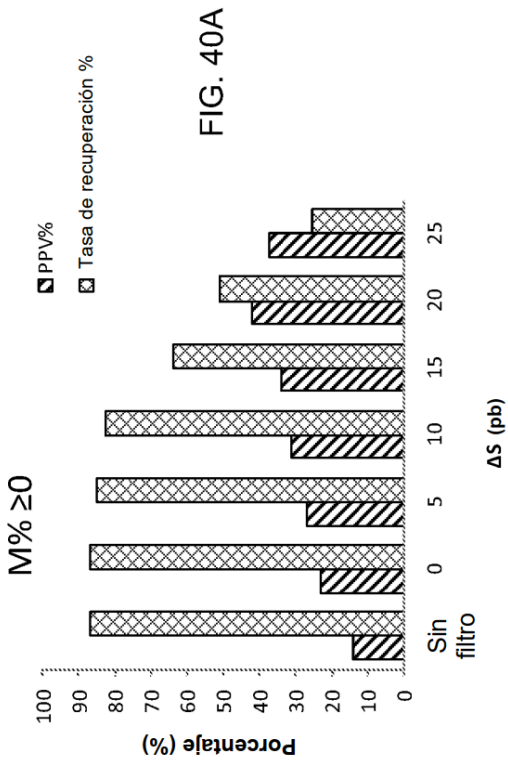


FIG. 40A

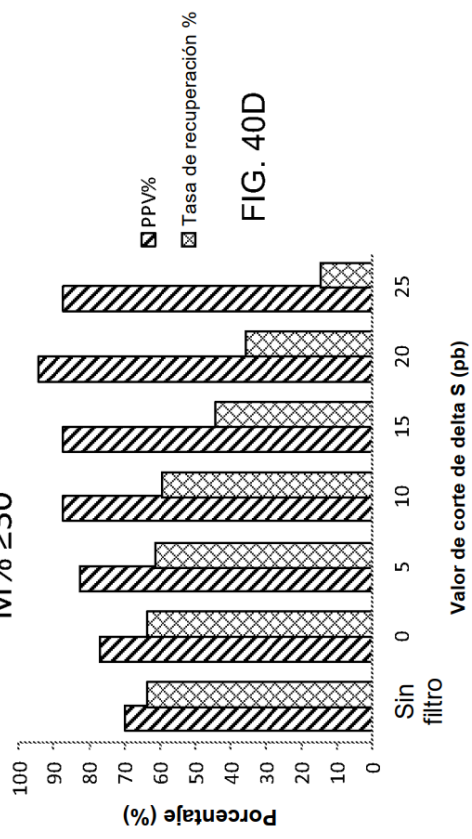


FIG. 40D

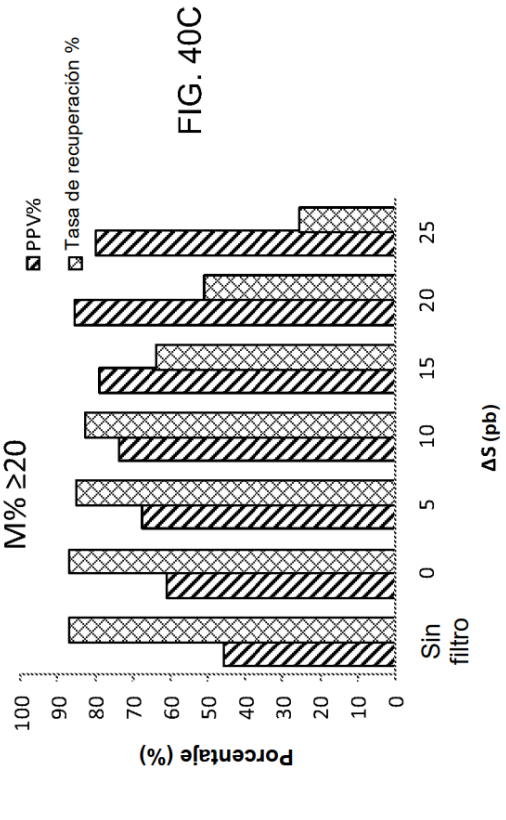


FIG. 40C

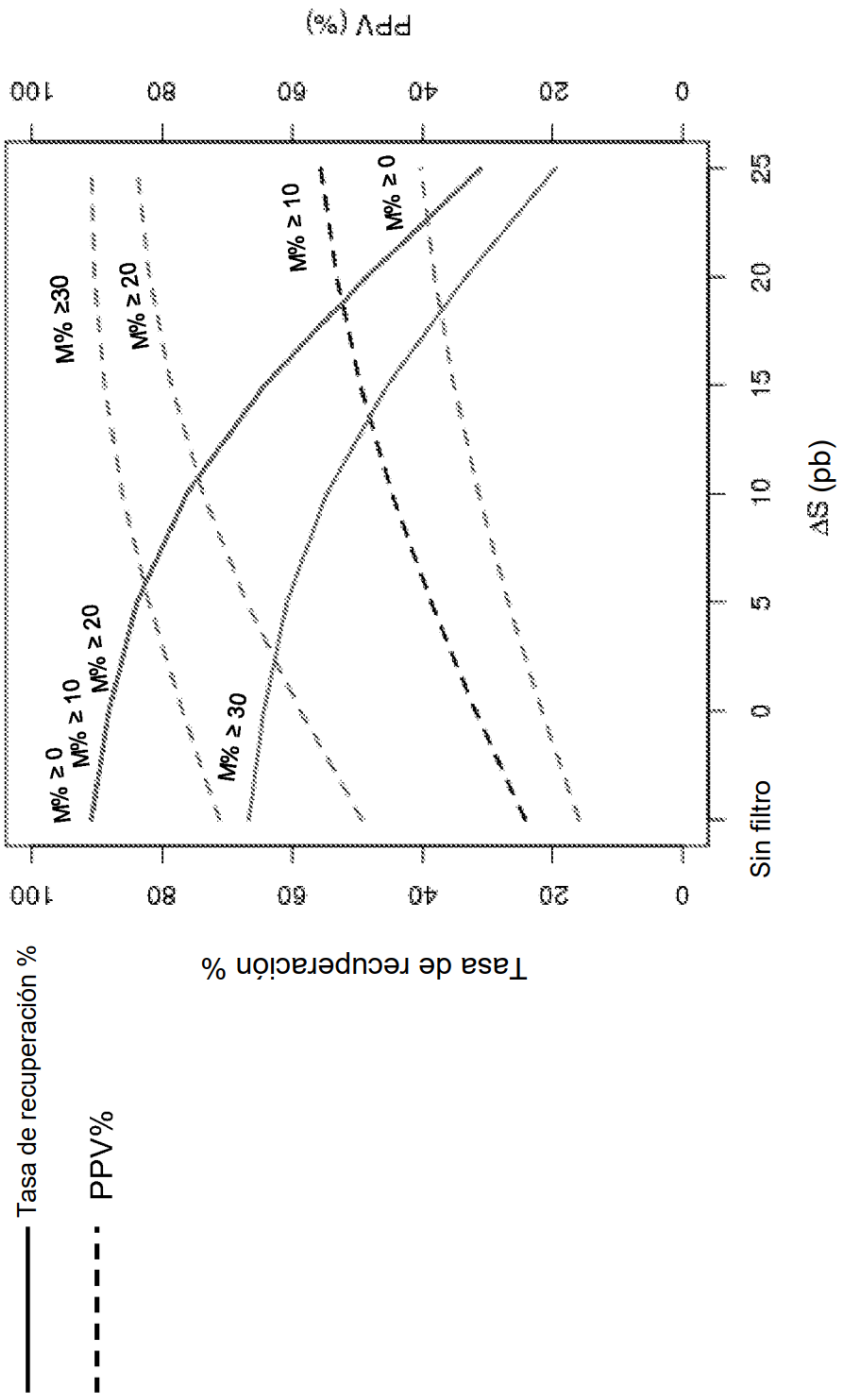

PPV (%)
100
80
60
40
20
0
M% ≥30
M% ≥ 20
M% ≥ 10
M% ≥ 0
M% ≥ 10
M% ≥ 20
M% ≥ 30
M% ≥ 0
Sin filtro
0
5
10
15
20
25
ΔS (pb)
Tasa de recuperación %
Tasa de recuperación %
PPV%

FIG. 41

Nombre	Posición en el cr.	Plasma materno (NGS)		Padre		Madre		Sangre del cordón umbilical		Placenta		Observación
		Genotipo	Recuento	NGS	Sanger seq.	NGS	Sanger seq.	NGS	Sanger seq.	NGS	Sanger seq.	
TP1	cr1:86987866	CT	259:44	CC	CC	CC	CC	CT	CT	CT	CT	
TP2	cr2:231414780	CT	239:53	CC	CC	CC	CC	CT	CT	CT	CT	
TP3	cr3:126060289	GA	246:50	GG	GG	GG	GG	GA	GA	GA	GA	
TP4	cr4:73534843	AC	204:45	AA	AA	AA	AA	AC	AC	AC	AC	
