(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 687 847**

(51) Int. CI.:

C12Q 1/68 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **14.06.2013 PCT/IB2013/054898**

(87) Fecha y número de publicación internacional: **27.12.2013 WO13190441**

(96) Fecha de presentación y número de la solicitud europea: **14.06.2013 E 13807105 (5)**

(97) Fecha y número de publicación de la concesión europea: **08.08.2018 EP 2864501**

(54) Título: **Análisis mutacional de ADN de plasma para la detección de cáncer**

(30) Prioridad:

21.06.2012 US 201261662878 P
13.08.2012 US 201261682725 P
31.08.2012 US 201261695795 P
08.10.2012 US 201261711172 P
13.03.2013 US 201313801748

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
29.10.2018

(73) Titular/es:

THE CHINESE UNIVERSITY OF HONG KONG (100.0%)
Shatin
New Territories, Hong Kong SAR, CN

(72) Inventor/es:

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

(74) Agente/Representante:

PONS ARIÑO, Ángel

DESCRIPCIÓN

Análisis mutacional de ADN de plasma para la detección de cáncer

Antecedentes

Se ha mostrado que el ADN derivado de tumor está presente en el plasma/suero libre de células de pacientes con cáncer (Chen XQ et al. Nat Med 1996; 2: 1033-1035). La mayoría de los métodos actuales se basan en el análisis directo de las mutaciones conocidas por estar asociadas con el cáncer (Diehl F et al. Proc Natl Acad Sci 2005; 102: 16368-16373; Forshew T et al. Sci Transl Med 2012; 4: 136ra68). Otro método ha investigado las variaciones en el número de copias asociadas con el cáncer detectadas mediante secuenciación aleatoria de ADN de plasma (Publicación de Patente de Estados Unidos 2013/0040824 de Lo et al.).

Se sabe que con el tiempo, más de una célula de cáncer adquirirá ventaja de crecimiento y producirá múltiples clones de células hijas. En última instancia, el crecimiento tumoral y/o sus focos metastásicos contendrían un conglomerado de grupos de células de cáncer clónales. Este fenómeno se conoce generalmente como heterogeneidad de tumor (Gerlinger M et al. N Engl J Med 2012; 366: 883-892; Yap TA et al. Sci Transl Med 2012; 4: 127psl0).

Se sabe que los cánceres son muy heterogéneos, es decir el perfil de mutación de los cánceres del mismo tipo de tejido puede variar ampliamente. Por lo tanto, el análisis directo de mutaciones específicas generalmente puede detectar sólo un subconjunto de los casos dentro de un tipo particular de cáncer conocido por estar asociado con esas mutaciones específicas. Además, el ADN derivado de tumor es generalmente la especie menor de ADN en el plasma humano; la concentración absoluta de ADN en el plasma es baja. Por lo tanto, la detección directa de una o un pequeño grupo de mutaciones asociadas con el cáncer en el plasma o suero pueden lograr sensibilidad analítica baja incluso entre los pacientes con cánceres que se sabe que albergan las mutaciones dirigidas. Además, se ha observado que existe una heterogeneidad intratumoral significativa en términos de mutaciones incluso dentro de un solo tumor. Las mutaciones se pueden encontrar en sólo una subpoblación de las células tumorales. La diferencia en los perfiles mutacionales entre el tumor primario y las lesiones metastásicas es aún más grande. Un ejemplo de heterogeneidad intratumoral y de metástasis primaria implica los genes *KRAS, BRAF* y *PIK3CA* en pacientes que sufren cánceres colorrectales (Baldus et al. Clin Cancer Research 2010. 16:790-9).

En un escenario en el que un paciente tiene un tumor primario (que lleva una mutación *KRAS* pero no una mutación *PIK3CA*) y una lesión metastásica oculta (que lleva una mutación *PIK3CA* pero no una mutación *KRAS*), si uno se centra en la detección de la mutación *KRAS* en el tumor primario, la lesión metastásica oculta puede no ser detectada. Sin embargo, si se incluye ambas mutaciones en el análisis, tanto en el tumor primario como la lesión metastásica oculta pueden ser detectadas. Por lo tanto, la prueba que involucra ambas mutaciones tendría una mayor sensibilidad en la detección de tejidos de tumor residuales. Tal ejemplo sencillo se vuelve más complejo cuando se está buscando el cáncer, y cuando uno tiene poca o ninguna idea de los tipos de mutaciones que pudieran ocurrir.

Por tanto, es deseable proporcionar nuevas técnicas para llevar a cabo una amplia selección, detección o evaluación del cáncer.

Sumario

De acuerdo con la presente invención, se proporciona un método para detectar el cáncer o cambio premaligno en un sujeto, comprendiendo el método: obtener una secuencia consenso de un genoma del sujeto, en el que la secuencia consenso se obtiene usando etiquetas de secuencia de una muestra del sujeto que contiene más del 50% de ADN de células sanas; recibir una o más etiquetas de secuencia para cada una de una pluralidad de fragmentos de ADN en una muestra biológica del sujeto, incluyendo la muestra biológica ADN libre de células; determinar posiciones genómicas para las etiquetas de secuencia; comparar las etiquetas de secuencia con la secuencia consenso para determinar un primer número de primeros loci, en el que: en cada uno de los primeros loci, un número de etiquetas de secuencia que tienen una variante de secuencia relativa a la secuencia consenso está por encima de un valor de corte, siendo el valor de corte mayor que uno; determinar un parámetro basado en un recuento de etiquetas de secuencia que tienen una variante de secuencia en los primeros loci; y comparar el parámetro con un valor umbral para determinar una clasificación de un nivel de cáncer en el sujeto, correspondiendo el valor umbral a un rango del parámetro para sujetos que tienen la clasificación del nivel de cáncer.

Las realizaciones pueden observar una frecuencia de mutaciones somáticas en una muestra biológica (por ejemplo, plasma o suero) de un sujeto sometido a detección o control del cáncer, cuando se compara con la del ADN constitucional del mismo sujeto. La secuenciación aleatoria se puede utilizar para determinar estas frecuencias. Un parámetro puede derivarse de estas frecuencias y utilizarse para determinar una clasificación de un nivel de cáncer. Los falsos positivos pueden ser filtrados requiriendo que cualquier locus variante tenga al menos un número especificado de lecturas de secuencia variante (etiquetas), proporcionando así un parámetro más preciso. Las frecuencias relativas para los diferentes loci variantes pueden ser analizadas para determinar un nivel de

heterogeneidad de los tumores en un paciente.

En una realización, el parámetro se puede comparar con el mismo parámetro derivado de un grupo de sujetos sin cáncer, o con un bajo riesgo de cáncer. Una diferencia significativa en el parámetro obtenido del sujeto de prueba y la del grupo de sujetos sin cáncer, o con un bajo riesgo de cáncer, puede indicar un riesgo aumentado de que el sujeto de prueba tenga cáncer o una afección premaligna o que desarrolle cáncer en el futuro. Por lo tanto, en una realización, el análisis de ADN de plasma puede llevarse a cabo sin la información genómica previa del tumor. Tal realización es, por lo tanto, especialmente útil para la detección de cáncer.

En otra realización, las realizaciones también se pueden utilizar para controlar un paciente con cáncer después del tratamiento y para ver si hay un tumor residual o si el tumor ha recidivado. Por ejemplo, un paciente con tumor residual o en el que el tumor ha recidivado tendría una mayor frecuencia de mutaciones somáticas que uno en el que no hay tumor residual o en el que no se observa recidiva tumoral. El control puede implicar la obtención de muestras de un paciente con cáncer en múltiples puntos temporales después del tratamiento para determinar las variaciones temporales de aberraciones genéticas asociadas con el tumor en los fluidos corporales u otras muestras con ácidos nucleicos sin células, por ejemplo, plasma o suero.

De acuerdo con una realización, un método detecta el cáncer o cambio premaligno en un sujeto. Se obtiene un genoma constitucional del sujeto. Se reciben una o más etiquetas de secuencia por cada uno de una pluralidad de fragmentos de ADN en una muestra biológica del sujeto, donde la muestra biológica incluye ADN libre de células. Las posiciones genómicas están determinadas por las etiquetas de secuencia. Las etiquetas de secuencia se comparan con el genoma constitucional para determinar un primer número de primeros loci. En cada uno de los primeros loci, un número de las etiquetas de secuencia que tienen una variante de secuencia respecto al genoma constitucional está por encima de un valor de corte, donde el valor de corte es mayor que uno. Un parámetro se determina basándose en un recuento de las etiquetas de secuencia que tienen una variante de secuencia en los primeros loci. El parámetro se compara con un valor umbral para determinar una clasificación de un nivel de cáncer en el sujeto.

De acuerdo con otra realización, un método analiza una heterogeneidad de uno o más tumores de un sujeto. Se obtiene un genoma constitucional del sujeto. Una o más etiquetas de secuencia se reciben para cada una de una pluralidad de fragmentos de ADN en una muestra biológica del sujeto, donde la muestra biológica incluye ADN libre de células. Las posiciones genómicas están determinadas por las etiquetas de secuencia. Las etiquetas de secuencia se comparan con el genoma constitucional para determinar un primer número de primeros loci. En cada uno de los primeros loci, un número de las etiquetas de secuencia que tienen una variante de secuencia respecto al genoma constitucional está por encima de un valor de corte, donde el valor de corte es mayor que uno. Una medida de la heterogeneidad de los uno o más tumores se calcula basándose en los primeros números respectivos del conjunto de primeras ubicaciones genómicas.

De acuerdo con otra realización, un método determina una concentración fraccional de ADN de tumor en una muestra biológica que incluye ADN libre de células. Una o más etiquetas de secuencia se reciben para cada una de una pluralidad de fragmentos de ADN en la muestra biológica. Las posiciones genómicas están determinadas por las etiquetas de secuencia. Para cada una de una pluralidad de regiones genómicas, una cantidad respectiva de fragmentos de ADN dentro de la región genómica se determina a partir de las etiquetas de secuencia que tienen una posición genómica dentro de la región genómica. La cantidad respectiva se normaliza para obtener una densidad respectiva. La densidad respectiva se compara con una densidad de referencia para identificar si la región genómica exhibe una pérdida de 1-copia o una ganancia de 1-copia. Una primera densidad se calcula a partir de las densidades respectivas identificadas por exhibir una pérdida de 1 copia o a partir de las densidades respectivas identificadas que exhiben una ganancia de 1-copia. La concentración fraccional se calcula comparando la primera densidad con otra densidad para obtener un diferencial, en el que el diferencial se normaliza con la densidad de referencia.

Otras realizaciones están dirigidas a sistemas y medios legibles por ordenador asociados con los métodos descritos en la presente memoria.

Una mejor comprensión de la naturaleza y ventajas de la presente invención puede ser obtenida con referencia a la siguiente descripción detallada y las figuras adjuntas.