TP5	cr4:75070975	CC	294:0	CC	CC	CC	CC	CT	CT	CC	CC	Específica de sangre de cordón umbilical
TP6	cr4:95295449	CA	261:35	CC	CC	CC	CC	CC	CC	CA	CA	Específica de placenta
TP7	cr5:83895390	GA	253:42	GG	GG	GG	GG	GA	GA	GA	GA	
TP8	cr6:54202751	CA	232:38	CC	CC	CC	CC	CA	CA	CA	CA	
TP9	cr7:41399864	TA	282:59	TT	TT	TT	TT	TA	TA	TA	TA	
TP10	cr8:117728725	CT	235:43	CC	CC	CC	CC	CT	CT	CT	CT	
TP11	cr10:22255065	AG	207:48	AA	AA	AA	AA	AG	AG	AG	AG	
TP12	cr11:76736964	TC	234:58	TT	TT	TT	TT	TC	TC	TC	TC	
TP13	cr12:106632793	GA	230:44	GG	GG	GG	GG	GA	GA	GA	GA	
TP14	cr14:57083665	TC	205:52	TT	TT	TT	TT	TC	TC	TC	TC	
TP15	cr17:66731200	CT	241:44	CC	CC	CC	CC	CT	CT	CT	CT	
TP16	cr19:33302746	GA	248:32	GG	GG	GG	GG	GG	GG	GA	GA	Específica de placenta
TP17	cr20:41416859	CT	165:49	CC	CC	CC	CC	CT	CT	CT	CT	
TP18	cr21:26925757	CT	220:47	CC	CC	CC	CC	CT	CT	CT	CT	

FIG. 42

ES 2 908 347 T3

Nombre	Posición en el cr.	Plasma materno (NGS)		Padre		Madre		Sangre del cordón umbilical		Placenta		Observación
		Genotipo	Recuento	NGS	Sanger seq.	NGS	Sanger seq.	NGS	Sanger seq.	NGS	Sanger seq.	
TP19	cr1:2816785	GA	216:38	GG	GG	GG	GG	AG	AG	GA	GA	
TP20	cr2:60295478	CT	310:60	CC	CC	CC	CC	CT	CT	CT	CT	
TP21	cr3:75762726	AG	331:39	AG (58:1)	AG	AG (62:1)	AG	AG (66:3)	AG	AG (27:1)	AG	Todas heterocigotas
TP22	cr3:77051026	CT	264:53	CC	CC	CC	CC	CT	CT	CT	CT	
TP23	cr4:34755555	CG	262:49	CC	CC	CC	CC	CG	CG	CG	CG	
TP24	cr4:54613139	CT	168:28	CC	CC	CC	CC	CT	CT	CT	CT	
TP25	cr4:68332287	AG	182:41	AA	AA	AA	AA	AG	AG	AG	AG	
TP26	cr4:113248266	AT	225:51	AA	AA	AA	AA	AT	AT	AT	AT	
TP27	cr4:181441802	TA	206:41	TT	TT	TT	TT	TA	TA	TA	TA	
TP28	cr4: 2366786	CT	233:43	CC	CC	CC	CC	CT	CT	CT	CT	
TP29	cr5:146803590	CG	270:44	CC	CC	CC	CC	CG	CG	CG	CG	
TP30	cr6:32603242	GA	133:16	GG (36)	GG	GG (24)	GG	GG (36)	GG	GA (21:3)	GG	Específica de placenta: región HLA
TP31	cr6: 55920652	TC	271:54	TT	TT	TT	TT	TC	TC	TC	TC	
TP32	cr6:106112398	GA	277:57	GG	GG	GG	GG	GA	GA	GA	GA	
TP33	cr6:169649375	TC	182:32	TT	TT	TT	TT	TC	TC	TC	TC	
TP34	cr6:168937294	AG	224:47	AA	AA	AA	AA	AG	AG	AG	AG	
TP35	cr8: 17334059	CG	281:36	CC	CC	CC	CC	CG	CG	CG	CG	
TP36	cr8:107961680	AC	309:49	AA	AA	AA	AA	AC	AC	AC	AC	
TP37	cr9: 36612823	TC	290:35	TT	TT	TT	TT	TC	TC	TC	TC	
TP38	cr10: 50117829	AT	247:65	AA	AA	AA	AA	AT	AT	AT	AT	
TP39	cr10:121730789	GA	204:41	GG	GG	GG	GG	GA	GA	GA	GA	
TP40	cr12: 57200848	TC	270:67	TT	TT	TT	TT	TC	TC	TC	TC	
TP41	cr12:130836521	CT	159:31	CC	CC	CC	CC	CT	CT	CT	CT	
TP42	cr14: 33427740	GA	240:44	GG	GG	GG	GG	GA	GA	GA	GA	
TP43	cr15: 27113786	TC	239:43	TT	TT	TT	TT	TC	TC	TC	TC	
TP44	cr18: 82769063	GA	132:18	GG (45)	GG	GG (22)	GA	GG (37)	GA	GA (11:1)	GA	Específica de placenta: resultados de NGS y sec. de Sanger no consistentes
TP45	cr17: 8434076	TC	276:47	TT	PCR fallida	TT	PCR fallida	TC	PCR fallida	TC	PCR fallida	Volver a optimizar PCR
TP46	cr17: 78880681	GA	223:39	GG	GG	GG	GG	GA	GA	GA	GA	
TP47	cr16: 10102981	GA	156:23	GG	GG	GG	GG	GA	GA	GA	GA	
TP48	cr20: 44745129	AG	191:34	AA	AA	AA	AA	GA	GA	GA	GA	