Breve descripción de los dibujos

La FIG. 1 es un diagrama de flujo de un método 100 para detectar cáncer o cambio premaligno en un sujeto de acuerdo con las realizaciones de la presente invención.

La FIG. 2 muestra un diagrama de flujo de un método para comparar el genoma de muestra (SG) directamente con el genoma constitucional (CG) de acuerdo con las realizaciones de la presente invención.

La FIG. 3 muestra un diagrama de flujo de un método 300 que compara el genoma de muestra (SG) con el genoma constitucional (CG), utilizando el genoma de referencia (RG) de acuerdo con las realizaciones de la presente invención.

La FIG. 4 es una tabla 400 que muestra el número de mutaciones de nucleótido único asociadas con el cáncer identificadas correctamente utilizando diferentes números de ocurrencias como el criterio de clasificación de una mutación como presente en la muestra de acuerdo con las realizaciones de la presente invención cuando la concentración fraccional de ADN derivado de tumor en la muestra se supone que es 10 %.

La FIG. 5 es una tabla que muestra el número esperado de loci falsos positivos y el número esperado de mutaciones identificadas cuando la concentración fraccional de ADN derivado de tumor en la muestra se supone que es 5 %.

La FIG. 6A es una gráfica 600 que muestra la velocidad de detección de mutaciones asociadas con el cáncer en el plasma con 10 % y 20 % de concentraciones fraccionales en plasma de ADN derivado de tumor y utilizando cuatro y seis ocurrencias (r) como criterios para las mutaciones potenciales asociadas con el cáncer. La FIG. 6B es una gráfica 650 que muestra el número esperado de posiciones de nucleótidos falsamente clasificadas de tener un cambio de nucleótido utilizando criterios de ocurrencia (r) de 4, 5, 6 y 7 frente a la profundidad de secuenciación.

La FIG. 7A es una gráfica 700 que muestra el número de verdaderos sitios de mutación asociados con el cáncer y sitios falsos positivos con profundidades de secuenciación de diferencia cuando la concentración fraccional de ADN derivado de tumor en la muestra se supone que es el 5 %. La FIG. 7B es una gráfica 750 que muestra el número predicho de sitios falsos positivos que implican el análisis del genoma completo (WG) y todos los exones.

La FIG. 8 es una tabla 800 que muestra los resultados de 4 pacientes con HCC antes y después del tratamiento, incluyendo concentraciones fraccionales de ADN derivado de tumor en plasma de acuerdo con las realizaciones de la presente invención.

La FIG. 9 es una tabla 900 que muestra la detección de los SNV asociados con HCC en 16 sujetos de control sanos de acuerdo con las realizaciones de la presente invención.

La FIG. 10A muestra una gráfica de distribución de las densidades de lectura de secuencias de la muestra de tumor de un paciente con HCC de acuerdo con las realizaciones de la presente invención. La FIG. 10B muestra una gráfica de distribución 1050 de las puntuaciones z para todas las agrupaciones en el plasma de un paciente con HCC de acuerdo con las realizaciones de la presente invención.

La FIG. 11 muestra una gráfica de distribución 1100 de las puntuaciones z para el plasma de un paciente con HCC de acuerdo con las realizaciones de la presente invención.

La FIG. 12 es un diagrama de flujo de un método 1200 para determinar una concentración fraccional de ADN de tumor en una muestra biológica que incluye ADN libre de células de acuerdo con las realizaciones de la presente invención.

La FIG. 13A muestra una tabla 1300 del análisis de mutaciones en el plasma del paciente con cáncer de ovarios y un cáncer de mama en el momento del diagnóstico de acuerdo con las realizaciones de la presente invención.

La FIG. 13B muestra una tabla 1350 del análisis de mutaciones en el plasma de un paciente con cáncer de ovarios bilateral y un cáncer de mama después de la resección del tumor de acuerdo con las realizaciones de la presente invención.

La FIG. 14A es una tabla 1400 que muestra la detección de variaciones de nucleótido único en el ADN de plasma para HCC1. La FIG. 14B es una tabla 1450 que muestra la detección de variaciones de nucleótido único en el ADN de plasma para HCC2.

La FIG. 15A es una tabla 1500 que muestra la detección de variaciones de nucleótido único en el ADN de plasma para HCC3. La FIG. 15B es una tabla 1550 que muestra la detección de variaciones de nucleótido único en el ADN de plasma para HCC4.

La FIG. 16 es una tabla 1600 que muestra la detección de variaciones de nucleótido único en el ADN de plasma para el paciente con cáncer de ovario (y mama).

La FIG. 17 es una tabla 1700 que muestra las sensibilidades predichas de diferentes requisitos de ocurrencia y profundidades de secuenciación.

La FIG. 18 es una tabla 1800 que muestra los números predichos de loci falsos positivos para diferentes puntos de corte y diferentes profundidades de secuenciación.

La FIG. 19 muestra un diagrama de árbol que ilustra el número de mutaciones detectadas en los diferentes sitios de tumor.

La FIG. 20 es una tabla 2000 que muestra el número de fragmentos que llevan las mutaciones derivadas de tumor en la muestra de plasma de pre-tratamiento y post-tratamiento.

La FIG. 21 es una gráfica 2100 que muestra las distribuciones de ocurrencia en el plasma para las mutaciones detectadas en un sitio de tumor único y mutaciones detectadas en los cuatro sitios de tumor.

La FIG. 22 es una gráfica 2200 que muestra la distribución predicha de ocurrencia en el plasma para las mutaciones procedentes de un tumor heterogéneo.

La FIG. 23 demuestra la especificidad de realizaciones para 16 sujetos control sanos que fueron reclutados.

La FIG. 24 es un diagrama de flujo de un método 24 00 para analizar una heterogeneidad de uno o más tumores de un sujeto de acuerdo con las realizaciones de la presente invención.

La FIG. 25 muestra un diagrama de bloques de un ejemplo de sistema informático que se puede utilizar con el sistema y métodos de acuerdo con las realizaciones de la presente invención.

Definiciones

Como se utiliza en la presente memoria, el término "*locus*" o su forma plural "*loci*" es una ubicación o dirección de cualquier longitud de nucleótidos (o pares de bases) que puede tener una variación través de los genomas. Una "*agrupación (bin)*" es una región de longitud predeterminada en un genoma. Una pluralidad de agrupaciones puede tener una misma primera longitud (resolución), mientras que una pluralidad diferente puede tener una misma segunda longitud. En una realización, las agrupaciones no se superponen entre sí.

La expresión "*secuenciación aleatoria*" como se utiliza en la presente memoria, se refiere a la secuenciación mediante la cual los fragmentos de ácidos nucleicos secuenciados no han sido específicamente identificados o predeterminados antes del procedimiento de secuenciación. No se requieren cebadores específicos de la secuencia para dirigirse a loci de genes específicos. El término "*secuenciación universal*" se refiere a la secuencia donde la secuenciación se puede iniciar en cualquier fragmento. En una realización, se añaden adaptadores al extremo de un fragmento, y los cebadores para la secuenciación se unen a los adaptadores. Por lo tanto, cualquier fragmento puede ser secuenciado con el mismo cebador, y por lo tanto la secuenciación puede ser aleatoria.

La expresión "*etiqueta de secuencia*" (también conocida como lectura de secuencia) como se utiliza en la presente memoria, se refiere a la cadena de nucleótidos secuenciados a partir de cualquier parte o la totalidad de una molécula de ácido nucleico. Por ejemplo, una etiqueta secuenciada puede ser una cadena corta de nucleótidos (por ejemplo, ~ 30) secuenciados de un fragmento de ácido nucleico, una cadena corta de nucleótidos en ambos extremos de un fragmento de ácido nucleico, o la secuenciación del fragmento de ácido nucleico completo 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) con las otras partes de la molécula de ácido nucleico más grandes.

La expresión "*genoma constitucional*" (también denominada como CG) se compone de los nucleótidos de consenso en los loci dentro del genoma, y por lo tanto se puede considerar una secuencia consenso. El CG puede abarcar el genoma completo del sujeto (por ejemplo, el genoma humano), o sólo partes del genoma. El genoma constitucional (CG) se puede obtener de ADN de las células, así como de ADN libre de células (por ejemplo, como se puede encontrar en el plasma). Idealmente, los nucleótidos de consenso deberían indicar que un locus es homocigoto para un alelo o heterocigoto para dos alelos. Un locus heterocigoto generalmente contiene dos alelos que son miembros de un polimorfismo genético. A modo de ejemplo, los criterios para determinar si un locus es heterocigoto puede ser un umbral de dos alelos cada uno apareciendo en al menos un porcentaje predeterminado (por ejemplo, 30 % o 40 %) de lecturas alineadas con el locus. Si un nucleótido aparece en un porcentaje suficiente (por ejemplo, 70 % o más), entonces el locus puede ser determinado como homocigoto en el CG. Aunque el genoma de una célula sana puede diferir del genoma de otra célula sana debido a mutaciones aleatorias que ocurren de forma espontánea durante la división celular, el CG no debe variar cuando se usa tal consenso. Algunas células pueden tener genomas con reordenaciones genómicas, por ejemplo, linfocitos B y T, tales como las que implican genes del receptor de linfocitos T y anticuerpos. Tales diferencias a gran escala todavía serían una población relativamente pequeña de la población total de células nucleadas en la sangre, y por lo tanto, tales reordenaciones no afectarían a la determinación del genoma constitucional con muestreo suficiente (por ejemplo, profundidad de secuenciación) de las células sanguíneas. Otros tipos de células, incluyendo células bucales, células de la piel, folículos pilosos, o biopsias de diversos tejidos corporales normales, también pueden servir como fuentes de CG.

La expresión "*ADN constitucional*" se refiere a cualquier fuente de ADN que es un reflejo de la composición genética con la que nace un sujeto. Para un sujeto, los ejemplos de "muestras constitucionales", a partir de las cuales se puede obtener el ADN constitucional, incluyen ADN de células sanguíneas sanas, ADN de células bucales y ADN de la raíz del cabello. El ADN de estas células sanas define el CG del sujeto. Las células pueden ser identificadas como sanas de distintas maneras, por ejemplo, cuando una persona se sabe que no tiene cáncer o la muestra puede ser obtenida de tejido que no es probable que contenga células de cáncer o premalignas (por ejemplo, ADN de raíz del cabello cuando se sospecha de cáncer de hígado). En otro ejemplo, una muestra de plasma se puede obtener cuando un paciente está libre de cáncer, y el ADN constitucional determinado se compara con los resultados de una muestra de plasma posterior (por ejemplo, un año o más). En otra realización, se puede utilizar una muestra biológica única que contiene < 50 % del ADN de tumor para deducir el genoma constitucional y las alteraciones genéticas asociadas con el tumor. En tal muestra, las concentraciones de mutaciones de nucleótido único asociadas con el tumor serían menores que las de cada alelo de SNP heterocigotos en el CG. Tal muestra puede ser la misma que la muestra biológica utilizada para determinar un genoma de muestra, descrito a continuación.