FIG. 43

ES 2 908 347 T3

Análisis de muestreo descendente

Mutaciones simuladas: 3000

Mutaciones *de novo:* 47

Fracción de ADN fetal (%): 32%

Prof de sec. (X)	Pasa el valor de corte dinámico	Pasa la realineación	Pasa el filtro de tamaño	Recuperadas de 47 mutaciones	Falsos positivos	Recuperadas de 3000 mutaciones simuladas	PPV%	Tasa de recuperación (%)
25	834	482	476	10	4	462	99,16	15,49
50	3496	1522	1499	28	10	1461	99,33	48,87
100	10132	2072	2021	35	32	1954	98,42	65,28

FIG. 44

El n.º de mutaciones = 3000										
Profundidad (X)	El número de mutaciones verdaderas detectadas en diferentes fracciones de ADN tumoral (%)									FP
	1%	2%	3%	4%	5%	10%	15%	20%	30%	
25	0	0	1	4	10	102	340	709	1587	4
50	0	0	3	10	24	311	961	1706	2664	10
60	0	1	6	21	51	541	1415	2187	2873	12
80	0	4	22	67	151	1113	2176	2736	2986	18
90	0	7	35	104	227	1411	2441	2860	2996	25
100	0	10	53	152	319	1692	2633	2929	2999	32
200	11	155	550	1113	1685	2921	2998	3000	3000	45
400	49	642	1667	2435	2806	3000	3000	3000	3000	na
600	250	1665	2658	2942	2992	3000	3000	3000	3000	na
800	643	2430	2941	2996	3000	3000	3000	3000	3000	na

FIG. 45A

El n.º de mutaciones = 6000										
Profundidad (X)	El número de mutaciones verdaderas detectadas en diferentes fracciones de ADN tumoral (%)									FP
	1%	2%	3%	4%	5%	10%	15%	20%	30%	
25	0	1	3	9	19	205	681	1418	3173	4
50	0	1	6	19	49	622	1923	3413	5327	10
60	0	2	13	42	103	1082	2829	4374	5746	12
80	0	8	43	134	302	2227	4352	5472	5972	18
90	1	13	70	209	453	2822	4882	5721	5991	25
100	1	21	106	305	638	3384	5266	5858	5997	32
200	21	310	1101	2227	3370	5841	5996	6000	6000	45
400	98	1284	3333	4869	5612	6000	6000	6000	6000	na
600	500	3331	5317	5883	5985	6000	6000	6000	6000	na
800	1287	4861	5882	5992	6000	6000	6000	6000	6000	na

FIG. 45B

El n.º de mutaciones = 9000										
Profundidad (X)	El número de mutaciones verdaderas detectadas en diferentes fracciones de ADN tumoral (%)									FP
	1%	2%	3%	4%	5%	10%	15%	20%	30%	
25	0	1	4	13	29	307	1021	2128	4760	4
50	0	1	8	29	73	933	2884	5119	7991	10
60	0	3	19	63	154	1623	4244	6561	8619	12
80	0	12	65	201	453	3340	6528	8208	8957	18
90	1	20	105	313	680	4233	7324	8581	8987	25
100	1	31	159	457	956	5076	7899	8787	8996	32
200	32	466	1651	3340	5056	8762	8994	9000	9000	45
400	147	1927	5000	7304	8418	9000	9000	9000	9000	na
600	751	4996	7975	8825	8977	9000	9000	9000	9000	na
800	1930	7291	8823	8988	8999	9000	9000	9000	9000	na

FIG. 45C

ES 2 908 347 T3

El n.º de mutaciones = 10000

Profundidad (X)	El número de mutaciones verdaderas detectadas en diferentes fracciones de ADN tumoral (%)									FP
	1%	2%	3%	4%	5%	10%	15%	20%	30%	
25	0	1	5	14	32	341	1135	2364	5289	4
50	0	1	9	32	81	1036	3204	5688	8879	10
60	0	3	21	70	171	1803	4716	7290	9576	12
80	1	13	72	224	504	3711	7254	9120	9953	18
90	1	22	117	348	755	4703	8137	9535	9985	25
100	2	34	177	508	1063	5640	8777	9763	9996	32
200	35	517	1835	3712	5617	9736	9994	10000	10000	45
400	163	2141	5555	8115	9353	9999	10000	10000	10000	na
600	834	5551	8861	9805	9974	10000	10000	10000	10000	na
800	2145	8101	9803	9987	9999	10000	10000	10000	10000	na

FIG. 46A

El n.º de mutaciones = 20000

Profundidad (X)	El número de mutaciones verdaderas detectadas en diferentes fracciones de ADN tumoral (%)									FP
	1%	2%	3%	4%	5%	10%	15%	20%	30%	
25	0	2	10	29	65	682	2269	4728	10578	4
50	0	3	18	64	163	2072	6408	11376	17758	10
60	0	7	42	140	342	3607	9432	14581	19153	12
80	1	26	144	447	1007	7422	14508	18241	19905	18
90	2	43	233	696	1510	9406	16275	19069	19971	25
100	3	69	354	1017	2125	11280	17554	19526	19991	32
200	71	1035	3669	7423	11235	19471	19988	20000	20000	45
400	326	4282	11111	16231	18706	19999	20000	20000	20000	na
600	1668	11103	17723	19611	19949	20000	20000	20000	20000	na
800	4290	16203	19606	19974	19999	20000	20000	20000	20000	na

FIG. 46B

El n.º de mutaciones = 30000

Profundidad (X)	El número de mutaciones verdaderas detectadas en diferentes fracciones de ADN tumoral (%)									FP
	1%	2%	3%	4%	5%	10%	15%	20%	30%	
25	0	3	15	43	97	1023	3404	7092	15866	4
50	0	4	28	96	244	3109	9613	17064	26637	10
60	0	10	63	211	513	5410	14147	21871	28729	12
80	2	39	216	671	1511	11134	21761	27361	29858	18
90	3	65	350	1044	2265	14110	24412	28604	29956	25
100	5	103	531	1525	3188	16921	26331	29289	29987	32
200	106	1552	5504	11135	16852	29207	29982	30000	30000	45
400	489	6422	16666	24346	28058	29998	30000	30000	30000	na
600	2502	16654	26584	29416	29923	30000	30000	30000	30000	na
800	6434	24304	29410	29962	29998	30000	30000	30000	30000	na

FIG. 46C

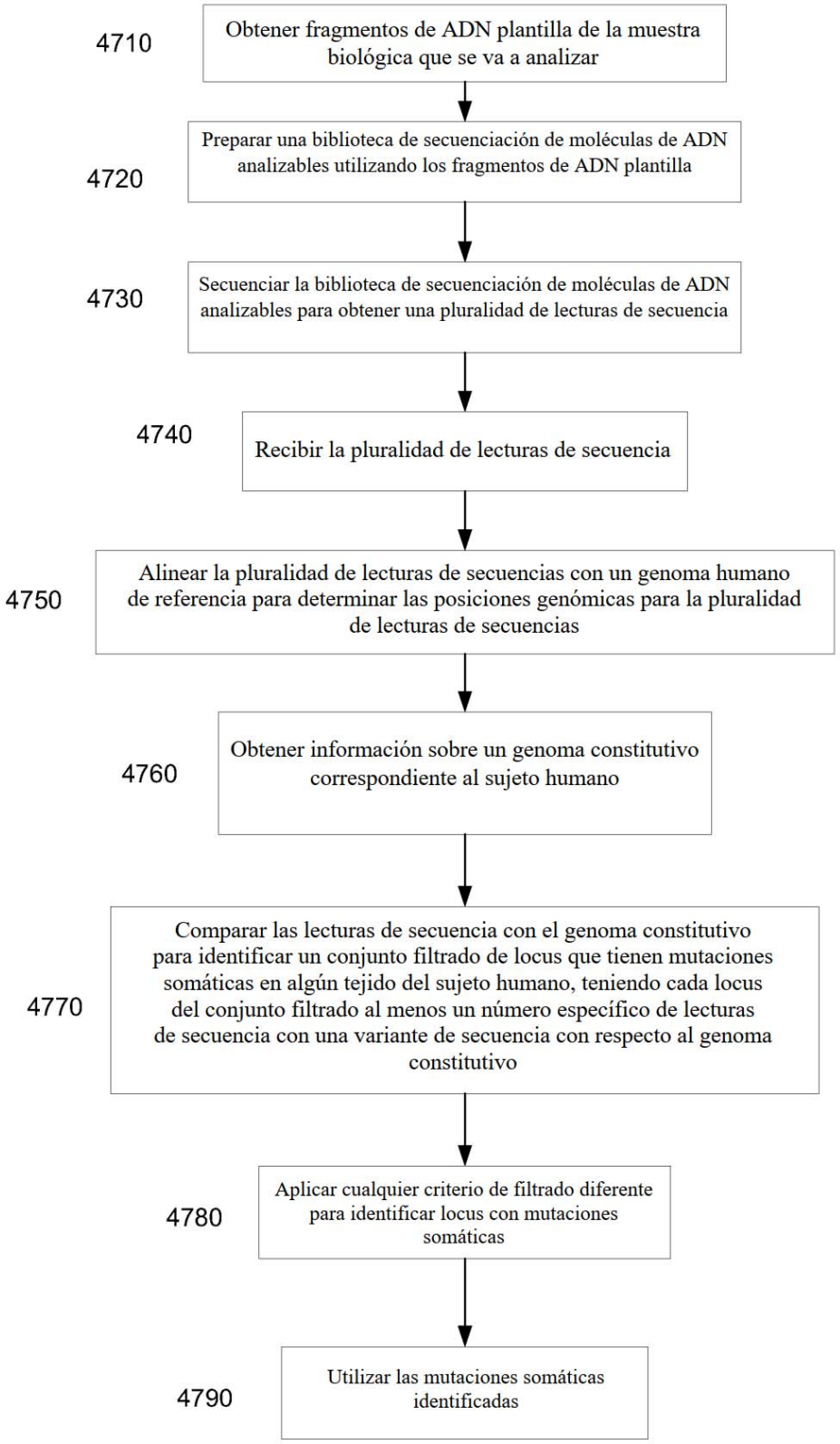

4700
4710
Obtener fragmentos de ADN plantilla de la muestra biológica que se va a analizar
4720
Preparar una biblioteca de secuenciación de moléculas de ADN analizables utilizando los fragmentos de ADN plantilla
4730
Secuenciar la biblioteca de secuenciación de moléculas de ADN analizables para obtener una pluralidad de lecturas de secuencia
4740
Recibir la pluralidad de lecturas de secuencia
4750