La expresión "*muestra biológica*" como se utiliza en la presente memoria, se refiere a cualquier muestra que se toma de un sujeto (por ejemplo, un ser humano, una persona con cáncer, una persona con sospecha de tener cáncer, u otros organismos) y que contiene una o más moléculas de ácido nucleico libres de las células de interés. Una muestra biológica puede incluir ADN libre de células, algunas de las cuales puede tener su origen en las células sanas y alguna de las células tumorales. Por ejemplo, el ADN de tumor se puede encontrar en la sangre u otros fluidos, por ejemplo, orina, fluido pleural, fluido ascítico, líquido peritoneal, saliva, lágrimas o líquido cefalorraquídeo. Un ejemplo de una muestra que no es un líquido es una muestra de heces, la cual puede ser mezclada con el fluido diarreico. Para algunas de las muestras, la muestra biológica puede obtenerse de forma no invasiva. En algunas realizaciones, la muestra biológica puede ser utilizada como una muestra constitucional.

La expresión "*genoma de muestra*" (también denominada SG) es una colección de lecturas de secuencia 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 en sólo un número suficiente de lecturas (por ejemplo, al menos 2 o 3, o valores de corte mayores). 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 nucleótido único" (también referida como una SNM). Otros tipos de mutaciones también se pueden detectar utilizando la presente invención, por ejemplo, mutaciones que implican dos o más nucleótidos, (como los que afectan al 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) y la inversión de secuencia.

La expresión "*genoma de referencia*" (también denominada RG) se refiere a un genoma haploide o diploide con el cual se pueden alinear y comparar las lecturas de secuencia de la muestra biológica y la muestra constitucional. Para un genoma haploide, sólo hay un nucleótido en cada locus. Para un genoma diploide, los loci heterocigotos pueden ser identificados, teniendo cada locus dos alelos, donde cualquier alelo 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 de tumor, y/u otra medida de la gravedad de un cáncer. El nivel de cáncer podría ser varios u otros caracteres. El nivel podría ser cero. El nivel de cáncer también incluye afecciones (estados) premalignas o precancerosas asociadas con las mutaciones o varias mutaciones. El nivel de cáncer se puede utilizar de varias maneras. Por ejemplo, la detección puede comprobar si el cáncer está presente en alguien que no se sabe de antemano que tenía cáncer. La evaluación puede investigar a alguien que ha sido diagnosticado con cáncer. La detección puede significar "cribar" o puede significar comprobar si alguien, con características sugestivas de cáncer (por ejemplo, síntomas u otras pruebas positivos), tiene cáncer.

Descripción detallada

Las realizaciones se proporcionan para la detección de cáncer mediante el análisis de una muestra biológica (por ejemplo, una muestra de plasma/suero de la sangre) que no se toma directamente de un tumor e incluye ácidos nucleicos libres de células. Los ácidos nucleicos libres de células pueden obtenerse a partir de varios tipos de tejidos de todo el cuerpo. De esta manera, se puede realizar un amplio análisis para la detección de varios cánceres.

Las aberraciones genéticas (incluyendo mutaciones de nucleótido único, deleciones, amplificaciones, y reordenaciones) se acumulan en las células tumorales durante el desarrollo de los cánceres. En realizaciones, la secuenciación masivamente paralela se puede utilizar para detectar y cuantificar las mutaciones de nucleótido único (SNM), también llamadas variaciones de nucleótido único (SNV), en fluidos corporales (por ejemplo, plasma, suero, saliva, fluido ascítico, fluido pleural y líquido cefalorraquídeo) para detectar y controlar los cánceres. Una cuantificación del número de SNM (u otros tipos de mutaciones) puede proporcionar un mecanismo para identificar estadios tempranos del cáncer como parte de las pruebas de detección. En diversas aplicaciones, se procura distinguir los errores de secuenciación y distinguir las mutaciones espontáneas que se producen en las células sanas (por ejemplo, requiriendo que múltiples SNM se identifiquen en un locus particular, por ejemplo, al menos 3, 4, o 5).

Algunas realizaciones también proporcionan métodos no invasivos para el análisis de la heterogeneidad del tumor, que pueden implicar células dentro del mismo tumor (es decir, heterogeneidad intratumoral) o células de diferentes tumores (ya sea del mismo sitio o de diferentes sitios) dentro de un cuerpo. Por ejemplo, se puede analizar de forma no invasiva la estructura clonal de tal heterogeneidad del tumor, incluyendo una estimación de la masa de células tumorales relativa que contiene cada mutación. Las mutaciones que están presentes en concentraciones relativas mayores están presentes en un mayor número de células malignas en el cuerpo, por ejemplo, las células que han aparecido previamente durante el proceso tumorigénico con respecto a otras células malignas presentes todavía en el cuerpo (Welch JS et al. Cell 2012; 150: 264-278). Tales mutaciones, debido a su mayor abundancia relativa, se espera que exhiban una mayor sensibilidad de diagnóstico para detectar ADN de cáncer que aquellas con menor abundancia relativa. Un control seriado del cambio de la abundancia relativa de mutaciones permitiría controlar de forma no invasiva el cambio en la arquitectura clonal de los tumores, ya sea de forma espontánea cuando la enfermedad progresa, o en respuesta al tratamiento. Tal información sería de utilidad en el pronóstico de evaluación o en la detección temprana de resistencia del tumor al tratamiento.

I. INTRODUCCIÓN

Las mutaciones pueden ocurrir durante la división celular debido a los errores en la replicación del ADN y/o la reparación del ADN. Un tipo de tales mutaciones implica la alteración de nucleótidos únicos, que pueden implicar múltiples secuencias de diferentes partes del genoma. Se cree generalmente que los cánceres son debidos a la expansión clonal de una célula de cáncer única que ha adquirido ventaja de crecimiento. Esta expansión clonal conduciría a la acumulación de mutaciones (por ejemplo mutaciones de nucleótido único) en todas las células de cáncer procedentes de la célula de cáncer ancestral. Estas células tumorales de la progenie compartirían un conjunto de mutaciones (por ejemplo, mutaciones de nucleótido único). Como se describe en la presente memoria, las mutaciones de nucleótido único asociadas con el cáncer son detectables en el plasma/suero de pacientes con cáncer.

Algunas realizaciones pueden detectar con eficacia todas las mutaciones en una muestra biológica (por ejemplo, el plasma o suero). Ya que el número de mutaciones no es fijo (se pueden detectar cientos, miles o millones de mutaciones asociadas con el cáncer de diferentes subpoblaciones), las realizaciones pueden proporcionar una mejor sensibilidad que las técnicas que detectan mutaciones específicas. El número de mutaciones se puede utilizar para detectar el cáncer.

Para proporcionar tal nivel de detección de muchas o todas las mutaciones, las realizaciones pueden realizar una búsqueda (por ejemplo, una búsqueda aleatoria) de las variaciones genéticas en una muestra biológica (por ejemplo, fluidos corporales, incluyendo plasma y suero), que podrían contener ADN derivado de tumor. El uso de una muestra, tal como plasma, evita la necesidad de realizar una biopsia invasiva del tumor o cáncer. Además, como la detección puede cubrir la totalidad o grandes regiones del genoma, la detección no se limita a algunas mutaciones enumerables y conocidas, sino que también puede utilizar la existencia de cualquier mutación. Además, dado que el número de mutaciones se suma a través de todas o grandes regiones del genoma, se puede obtener una mayor sensibilidad.

Sin embargo, hay sitios polimórficos, incluyendo polimorfismos de nucleótido único (SNP), en el genoma humano, que no deben ser contados en las mutaciones. Las realizaciones pueden determinar si es probable que las variaciones genéticas que se han detectado sean mutaciones asociadas con el cáncer o sean polimorfismos en el genoma. Por ejemplo, como parte de la determinación entre las mutaciones asociadas con el cáncer y polimorfismos en el genoma, las realizaciones pueden determinar un genoma constitucional, que puede incluir polimorfismos. Los polimorfismos del genoma constitucional (CG) pueden limitarse a polimorfismos que se exhiben con un porcentaje suficientemente alto (por ejemplo, 30-40 %) en los datos de secuenciación.

Las secuencias obtenidas de la muestra biológica pueden a continuación ser alineadas con el genoma constitucional y se identifican las variaciones que son mutaciones de nucleótido único (SNM), u otros tipos de mutaciones. Estas SNM serían variaciones que no están incluidas en los polimorfismos conocidos, y por lo tanto pueden ser marcadas como asociadas con el cáncer, y no parte del genoma constitucional. Una persona sana puede tener un cierto número de SNM debido a las mutaciones aleatorias entre las células sanas, por ejemplo, creadas durante la división celular, pero una persona con cáncer tendría más.

Por ejemplo, en una persona con cáncer, el número de SNM detectables en un fluido corporal sería mayor que los polimorfismos presentes en el genoma constitucional de la misma persona. Se puede hacer una comparación entre las cantidades de las variaciones detectadas en una muestra de fluido corporal que contiene ADN derivado de tumor y una muestra de ADN que contiene mayoritariamente ADN constitucional. En una realización, el término “mayoritariamente” significaría más de 90 %. En otra realización preferida, el término “mayoritariamente” significaría más de 95, 97 %, 98 %, o 99 %. Cuando la cantidad de variaciones en el fluido corporal es superior a la de la muestra con ADN mayoritariamente constitucional, existe una mayor probabilidad de que el fluido corporal pueda contener ADN derivado de tumor.

Un método que podría utilizarse para buscar aleatoriamente las variaciones en las muestras de ADN es la secuenciación aleatoria o secuenciación de disparo de pistola (por ejemplo, utilizando secuenciación masivamente paralela). Cualquier plataforma de secuenciación masivamente paralela se puede usar, incluyendo una plataforma de secuenciación por ligación (por ejemplo, la plataforma Life Technologies SOLiD), la Ion Torrent/Ion Proton, la secuenciación por semiconductores, Roche 454 y las plataformas de secuenciación molecular única (por ejemplo Helicos, Pacific Biosciences y Nanopore). Sin embargo, se sabe que pueden producirse errores de secuenciación y pueden ser mal interpretados como una variación en el ADN constitucional o como mutaciones derivadas de ADN de tumor. Por lo tanto, para mejorar la especificidad de nuestro enfoque propuesto, la probabilidad del error de secuenciación u otros componentes de errores analíticos se puede explicar, por ejemplo, mediante el uso de una profundidad de secuenciación apropiada junto con el requisito de al menos un número especificado (por ejemplo, 2 o 3) de alelos detectados en un locus para que se cuente como una SNM.