Alinear la pluralidad de lecturas de secuencias con un genoma humano de referencia para determinar las posiciones genómicas para la pluralidad de lecturas de secuencias
4760
Obtener información sobre un genoma constitutivo correspondiente al sujeto humano
4770
Comparar las lecturas de secuencia con el genoma constitutivo para identificar un conjunto filtrado de locus que tienen mutaciones somáticas en algún tejido del sujeto humano, teniendo cada locus del conjunto filtrado al menos un número específico de lecturas de secuencia con una variante de secuencia con respecto al genoma constitutivo
4780
Aplicar cualquier criterio de filtrado diferente para identificar locus con mutaciones somáticas
4790
Utilizar las mutaciones somáticas identificadas

FIG. 47

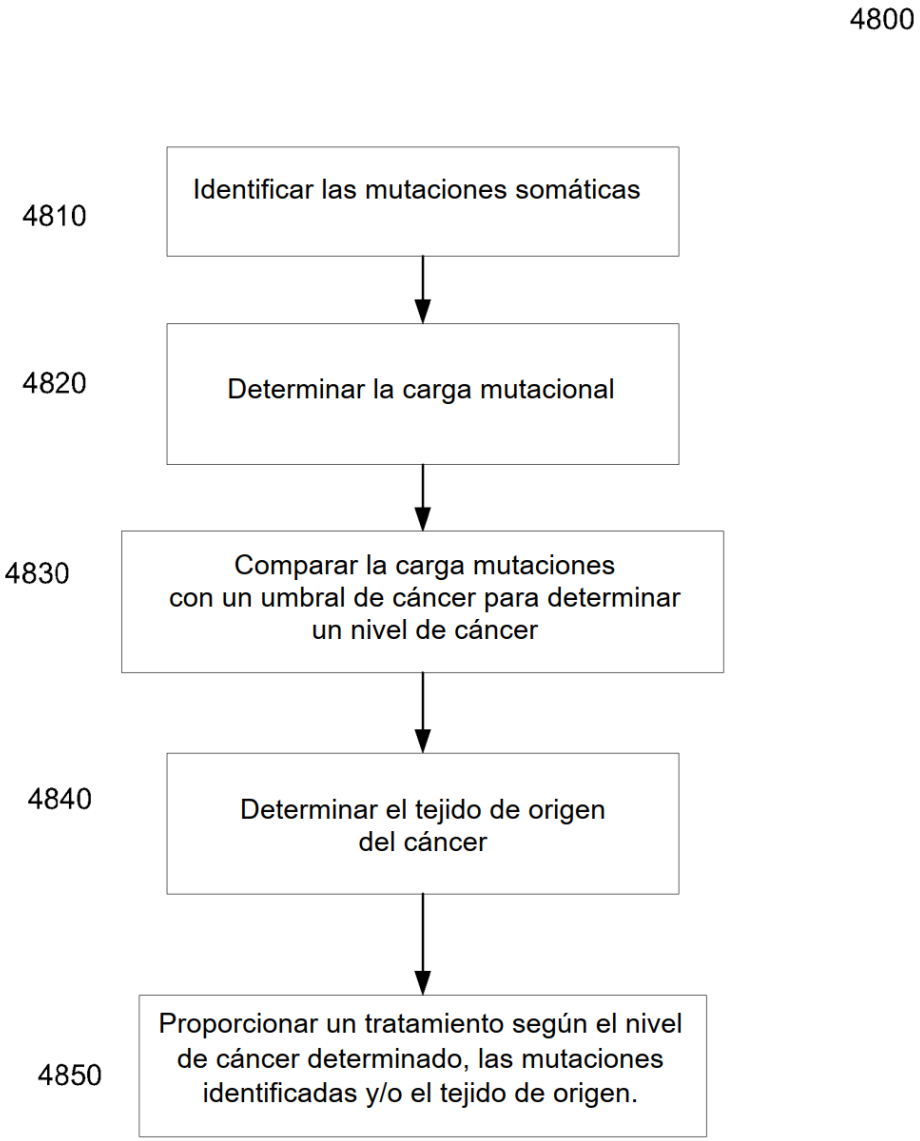

4800
4810
Identificar las mutaciones somáticas
4820
Determinar la carga mutacional
4830
Comparar la carga mutaciones con un umbral de cáncer para determinar un nivel de cáncer
4840
Determinar el tejido de origen del cáncer
4850
Proporcionar un tratamiento según el nivel de cáncer determinado, las mutaciones identificadas y/o el tejido de origen.