Como se describe en la presente memoria, las realizaciones pueden proporcionar evidencia de la presencia de ADN derivado de tumor en una muestra biológica (por ejemplo, un fluido corporal) cuando la cantidad de variaciones genéticas aleatoriamente detectadas presentes en la muestra supera lo esperado para el ADN constitucional y a las variaciones que pueden ser inadvertidamente detectadas debido a los errores analíticos (por ejemplo, errores de secuenciación). La información podría ser utilizada para la detección, diagnóstico, pronóstico y control de cánceres. En las siguientes secciones, se describen las etapas analíticas que se pueden utilizar para la detección de mutaciones de nucleótido único en el plasma/suero u otras muestras (por ejemplo, fluidos corporales). Los fluidos corporales podrían incluir plasma, suero, líquido cefalorraquídeo, fluido pleural, fluido ascítico, secreción del pezón, saliva, fluido de lavado broncoalveolar, esputo, lágrimas, sudor y orina. Además de los fluidos corporales, la tecnología también se puede aplicar a muestras de heces, ya que se ha demostrado que estas últimas contienen ADN de tumor de cáncer colorrectal (Berger BM, Ahlquist DA. Pathology 2012; 44: 80- 5 88).

II. MÉTODO DE DETECCIÓN GENERAL

La FIG. 1 es un diagrama de flujo de un método 100 para detectar cáncer o un cambio premaligno en un sujeto de acuerdo con las realizaciones de la presente invención. Las realizaciones pueden analizar el ADN libre de células en una muestra biológica del sujeto para detectar las variaciones en el ADN libre de células que probablemente sean el resultado de un tumor. El análisis puede utilizar un genoma constitucional del sujeto para explicar los polimorfismos que son parte de las células sanas, y puede explicar los errores de secuenciación. El método 100 y cualquiera de los métodos descritos en la presente memoria pueden ser realizados total o parcialmente con un sistema informático que incluye uno o más procesadores.

En la etapa 110, se obtiene un genoma constitucional del sujeto. El genoma constitucional (CG) se puede determinar a partir del ADN constitucional del sujeto analizado. En diversas realizaciones, el CG se puede leer de la memoria o se determina de forma activa, por ejemplo, analizando las lecturas de secuencia del ADN constitucional, que pueden estar en las células de la muestra que incluye el ADN libre de células. Por ejemplo, cuando se sospecha de una neoplasia maligna no hematológica, las células sanguíneas pueden ser analizadas para determinar el ADN constitucional del sujeto.

En diversas aplicaciones, el análisis del ADN constitucional podría llevarse a cabo utilizando secuenciación masivamente paralela, hibridación basada en matriz, hibridación en solución basada en sonda, ensayos basados en ligación, ensayos de reacción de extensión de cebador, y espectrometría de masas. En una realización, el CG se puede determinar en un momento en la vida de un sujeto, por ejemplo, al nacer o incluso en el período prenatal (que se podría hacer usando células fetales o a través del fragmento de ADN libre de células, véase la Publicación de Estados Unidos 2011/0105353), y comparar después cuando los fluidos corporales u otras muestras se obtienen en otros momentos en la vida del sujeto. Por lo tanto, el CG simplemente puede leerse de la memoria del ordenador. El genoma constitucional puede leerse como una lista de loci donde el genoma constitucional difiere de un genoma de referencia.

En la etapa 120 se reciben una o más etiquetas de secuencia para cada una de una pluralidad de fragmentos de ADN en una muestra biológica del sujeto, donde la muestra biológica incluye ADN libre de células. En una realización, se generan una o más etiquetas de secuencia de una secuenciación aleatoria de fragmentos de ADN en la muestra biológica. Más de una etiqueta de secuencia se puede obtener cuando se realiza la secuenciación del extremo apareado. Una etiqueta correspondería a cada extremo del fragmento de ADN.

El ADN libre de células en la muestra (por ejemplo, plasma, suero u otro fluido corporal) se puede analizar para buscar variaciones genéticas. El ADN libre de células puede ser analizado utilizando la misma plataforma analítica que la que se ha utilizado para analizar el ADN constitucional. En otra alternativa, podría usarse una plataforma analítica diferente. Por ejemplo, la muestra de ADN libre de células puede ser secuenciada utilizando secuenciación masivamente paralela o partes del genoma podrían ser capturadas o enriquecidas antes de la secuenciación masivamente paralela. Si se utiliza el enriquecimiento, se podría, por ejemplo, utilizar captura en fase de solución o en fase sólida de partes seleccionadas del genoma. A continuación, la secuenciación masivamente paralela puede llevarse a cabo en el ADN capturado.

En la etapa 130, se determinan las posiciones genómicas para las etiquetas de secuencia. En una realización, las etiquetas de secuencia se alinean con un genoma de referencia, que se obtiene de uno o más de otros sujetos. En otra realización, las etiquetas de secuencia genómica están alineadas con el genoma constitucional del sujeto analizado. La alineación se puede realizar utilizando técnicas conocidas por un experto en la materia, por ejemplo, utilizando la Herramienta de Búsqueda de Alineación Local Básica (BLAST).

En la etapa 140, se determinan un primer número de loci donde al menos N etiquetas de secuencia tienen una variante de secuencia respecto al genoma constitucional (CG) . N es igual o mayor que dos. Como se describe más detalladamente a continuación, los errores de secuenciación, así como las mutaciones somáticas que ocurren aleatoriamente en las células (por ejemplo, debido a la división celular) pueden ser eliminados cuando tienen N igual a 2, 3, 4, 5, o mayor. Los loci que satisfacen uno o más criterios especificados pueden ser identificados como una mutación (variante) o loci de mutación (loci variantes), mientras que un locus que tiene una variante pero que no cumple uno o más criterios (por ejemplo, simplemente como una etiqueta de secuencia variante) se considera como una mutación potencial o putativa. La variante de secuencia podría ser solo un nucleótido o varios nucleótidos.

N puede determinarse como porcentaje de etiquetas totales para un locus, en lugar de un valor absoluto. Por ejemplo, un locus variante puede ser identificado cuando la concentración fraccional de ADN de tumor inferida de las lecturas variantes se determina que es igual o mayor que 10 % (o algún otro porcentaje). En otras palabras, cuando el locus está cubierto por 200 lecturas de secuencia, puede necesitarse un criterio de al menos 10 lecturas de secuencia que muestra el alelo variante para definir la variante como una mutación. Las 10 lecturas de secuencia del alelo variante y 190 lecturas del alelo tipo silvestre daría una concentración fraccional de ADN de tumor de 10 % (2x10/(10 + 190)).

En una realización, las etiquetas de secuencia (denominadas colectivamente como el genoma de muestra) se pueden comparar directamente con el CG para determinar las variantes. En otra realización, el genoma de muestra (SG) se compara con el CG a través de un genoma de referencia (RG) para determinar las variantes. Por ejemplo, tanto el CG como SG pueden ser comparados con el RG para determinar los números respectivos (por ejemplo, conjuntos) de loci que exhiben variantes, y luego se puede considerar una diferencia para obtener el primer número de loci. El primer número simplemente se puede obtener como un número o puede corresponder a un conjunto específico de loci, que luego pueden ser analizados adicionalmente para determinar un parámetro de las etiquetas de secuencia en los primeros loci.

En una aplicación, los resultados de la secuenciación de ADN constitucional y ADN de plasma se comparan para determinar si una mutación de nucleótido único está presente en el ADN de plasma. Las regiones en las que el ADN constitucional es homocigoto pueden ser analizadas. Para fines ilustrativos, se supone que el genotipo de un locus particular es homocigoto en el ADN constitucional y es AA. Luego en el plasma, la presencia de un alelo distinto de A indicaría la presencia potencial de una mutación de nucleótido único (SNM) en el locus particular. Los loci que indican la presencia potencial de una SNM pueden formar el primer número de loci en la etapa 140.

En una realización, podría ser útil seleccionar las partes del genoma que se sabe que son particularmente propensas a la mutación en un tipo de cáncer particular o en un subconjunto particular de la población. Es importante en este último aspecto, que las realizaciones puedan buscar tipos de mutaciones que son particularmente frecuentes en un grupo de población específica, por ejemplo, mutaciones que son especialmente comunes en los sujetos que son portadores del virus de la hepatitis B (para cáncer de hígado) o virus del papiloma humano (para cáncer cervical) o que tienen predisposición genética a mutaciones somáticas o sujetos con mutaciones de la línea germinal en un gen de reparación del mal apareamiento de ADN. La tecnología también podría ser útil para detectar mutaciones en los cánceres de ovario y mama en sujetos con mutaciones *BRCA1* o *BRCA2*. La tecnología sería igualmente útil para detectar mutaciones en el cáncer colorrectal en sujetos con mutaciones en *APC*.

En la etapa 150, se determina un parámetro en función del recuento de las etiquetas de secuencia que tienen una variante de secuencia en los primeros loci. En un ejemplo, el parámetro es el primer número de loci donde al menos N fragmentos de ADN tienen una variante de secuencia en un locus respecto al genoma constitucional. Por lo tanto, el recuento puede ser utilizado simplemente para asegurar que un locus tiene más de N copias de una variante particular identificada antes de ser incluida en el primer número. En otra realización, el parámetro puede ser o incluir el número total de etiquetas de secuencia que tienen una variante de secuencia respecto al genoma constitucional en los primeros loci.

En la etapa 160, el parámetro para el sujeto se compara con un valor umbral (por ejemplo, derivado de uno o más otros sujetos) para determinar una clasificación de un nivel de cáncer en el sujeto. Los ejemplos de un nivel de cáncer incluyen si el sujeto tiene cáncer o una afección premaligna, o un aumento de la probabilidad de desarrollar cáncer. En una realización, el valor umbral puede determinarse de una muestra obtenida previamente del sujeto.

En otra realización, se puede determinar que uno o más otros sujetos no tienen cáncer o un bajo riesgo de cáncer. Por lo tanto, el valor umbral puede ser un valor normal, un rango normal, o indicar una desviación estadísticamente significativa de un valor o rango normal. Por ejemplo, el número de mutaciones con relación al CG de un sujeto

específico, detectable en el plasma de sujetos sin un cáncer o con un bajo riesgo de cáncer, se puede utilizar como el rango normal para determinar si el número de mutaciones detectadas en el sujeto analizado es normal. En otra realización, se podría conocer que otros sujetos tienen cáncer, y por lo tanto un número similar de mutaciones pueden indicar cáncer.