FIG. 48

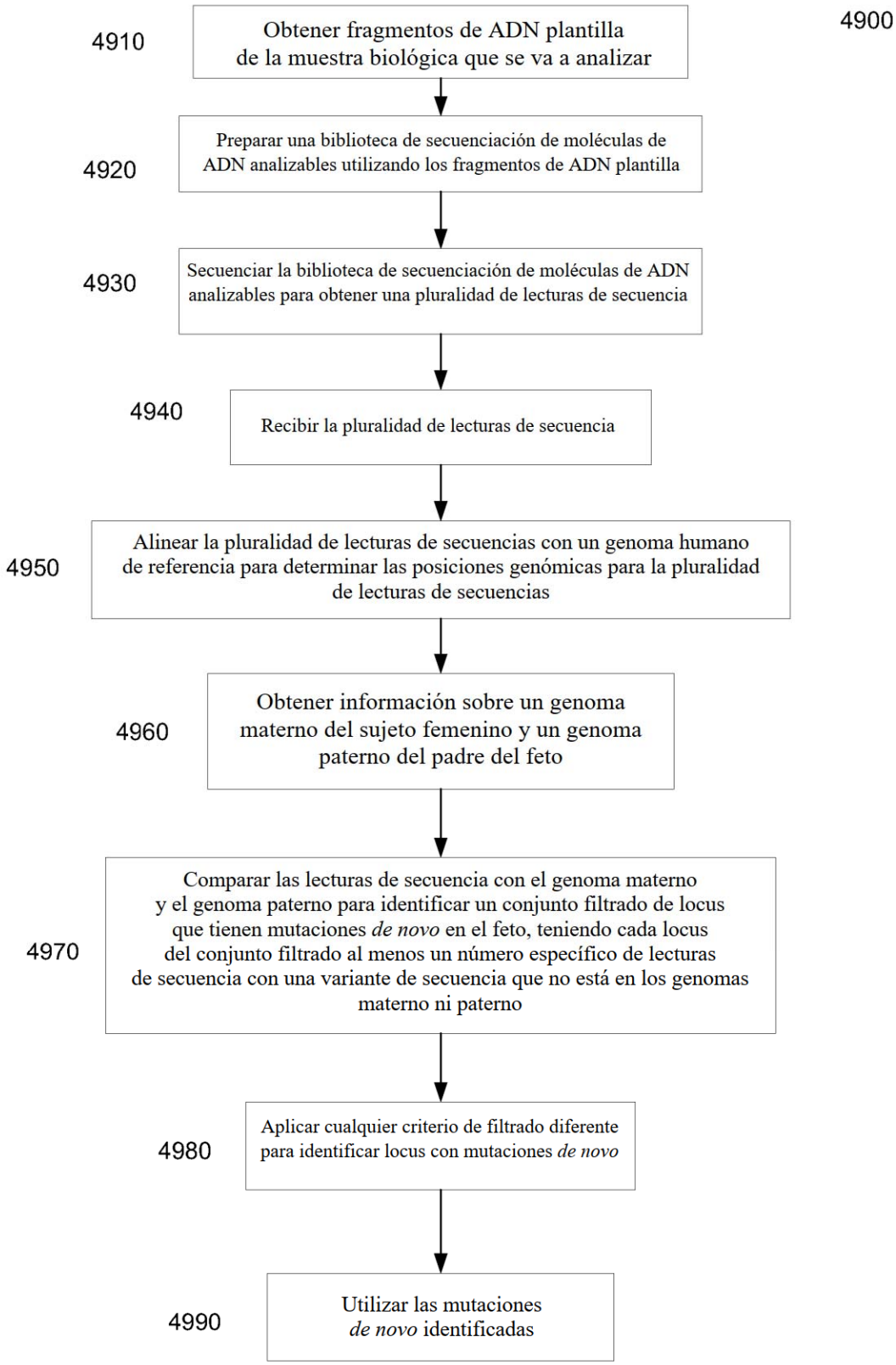

4900
4910
Obtener fragmentos de ADN plantilla de la muestra biológica que se va a analizar
4920
Preparar una biblioteca de secuenciación de moléculas de ADN analizables utilizando los fragmentos de ADN plantilla
4930
Secuenciar la biblioteca de secuenciación de moléculas de ADN analizables para obtener una pluralidad de lecturas de secuencia
4940
Recibir la pluralidad de lecturas de secuencia
4950
Alinear la pluralidad de lecturas de secuencias con un genoma humano de referencia para determinar las posiciones genómicas para la pluralidad de lecturas de secuencias
4960
Obtener información sobre un genoma materno del sujeto femenino y un genoma paterno del padre del feto
4970
Comparar las lecturas de secuencia con el genoma materno y el genoma paterno para identificar un conjunto filtrado de locus que tienen mutaciones *de novo* en el feto, teniendo cada locus del conjunto filtrado al menos un número específico de lecturas de secuencia con una variante de secuencia que no está en los genomas materno ni paterno
4980
Aplicar cualquier criterio de filtrado diferente para identificar locus con mutaciones *de novo*
4990
Utilizar las mutaciones *de novo* identificadas