En una aplicación, los otros sujetos se pueden seleccionar de modo que tengan unas características clínicas que sean coincidentes con las del sujeto de prueba, por ejemplo, sexo, edad, dieta, antecedentes de tabaquismo, historial de consumo de drogas, enfermedad previa, antecedentes familiares, genotipos de loci genómicos seleccionados, estado de las infecciones virales (por ejemplo infección por virus de hepatitis B o C o virus del papiloma humano o virus de inmunodeficiencia humana o virus de Epstein-Barr) o infecciones con otros agentes infecciosos (tales como bacterias (por ejemplo, *Helicobacter pylori*) y parásitos (por ejemplo, *Clonorchis sinensis*), etc. Por ejemplo, los sujetos que son portadores del virus de la hepatitis B o C tienen un mayor riesgo de desarrollar carcinoma hepatocelular. Por lo tanto, los sujetos de prueba que tienen un número similar o patrón de mutaciones como un portador de la hepatitis B o C se puede considerar que tienen un aumento del riesgo de desarrollar carcinoma hepatocelular. Por otro lado, un paciente con hepatitis B o C que exhibe más mutaciones que otro paciente con hepatitis puede ser identificado apropiadamente y ser incluido en un nivel de clasificación del cáncer mayor, ya que se utiliza el valor basal apropiado (es decir, en relación con otro paciente con hepatitis). Del mismo modo, los sujetos que son portadores de la infección por virus del papiloma humano tienen mayor riesgo de cáncer cervical, y cáncer de cabeza y cuello. La infección con el virus de Epstein-Barr se ha asociado con el carcinoma nasofaríngeo, cáncer gástrico, linfoma de Hodgkin y linfoma no Hodgkin. La infección con *Helicobacter pylori* se ha asociado con el cáncer gástrico. La infección con *Clonorchis sinensis* se ha asociado con colangiocarcinoma.

El control de los cambios del número de mutaciones en diferentes puntos temporales se puede utilizar para controlar el progreso del cáncer y la respuesta al tratamiento. Tal control también se puede utilizar para documentar el progreso de una afección premaligna o cambio en el riesgo de que un sujeto desarrolle cáncer.

La cantidad de etiquetas de secuencia que muestran variaciones también se puede utilizar para el control. Por ejemplo, se puede usar una concentración fraccional de las lecturas variantes en un locus. En una realización, un aumento en las concentraciones fraccionales de aberraciones genéticas asociadas con el tumor en las muestras durante el control serial puede significar la progresión de la enfermedad o la recidiva inminente. Del mismo modo, una disminución en las concentraciones fraccionales de aberraciones genéticas asociadas con el tumor en las muestras durante el control serial puede significar la respuesta al tratamiento y/o remisión y/o buen pronóstico.

III. DETERMINACIÓN DE GENOMAS

Los diversos genomas discutidos anteriormente se explican con más detalle a continuación. Por ejemplo, se describen el genoma de referencia, el genoma constitucional y el genoma de muestra.

A. Genoma de referencia

El genoma de referencia (RG) se refiere a un genoma haploide o diploide de un sujeto o consenso de una población. El genoma de referencia es conocido y por lo tanto puede ser utilizado para comparar las lecturas de secuenciación de nuevos pacientes. Las lecturas de secuenciación de una muestra de un paciente pueden ser alineadas y comparadas para identificar variaciones en las lecturas del RG. Para un genoma haploide, sólo hay un nucleótido en cada locus, y por lo tanto cada locus puede ser considerado hemicigoto. Para un genoma diploide, los loci heterocigotos pueden ser identificados, teniendo tal locus dos alelos, donde cualquier alelo puede permitir una coincidencia para la alineación al locus.

Un genoma de referencia puede ser el mismo entre una población de sujetos. Este mismo genoma de referencia se puede utilizar para los sujetos sanos para determinar el umbral adecuado para ser utilizado para clasificar el paciente (por ejemplo, tiene cáncer o no). Sin embargo, se pueden usar diferentes genomas de referencia para diferentes poblaciones, por ejemplo, para diferentes grupos étnicos o incluso para diferentes agrupaciones.

B. Genoma constitucional

El genoma constitucional (CG) de un sujeto (por ejemplo, un ser humano u otro organismo diploide) se refiere a un genoma diploide del sujeto. El CG puede especificar loci heterocigotos donde un primer alelo es de un primer haplotipo y un segundo alelo diferente es de un segundo haplotipo. Hay que tener en cuenta que las estructuras de dos haplotipos que cubren dos loci heterocigotos no necesitan ser conocidas, es decir, qué alelo en un locus heterocigoto está en el mismo haplotipo ya que un alelo de otro locus heterocigoto no necesita ser conocido. Sólo la existencia de los dos alelos en cada locus heterocigoto puede ser suficiente.

El CG puede diferir del RG debido a los polimorfismos. Por ejemplo, un locus en el RG puede ser homocigoto para T, pero el CG es heterocigoto para T/A. Por lo tanto, el CG exhibiría una variación en este locus. El CG también puede ser diferente del RG debido a mutaciones heredadas (por ejemplo, transmitidas en la familia) o mutaciones de novo (que se producen en un feto, pero que no están presentes en sus padres). La mutación heredada se suele

llamar 'mutación de la línea germinal'. Algunas de tales mutaciones están asociadas con la predisposición al cáncer, tal como una mutación *BRCA1* que se transmite en la familia. Tales mutaciones son diferentes de las 'mutaciones somáticas' que pueden ocurrir debido a la división celular durante la vida de una persona y que pueden empujar a una célula y su progenie a convertirse en un cáncer.

Un objetivo de la determinación del CG es eliminar tales mutaciones de la línea germinal y las mutaciones de novo de las mutaciones del genoma de muestra (SG) para identificar las mutaciones somáticas. La cantidad de mutaciones somáticas en el SG se puede utilizar después para evaluar la probabilidad de cáncer en el sujeto. Estas mutaciones somáticas pueden ser filtradas adicionalmente para eliminar errores de secuenciación, y potencialmente eliminar mutaciones somáticas que ocurren rara vez (por ejemplo, sólo una lectura que muestra una variante), ya que tales mutaciones somáticas no están probablemente relacionadas con el cáncer.

En una realización, un CG puede ser determinado utilizando células (ADN de capa leucocitaria). Sin embargo, el CG también se puede determinar a partir del ADN libre de células (por ejemplo, plasma o suero). Para un tipo de muestra en la que la mayoría de las células son no malignas, por ejemplo, la capa leucocitaria de un sujeto sano, el genoma mayoritario o consenso es el CG. Para el CG, cada locus genómico consiste en la secuencia de ADN poseída por la mayoría de las células en el tejido muestreado. La profundidad de la secuenciación debe ser suficiente para dilucidar sitios heterocigotos dentro del genoma constitucional.

En otro ejemplo, el plasma puede ser utilizado como la muestra constitucional para determinar el CG. Por ejemplo, para casos en los que el ADN de tumor en plasma es menor de 50 % y una SNM está en un estado heterocigoto, por ejemplo, la mutación es la adición de un nuevo alelo, entonces el nuevo alelo puede tener una concentración de menos de 25 %. Si bien. la concentración de los alelos heterocigotos de SNP en el CG debería ascender a aproximadamente 50 %. Por lo tanto, se puede hacer una distinción entre una mutación somática y un polimorfismo del CG. En una aplicación, un punto de corte adecuado puede estar entre 30-40 % para determinar una mutación somática de un polimorfismo cuando se usa plasma, u otras mezclas con una concentración significativa del tumor. Una medición de la concentración de ADN de tumor puede ser útil para asegurar que el ADN de tumor en plasma es menos de 50 %. Los ejemplos de la determinación de una concentración de ADN de tumor se describen en la presente memoria.

C. Genoma de muestra

El genoma de muestra (SG) no es simplemente un genoma haploide o diploide como es el caso del RG y del CG. El SG es una colección de lecturas de la muestra, y puede incluir: lecturas de ADN constitucional que corresponden al CG, lecturas de ADN de tumor, lecturas de células sanas que muestran mutaciones aleatorias relacionadas con el CG (por ejemplo, debido a mutaciones que resultan de la división celular), y errores de secuenciación. Se pueden utilizar varios parámetros para controlar exactamente que lecturas están incluidas en el SG. Por ejemplo, requerir un alelo para mostrar al menos 5 lecturas puede disminuir los errores de secuenciación presentes en el SG, así como disminuir las lecturas debido a las mutaciones aleatorias.

A modo de ejemplo, supongamos que el sujeto está sano, es decir, no tiene cáncer. Con fines ilustrativos, el ADN de 1000 células está en 1 ml de plasma (es decir, 1000 equivalentes de genoma de ADN) obtenido de este sujeto. El ADN de plasma consiste generalmente en fragmentos de ADN de aproximadamente 150 pb. Dado que el genoma humano tiene $3x10^9$ pb, habría aproximadamente $2x10^7$ fragmentos de ADN por genoma haploide. Ya que el genoma humano es diploide, habría aproximadamente $4x10^7$ fragmentos de ADN por ml de plasma.

Ya que de millones a miles de millones de células están liberando su ADN en el plasma por unidad de tiempo y los fragmentos de estas células se mezclarían juntos durante la circulación, los $4x10^7$ fragmentos de ADN podrían venir de $4x10^7$ células diferentes. Si estas células no tienen una relación clonal entre sí reciente (es decir, que no comparten una célula ancestral reciente) (en oposición a distante, por ejemplo, el cigoto original) entonces es estadísticamente probable que ninguna mutación se vea más de una vez entre estos fragmentos.

Por otro lado, si entre los 1000 equivalentes de genoma por ml de ADN de plasma hay un cierto porcentaje de células que comparten una célula ancestral reciente (es decir, están relacionadas entre sí por clonación), entonces se podría ver que las mutaciones de este clon están representadas preferentemente en el ADN de plasma (por ejemplo, que exhiben un perfil mutacional clonal en el plasma). Tales células clonalmente relacionadas podrían ser células de cáncer o células que están en el proceso de convertirse en un cáncer, pero que aún no lo son (es decir, pre-neoplásicas). Por lo tanto, requerir que una mutación se muestre más de una vez puede eliminar esta variación natural en las “mutaciones” identificadas en la muestra, lo cual puede dejar más mutaciones relacionadas con las células de cáncer o células pre-neoplásicas, permitiendo así la detección, especialmente la detección temprana del cáncer o condiciones precancerosas.

En una aproximación, se ha establecido que, en promedio, una mutación se acumulará en el genoma después de cada división celular. Los trabajos anteriores han mostrado que la mayor parte del ADN de plasma es de células hematopoyéticas (Lui YY et al. Clin Chem 2002: 48: 421-427). Se ha estimado que las células madre hematopoyéticas se replican una vez cada 25-50 semanas (Catlin SN, et al. Blood 2011; 117: 4460-4466). Por lo

tanto, como una aproximación simplista, un sujeto sano de 40 años de edad habría acumulado unas 40 a 80 mutaciones por células madre hematopoyéticas.

Si hay 1000 equivalentes de genoma por ml en el plasma de esta persona, y si cada una de estas células se deriva de una célula madre hematopoyética diferente, entonces podrían esperarse de 40.000 a 80.000 mutaciones entre los $4x10^{10}$ fragmentos de ADN (es decir, $4x10^{7}$ fragmentos de ADN por genoma, y 1000 equivalentes de genoma por ml de plasma). Sin embargo, como cada mutación se vería una vez, cada mutación puede seguir por debajo de un límite de detección (por ejemplo, si el valor de corte N es mayor que 1) , y por lo tanto, estas mutaciones se pueden filtrar, permitiendo de ese modo que el análisis se enfoque en las mutaciones que son más probables que resulten de afecciones cancerosas. El valor de corte puede ser cualquier valor (número entero o no entero) mayor que uno, y puede ser dinámico para diferentes loci y regiones. La profundidad de secuenciación y concentración fraccional de ADN de tumor también pueden afectar a la sensibilidad de la detección de mutaciones (por ejemplo, porcentaje de mutaciones detectables) de las células de cáncer o células pre-neoplásicas.

IV. COMPARACIÓN DEL SG DIRECTAMENTE CON EL CG

Algunas realizaciones pueden identificar posiciones de nucleótidos donde el CG es homocigoto, pero donde una especie minoritaria (es decir, el ADN de tumor) en el SG es heterocigota. Cuando se secuencia una posición con una elevada profundidad (por ejemplo, cobertura de más de 50 veces), se puede detectar si hay uno o dos alelos en esta posición en la mezcla de ADN de células sanas y cancerosas. Cuando hay dos alelos detectados, o bien (1) el CG es heterocigoto o (2) el CG es homocigoto pero el SG es heterocigoto. Estos dos escenarios se pueden diferenciar observando los recuentos relativos de los alelos mayoritarios y minoritarios. En el primer escenario, los dos alelos tendrían números similares de recuentos; pero para el último escenario, habría una gran diferencia en sus números de recuentos. Esta comparación de los recuentos de alelos relativos de las lecturas de la muestra de prueba es una realización para comparar las etiquetas de secuencia con el genoma constitucional. Los primeros loci del método 100 se pueden determinar como loci donde el número de alelos está por debajo de un umbral superior (umbral correspondiente a un polimorfismo en el CG) y por encima de un umbral inferior (umbral correspondiente a errores y mutaciones somáticas que ocurren a una velocidad suficientemente baja al no estar asociadas con una afección cancerosa). Por lo tanto, el genoma constitucional y los primeros loci pueden determinarse al mismo tiempo.

En otra realización, un proceso para identificar mutaciones puede determinar el CG primero, y luego determinar loci que tienen un número suficiente de mutaciones con relación al CG. El CG se puede determinar de una muestra constitucional que es diferente de la muestra de prueba.

La FIG. 2 muestra un diagrama de flujo de un método 200 que compara el genoma de muestra (SG) directamente con el genoma constitucional (CG) de acuerdo con las realizaciones de la presente invención. En el bloque 210, se obtiene un genoma constitucional del sujeto. El genoma constitucional puede obtenerse, por ejemplo, de una muestra tomada previamente a tiempo o una muestra constitucional que se obtiene y se analiza justo antes de que el método 200 se implemente.

En el bloque 220, una o más etiquetas de secuencia son recibidas para cada una de una pluralidad de fragmentos de ADN en una muestra biológica del sujeto. La secuenciación se puede realizar usando diversas técnicas, como se ha mencionado en la presente memoria. Las etiquetas de secuencia son una medida de lo que se cree que es la secuencia de un fragmento. Pero, una o más bases de una etiqueta de secuencia pueden ser erróneas.

En el bloque 230, al menos una porción de las etiquetas de secuencia están alineadas con el genoma constitucional. La alineación puede explicar que el CG sea heterocigoto en varios loci. La alineación no requeriría una coincidencia exacta de modo que se podrían detectar las variantes.

En el bloque 240, se identifican las etiquetas de secuencia que tienen una variante de secuencia en un locus respecto al genoma constitucional. Es posible que una etiqueta de secuencia pudiera tener más de una variante. Las variantes para cada locus y para cada etiqueta de secuencia pueden ser rastreadas. Una variante podría ser cualquier alelo que no está en el CG. Por ejemplo, el CG podría ser heterocigoto para A/T y la variante podría ser G o C.

En el bloque 250, para cada locus con una variante, un sistema informático puede contar un primer número respectivo de etiquetas de secuencia que se alinean con el locus y tienen una variante de secuencia en el locus. Por lo tanto, cada locus puede tener un recuento asociado del número de variantes vistas en el locus. Generalmente, un menor número de variantes serán vistas en un locus en comparación con las etiquetas de secuencia que corresponden al CG, por ejemplo, debido a que la concentración de ADN de tumor es menor que 50 %. Sin embargo, algunas muestras pueden tener una concentración de ADN de tumor que es mayor que 50 %.

En el bloque 260, un parámetro se determina basándose en los primeros números respectivos. En una realización, si un número respectivo es mayor que un valor de corte (por ejemplo, mayor que dos), entonces el número respectivo se puede añadir a una suma, que es el parámetro o se utiliza para determinar el parámetro. En otra realización, el

número de loci que tienen un número respectivo mayor que el valor de corte se utiliza como el parámetro.

En el bloque 270, el parámetro se compara con un valor umbral para clasificar a un nivel de cáncer. Como se describió anteriormente, el valor umbral puede determinarse a partir del análisis de las muestras de otros sujetos. Dependiendo del estado sano o con cáncer de estos otros sujetos, se puede determinar la clasificación. Por ejemplo, si los otros sujetos tenían cáncer en estadio 4, entonces si el parámetro actual estaba cerca (por ejemplo, dentro de un rango específico) al valor del parámetro obtenido de los otros sujetos, entonces el sujeto actual podría ser clasificado como que tiene cáncer de estadio 4. Sin embargo, si el parámetro supera el umbral (es decir, mayor o menor, dependiendo de cómo se define el parámetro), entonces la clasificación puede ser identificada como que es menor de estadio 4. Un análisis similar se puede hacer cuando los otros sujetos no tienen cáncer.

Se pueden utilizar múltiples umbrales para determinar la clasificación, donde cada umbral se determina de un conjunto diferente de sujetos. Cada conjunto de sujetos puede tener un nivel común de cáncer. Por lo tanto, el parámetro actual se puede comparar con los valores de cada conjunto de sujetos, lo que puede proporcionar una coincidencia con uno de los conjuntos o proporcionar un rango. Por ejemplo, el parámetro puede ser aproximadamente igual al parámetro obtenido para los sujetos que son precancerosos o en estadio 2. En otro ejemplo, el parámetro actual puede entrar en un rango que posiblemente puede coincidir con varios niveles diferentes de cáncer. Por lo tanto, la clasificación puede incluir más de un nivel de cáncer.

V. USO DEL GENOMA DE REFERENCIA

Las secuencias genómicas tanto del ADN constitucional como del ADN de la muestra biológica pueden ser comparadas con el genoma de referencia humano. Cuando hay más cambios en la muestra de plasma que en el ADN constitucional en comparación con el genoma de referencia, entonces hay una probabilidad más alta de cáncer. En una realización, se estudian los loci homocigotos en el genoma de referencia. Las cantidades de loci heterocigotos tanto en el ADN constitucional como en el ADN de la muestra biológica se comparan. Cuando la cantidad de sitios heterocigotos detectados del ADN de la muestra biológica supera a la del ADN constitucional, hay una mayor probabilidad de cáncer.

El análisis también podría limitarse a loci que son homocigotos en el GC. Las SNM también se pueden definir para loci heterocigotos, pero esto generalmente requeriría la generación de una tercera variante. En otras palabras, si el locus heterocigoto es A/T, una nueva variante sería C o G. La identificación de SNM para loci homocigotos es generalmente más fácil.

El grado de aumento en la cantidad de loci heterocigotos en el ADN de la muestra biológica respecto al ADN constitucional puede ser indicativo de cáncer o de un estado premaligno cuando se compara con la velocidad de cambio observado en sujetos sanos. Por ejemplo, si el grado de aumento de tales sitios es superior al observado en sujetos sanos en un cierto umbral, se puede considerar que los datos son indicativos de cáncer o de un estado premaligno. En una realización, la distribución de las mutaciones en sujetos sin cáncer se determina y se puede considerar un umbral como un cierto número de desviaciones estándar (por ejemplo, 2 o 3 desviaciones estándar).

Una realización puede requerir al menos un número especificado de variantes en un locus antes de que se cuente el locus. Otra realización proporciona una prueba incluso para los datos basándose en la observación de un cambio una vez. Por ejemplo, cuando el número total de variaciones (errores + mutaciones o polimorfismos auténticos) vistas en plasma es estadísticamente de manera significativa mayor que en el ADN constitucional, entonces no hay evidencia de cáncer.

La FIG. 3 muestra un diagrama de flujo de un método 300 que compara el genoma de muestra (SG) con el genoma constitucional (CG), utilizando el genoma de referencia (RG) de acuerdo con las realizaciones de la presente invención. El método 300 asume que el RG ya se ha obtenido, y que ya se han recibido las etiquetas de secuencia para la muestra biológica.

En el bloque 310, al menos una porción de las etiquetas de secuencia están alineadas con el genoma de referencia. La alineación puede permitir que se detecten malos apareamientos como variaciones. El genoma de referencia puede ser de una población similar a la del sujeto. Las etiquetas de secuencia alineadas comprenden efectivamente el genoma de muestra (SG).

En el bloque 320, se identifica un primer número (A) de variantes potenciales, por ejemplo, mutaciones de nucleótido único (SNM). Las SNM potenciales son loci donde una etiqueta de secuencia del SG muestra un nucleótido que es diferente del RG. Se pueden usar otros criterios, por ejemplo, el número de etiquetas de secuencia que muestran una variación debe ser mayor que un valor de corte y si un locus es homocigoto en el RG. El conjunto de SNM potenciales puede ser representado como conjunto A cuando los loci específicos son identificados y rastreados mediante el almacenamiento de los loci en la memoria. Los loci específicos se pueden determinar o simplemente se puede determinar un número de tales SNM.

En el bloque 330, un genoma constitucional se determina mediante la alineación de las etiquetas de secuencia obtenidas mediante la secuenciación de fragmentos de ADN de una muestra constitucional respecto a un genoma de referencia. Esta etapa podría haber sido llevada a cabo en cualquier momento con anterioridad y utilizando una muestra constitucional obtenida en cualquier momento con anterioridad. El CG simplemente se podría leer de la memoria, donde la alineación se hizo anteriormente. En una realización, la muestra constitucional podría ser células sanguíneas.

En el bloque 340, se identifican un segundo número (B) de loci donde una etiqueta de secuencia alineada del CG tiene una variante (por ejemplo, una SNM) en un locus respecto al genoma de referencia. Si un conjunto de loci se rastrea específicamente, entonces B puede representar el conjunto, en lugar de sólo un número.

En el bloque 350, el conjunto B se resta del conjunto A para identificar variantes (SNM) que están presentes en el genoma de muestra, pero no en el CG. En una realización, el conjunto de SNM puede limitarse a las posiciones de nucleótidos donde el CG es homocigoto. Para lograr esta filtración, los loci específicos donde el CG es homocigoto pueden ser identificados en el conjunto C. En otra realización, un locus no se cuenta en el primer número A o el segundo número B, si el CG no es homocigoto en el locus. En otra realización, se puede filtrar cualquier polimorfismo conocido (por ejemplo, en virtud de su presencia en una base de datos de SNP).