FIG. 49

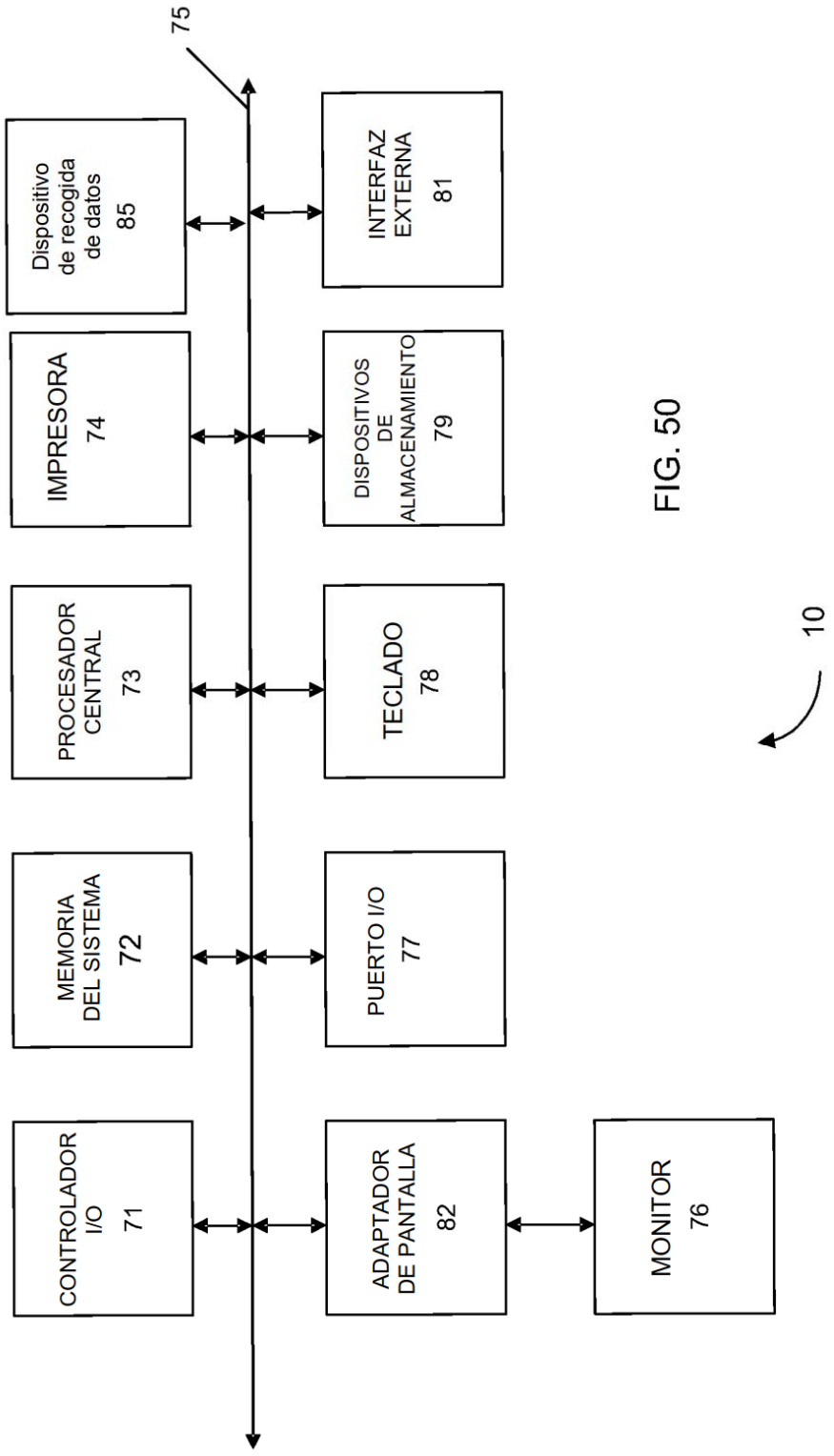

75
Dispositivo de recogida de datos 85
INTERFAZ EXTERNA 81
IMPRESORA 74
DISPOSITIVOS DE ALMACENAMIENTO 79
PROCESADOR CENTRAL 73
TECLADO 78
MEMORIA DEL SISTEMA 72
PUERTO I/O 77
CONTROLADOR I/O 71
ADAPTADOR DE PANTALLA 82
MONITOR 76
10

FIG. 50