En una realización, la sustracción en el bloque 350 puede ser simplemente una sustracción de números, y por lo tanto las SNM potenciales específicas no se eliminan, sino que simplemente se resta un valor. En otra realización, la sustracción considera una diferencia entre el conjunto A y el conjunto B (por ejemplo, donde el conjunto B es un subconjunto del conjunto A) para identificar las SNM específicas que no están en el conjunto B. En valores lógicos, esto puede expresarse como [A Y NO (B)]. El conjunto resultante de variantes identificadas puede ser marcado como C. El parámetro puede determinarse como el número C o determinarse a partir del conjunto C.

En algunas realizaciones se puede tener en cuenta la naturaleza de las mutaciones y diferentes ponderaciones atribuidas a diferentes clases de mutaciones. Por ejemplo, las mutaciones que se asocian comúnmente con el cáncer se pueden atribuir a una ponderación mayor (también llamado un valor de importancia cuando se hace referencia a las ponderaciones relativas de loci). Tales mutaciones se pueden encontrar en las bases de datos de mutaciones asociadas con el tumor, por ejemplo, el Catálogo de Mutaciones Somáticas en Cáncer (COSMIC) (www.sanger.ac.uk/genetics /CGP/cosmic/). En otro ejemplo, las mutaciones asociadas con cambios no sinónimos se pueden atribuir a una ponderación mayor.

Por lo tanto, el primer número A podría ser determinado como una suma ponderada, donde el recuento de etiquetas que muestra una variante en un locus puede tener una ponderación diferente que el recuento de etiquetas en otro locus. El primer número A puede reflejar esta suma ponderada. Un cálculo similar se puede realizar para B, y por lo tanto el número C y el parámetro pueden reflejar esta ponderación. En otra realización, las ponderaciones se cuentan cuando se determina un conjunto C de loci específicos. Por ejemplo, una suma ponderada puede determinarse para los recuentos de los loci del conjunto C. Tales ponderaciones pueden ser utilizadas para otros métodos descritos en la presente memoria.

Por consiguiente, el parámetro que se compara con un umbral para determinar la clasificación de un nivel de cáncer puede ser el número de loci que exhiben una variación para el SG y el CG con respecto al RG. En otras realizaciones, se puede contar el número total de fragmentos de ADN (como se contó con las etiquetas de secuencia) que muestra una variación. En otras realizaciones, tales números se pueden utilizar en otra fórmula para obtener el parámetro.

En una realización, la concentración de la variante en cada locus puede ser un parámetro y se compara con un umbral. Este umbral puede ser utilizado para determinar si un locus es un locus variante potencial (además del punto de corte de un número específico de lecturas que muestra la variante), y posteriormente se contará el locus. La concentración también se podría utilizar como un factor de ponderación en una suma de las SNM.

VI. DISMINUCIÓN DE FALSOS POSITIVOS USANDO VALORES DE CORTE

Como se mencionó anteriormente, las mutaciones de nucleótido único pueden ser reconocidas en un gran número de fragmentos de ADN libres de células (por ejemplo, ADN circulante en plasma) para una región genómica grande (por ejemplo, el genoma completo) o un número de regiones genómicas para mejorar la sensibilidad del enfoque. Sin embargo, los errores analíticos, tales como errores de secuenciación pueden afectar a la viabilidad, precisión y la especificidad de este enfoque. Aquí, utilizamos la plataforma de secuenciación masivamente paralela como un ejemplo para ilustrar la importancia de los errores de secuenciación. La tasa de error de secuenciación de la plataforma Illumina de secuenciación por síntesis es de aproximadamente 0,1 % a 0,3 % por nucleótido secuenciado (Minoche et al. Genome Biol 2011, 12:R112). Cualquier plataforma de secuenciación masivamente paralela se puede usar, incluyendo una plataforma de secuenciación por ligación (por ejemplo, la plataforma de Life Technologies SOLiD), Ion Torrent/Ion Proton, la secuenciación por semiconductores, Roche 454, las plataformas de secuenciación molecular única (por ejemplo Helicos, Pacific Biosciences y Nanopore).

En un estudio previo sobre el carcinoma hepatocelular, se mostró que existen aproximadamente 3.000 mutaciones de nucleótido único para el genoma completo del cáncer (Tao Y et al. 2011 Proc Natl Acad Sci USA; 108: 12042-12047). Suponiendo que sólo el 10 % del ADN total en la circulación se deriva de las células tumorales y el ADN de plasma es secuenciado con una profundidad de secuenciación media de cobertura del genoma haploide de una vez, nos encontraríamos 9 millones ($3 \times 10^9 \times 0{,}3$ %) de variaciones de nucleótido único (SNV) debido a los errores de secuenciación. Sin embargo, la mayoría de las mutaciones de nucleótido único se espera que ocurran en sólo uno de los dos cromosomas homólogos. Con una profundidad de secuenciación de cobertura del genoma haploide de una vez de una muestra con 100 % de ADN de tumor, esperaríamos detectar sólo la mitad de las 3.000 mutaciones, es decir, 1.500 mutaciones. Cuando secuenciamos la muestra de plasma que contiene 10 % de ADN derivado de tumor a una cobertura del genoma haploide, esperaríamos detectar sólo 150 (1.500 x 10 %) mutaciones de nucleótido único asociadas con el cáncer. Por lo tanto, la relación señal-ruido para la detección de mutaciones asociadas con el cáncer es 1 en 60.000. Esta relación señal-ruido muy baja sugiere que la exactitud de la utilización de este procedimiento para diferenciar los casos normales y cancerosos sería muy baja si simplemente se usaran todos los cambios de nucleótido único en la muestra biológica (por ejemplo, plasma) como un parámetro.

Se espera que con los avances en las tecnologías de secuenciación, se produzca una reducción continua de la tasa de error de secuenciación. También se puede analizar la misma muestra utilizando más de una plataforma de secuenciación y por medio de una comparación de los resultados de secuenciación de plataforma cruzada, localizar las lecturas que puedan resultar afectadas por los errores de secuenciación. Otro enfoque es analizar dos muestras tomadas en diferentes momentos del mismo sujeto. Sin embargo, estos procedimientos llevan mucho tiempo.

En una realización, una forma de aumentar la relación señal-ruido en la detección de mutaciones de nucleótido único en el plasma de pacientes con cáncer es contar una mutación sólo si hay múltiples ocurrencias de la misma mutación en la muestra. En las plataformas de secuenciación seleccionadas, los errores de secuenciación que implican sustituciones de nucleótidos particulares pueden ser más comunes y afectarían a los resultados de la secuenciación de la muestra de prueba y la muestra de ADN constitucional tanto del sujeto de prueba como de los sujetos de control. Sin embargo, en general, los errores de secuenciación se producen aleatoriamente.

La posibilidad de tener un error de secuenciación es exponencialmente más baja cuando se observa el mismo cambio en la misma posición de nucleótido en múltiples fragmentos de ADN. Por otro lado, la posibilidad de detectar un cambio mutacional genuino asociado con el cáncer en la muestra se ve afectado por la profundidad de secuenciación y la concentración fraccional del ADN de tumor en la muestra. La posibilidad de observar la mutación en múltiples fragmentos de ADN aumentaría con la profundidad de secuenciación y la concentración fraccional del ADN de tumor. En diversas realizaciones que utilizan muestras con ADN de tumor libre de células (tal como en el plasma), la concentración fraccional puede ser 5 %, 10 %, 20 % y 30 %. En una realización, la concentración fraccional es menos de 50 %.

La FIG. 4 es una tabla 400 que muestra el número de mutaciones de nucleótido único asociadas con el cáncer identificadas correctamente usando diferente número de ocurrencias como el criterio para clasificar una mutación como que está presente en la muestra de acuerdo con las realizaciones de la presente invención. También se muestran los números de posiciones de nucleótidos que se identifican falsamente de tener una mutación debido a un error de secuenciación basándose en los mismos criterios de clasificación. Se supone que la tasa de error de secuenciación es 0,1 % (Minoche et al. Genome Bio 2011, 12:R112). La concentración fraccional de ADN derivado de tumor en la muestra se supone que es 10 %.

La FIG. 4 muestra que la relación entre el número de mutaciones asociadas con el cáncer detectadas en el plasma y el número de resultados falsos positivos aumentaría exponencialmente con el aumento del número de veces que el mismo cambio se ve en la muestra para definir una mutación, cuando la concentración fraccional de ADN derivado de tumor en la muestra se supone que es del 10 %. En otras palabras, tanto la sensibilidad como la especificidad para la detección de mutaciones del cáncer mejorarían. Además, la sensibilidad para la detección de las mutaciones asociadas con el cáncer se ve afectada por la profundidad de la secuenciación. Con cobertura del genoma haploide de 100 veces de secuenciación, se pueden detectar 2.205 (73,5 %) de las 3.000 mutaciones incluso utilizando el criterio de la ocurrencia de la mutación particular en al menos 4 fragmentos de ADN en la muestra. Se pueden utilizar otros valores para el número mínimo de fragmentos, tales como 3, 5, 8, 10, y mayor que 10.

La FIG. 5 es una tabla 500 que muestra el número esperado de loci falsos positivos y el número esperado de mutaciones identificadas cuando la concentración fraccional de ADN derivado de tumor en la muestra se supone que es del 5 %. Con una concentración fraccional inferior de ADN derivado de tumor en la muestra, se requiere una profundidad de secuenciación mayor para lograr la misma sensibilidad de la detección de las mutaciones asociadas con el cáncer. Un criterio más riguroso también sería necesario para mantener la especificidad. Por ejemplo, tendría que ser utilizado el criterio de la ocurrencia de la mutación particular en al menos 5 fragmentos de ADN, en lugar del criterio de al menos 4 ocurrencias en la situación del 10 % de fracción de ADN de tumor, en la muestra. Las Tablas 400 y 500 proporcionan guía para el valor de corte a utilizar dada la cobertura de veces y una concentración de ADN de tumor, que puede suponerse o medirse como se describe en la presente memoria.

Otra ventaja del uso de los criterios de la detección de un cambio de nucleótido único más de una vez para definir una mutación es que se espera minimizar la detección de falsos positivos debido a los cambios de nucleótido único en los tejidos no malignos. Como pueden producirse cambios de nucleótidos durante la mitosis de las células normales, cada célula sana en el cuerpo puede albergar un número de cambios de nucleótido único. Estos cambios pueden potencialmente conducir a resultados falsos positivos. Sin embargo, los cambios de una célula estarían presentes en el plasma/suero cuando la célula muere. Aunque se espera que diferentes células normales lleven diferentes conjuntos de mutaciones, las mutaciones que ocurren en una célula es poco probable que estén presentes en numerosas copias en el plasma/suero. Esto contrasta con las mutaciones dentro de las células tumorales donde se espera ver múltiples copias en el plasma/suero porque el crecimiento del tumor es de naturaleza clonal. Por lo tanto, múltiples células de un clon morirían y liberarían las mutaciones de firma representativas de los clones.

En una realización, el enriquecimiento dirigido para las regiones genómicas especificas se puede realizar antes de la secuenciación. Esta etapa de enriquecimiento dirigido puede aumentar la profundidad de secuenciación de las regiones de interés con la misma cantidad total de secuenciación realizada. En otra realización más, primero se puede realizar una ronda de secuenciación con relativamente poca profundidad de secuenciación. A continuación, las regiones que muestran al menos un cambio de nucleótido único pueden ser enriquecidas por una segunda ronda de secuenciación que tiene una cobertura de veces mayor. Después, el criterio de múltiples ocurrencias se puede aplicar para definir una mutación de los resultados de secuenciación con un enriquecimiento dirigido.

VII. PUNTOS DE CORTE DINÁMICOS

Como se describió anteriormente, se puede utilizar un valor de corte N para el número de lecturas que soportan una variante (mutación potencial) para determinar si un locus califica como una mutación (por ejemplo, una SNM) a ser contada. La utilización de un punto de corte tal puede reducir los falsos positivos. La siguiente descripción proporciona métodos para seleccionar un punto de corte para diferentes loci. En las siguientes realizaciones, se supone que hay un único clon de cáncer predominante. Un análisis similar puede llevarse a cabo para los escenarios que implican múltiples clones de células de cáncer que liberan diferentes cantidades de ADN de tumor en el plasma.

A. Número de mutaciones asociadas con el cáncer detectadas en el plasma

El número de mutaciones asociadas con el cáncer detectables en el plasma puede verse afectado por diversos parámetros, por ejemplo: (1) El número de mutaciones en el tejido de tumor (N_T) - el número total de mutaciones presentes en el tejido de tumor es el número máximo de mutaciones asociadas con el tumor detectables en el plasma del paciente; (2) La concentración fraccional de ADN derivado de tumor en plasma (f) - cuanto mayor es la concentración fraccional de ADN derivado de tumor en plasma, mayor sería la posibilidad de detectar las mutaciones asociadas con el tumor en el plasma; (3) Profundidad de secuenciación (D) - la profundidad de secuenciación se refiere al número de veces que la región secuenciada está cubierta por las lecturas de secuencia. Por ejemplo, una profundidad de secuenciación promedio de 10 veces significa que cada nucleótido dentro de la región secuenciada está cubierta en un promedio por 10 lecturas de secuencia. La posibilidad de detectar una mutación asociada con el cáncer aumentaría cuando se aumenta la profundidad de secuenciación; y (4) El número mínimo de veces que un cambio de nucleótido se detecta en el plasma para definir como una mutación potencial asociada con el cáncer (r) , que es un valor de corte utilizado para discriminar los errores de secuenciación de mutaciones reales asociadas con el cáncer.

En una aplicación se usa la distribución de Poisson para predecir el número de mutaciones asociadas con el cáncer detectadas en el plasma. Suponiendo que una mutación está presente en una posición de nucleótido en uno de los dos cromosomas homólogos, con una profundidad de secuenciación de D, el número esperado de veces que una mutación está presente en el plasma (Mp) se calcula como: Mp = D x f/2.

La probabilidad de detectar la mutación en el plasma (Pb) en un sitio de mutación particular se calcula como:

$$Pb = 1 - \sum_{i=0}^{r-1} Poisson(i, M_P)$$

donde r (valor de corte) es el número de veces que un cambio de nucleótido se ve en el plasma para definir como una mutación potencial asociada con el tumor; Poisson (i,M_P) es la probabilidad de la distribución de Poisson que tiene i ocurrencias con un número promedio de M_P.

El número total de mutaciones asociadas con el cáncer esperadas para ser detectadas en el plasma (NP) se puede calcular como: $N_P = N_T$ x Pb, donde NT es el número de mutaciones presentes en el tejido de tumor. Las siguientes gráficas muestran los porcentajes de mutaciones asociadas con el tumor esperadas para ser detectadas en el plasma utilizando diferentes criterios de ocurrencias (r) para detectar una mutación potencial y diferentes

profundidades de secuenciación.

La FIG. 6A es una gráfica 600 que muestra la tasa de detección de mutaciones asociadas con el cáncer en el plasma con 10 % y 20 % de concentraciones fraccionales en plasma de ADN derivado de tumor y utilizando cuatro y seis ocurrencias (r) como criterios para detectar las mutaciones potenciales asociadas con el cáncer. Con la misma r, una concentración fraccional mayor de ADN derivado de tumor en plasma resultaría en un mayor número de mutaciones asociadas con el cáncer detectables en el plasma. Con la misma concentración fraccional de ADN derivado de tumor en plasma, una mayor r tendría como resultado un menor número de mutaciones detectadas.

B. Número de falsos positivos únicos detectados debido a errores

Los cambios de nucleótido único en los datos de secuenciación de ADN de plasma pueden ocurrir debido a errores de secuenciación y alineación. El número de posiciones de nucleótidos con cambios de nucleótido único falsos positivos se puede predecir matemáticamente basándose en una distribución binomial. Los parámetros que afectan al número de sitios falsos positivos (N_{FP}) pueden incluir: (1) Tasa de error de secuenciación (E) – La tasa de error de secuenciación se define como la proporción de nucleótidos secuenciados que es incorrecta; (2) Profundidad de secuenciación (D) - Con una profundidad de secuenciación mayor, el número de posiciones de nucleótidos que muestran un error de secuenciación aumentaría; (3) El número mínimo de ocurrencias del mismo cambio de nucleótido para la definición de una mutación potencial asociada con el cáncer (r) ; y (4) El número total de posiciones de nucleótidos dentro de la región de interés (N_I) .

La ocurrencia de mutaciones puede considerarse en general como un proceso aleatorio. Por lo tanto, con el aumento de los criterios de ocurrencia para la definición de una mutación potencial, el número de posiciones de nucleótidos falsas positivas disminuiría exponencialmente con r. En algunas de las plataformas de secuenciación existentes, ciertos contextos de secuencia son más propensos a tener errores de secuenciación. Los ejemplos de tales contextos de secuenciación incluyen el motivo GGC, homopolímeros (por ejemplo AAAAAAA), y repeticiones simples (por ejemplo ATATATATAT). Estos contextos de secuencia aumentarán sustancialmente el cambio de nucleótido único o artefactos de inserción/deleción (Nakamura K et al. Nucleic Acids Res 2011; 39,e90 y Minoche AE et al. Genome Biol 2011; 12,R112). Además, las secuencias de repetición, tales como homopolímeros y repeticiones simples, introducirían computacionalmente ambigüedades en la alineación y, por lo tanto, conducirían a resultados falsos positivos para las variaciones de nucleótido único.

Cuanto más grande es la región de interés, mayor es el número de posiciones de nucleótidos falsas positivas. Si se está buscando mutaciones en el genoma completo, entonces la región de interés sería el genoma completo y el número de nucleótidos implicados sería de 3 mil millones. Por otro lado, si se centra en los exones, entonces el número de nucleótidos que codifican los exones, es decir, aproximadamente 45 millones, constituiría la región de interés.

El número de posiciones de nucleótidos falsas positivas asociadas con los errores de secuenciación se puede determinar basándose en los siguientes cálculos. La probabilidad (P_{Er}) de tener el mismo cambio de nucleótido en la misma posición debido a errores de secuenciación se puede calcular como:

$$P_{Er} = C(D,r)E\left(\frac{E}{3}\right)^{r-1}$$

donde *C(D, r)* es el número de combinaciones posibles para la elección de elementos r de un total de elementos D; r es el número de ocurrencias para definir una mutación potencial; D es la profundidad de secuenciación; y E es la tasa de error de secuenciación. *C(D, r)* se puede calcular como:

$$C(D,r) = \frac{\mathrm{D!}}{\mathrm{r!\,(D-r)!}}$$

El número de posiciones de nucleótidos (N_{FP}) que son falsos positivos para mutaciones se puede calcular como:

$$N_{FP} = N_I P_{Er}$$

donde N_I es el número total de posiciones de nucleótidos en la región de interés.

La FIG. 6B es una gráfica 650 que muestra el número esperado de posiciones de nucleótidos falsamente clasificadas como que tienen un cambio de nucleótido utilizando criterios de ocurrencia (r) de 4, 5, 6 y 7 frente a la profundidad de secuenciación. En este cálculo, la región de interés se supone que es el genoma completo (3 mil millones de posiciones de nucleótidos). Se supone que la tasa de error de secuenciación es 0,3 % de los nucleótidos secuenciados. Como se puede ver, el valor de r tiene un impacto significativo en los falsos positivos. Pero, como se puede ver de la FIG. 6A, un valor mayor de *r* también reduce el número de mutaciones detectadas, al menos hasta que se usen profundidades de secuenciación significativamente mayores.

C. Elección de la ocurrencia mínima (r)

Como se ha descrito anteriormente, el número de sitios verdaderos de mutación asociada con el cáncer y los sitios falsos positivos debido a errores de secuenciación aumentaría con la profundidad de secuenciación. Sin embargo, sus tasas de aumento serían diferentes. Por lo tanto, es posible utilizar la elección de la profundidad de secuencia y el valor de r para maximizar la detección de mutaciones verdaderas asociadas con el cáncer mientras se mantiene el número de sitios falsos positivos en un valor bajo.

La FIG. 7A es una gráfica 700 que muestra el número de sitios verdaderos de mutación asociada con el cáncer y sitios falsos positivos con profundidades de secuenciación de diferencia. El número total de mutaciones asociadas con el cáncer en el tejido de tumor se supone que es 3.000 y la concentración fraccional de ADN derivado de tumor en plasma se supone que es 10 %. Se supone que la tasa de error de secuenciación es 0,3 %. En la leyenda, TP denota los sitios verdaderos positivos en los que una mutación correspondiente está presente en el tejido de tumor, y FP denota sitios falsos positivos en los que ninguna mutación correspondiente está presente en el tejido del tumor y los cambios de nucleótidos presentes en los datos de secuenciación son debido a errores de secuenciación.

De la gráfica 700, a una profundidad de secuenciación de 110 veces, aproximadamente 1.410 mutaciones verdaderas asociadas con el cáncer se detectarían si usamos la ocurrencia mínima de 6 como el criterio (r = 6) para definir un sitio de mutación potencial en el plasma. Usando este criterio, sólo aproximadamente 20 sitios falsos positivos serían detectados. Si utilizamos el mínimo de 7 ocurrencias (r = 7) como criterio para definir una mutación potencial, el número de mutaciones asociadas con el cáncer que podrían ser detectadas se reduciría en 470 a aproximadamente 940. Por lo tanto, el criterio de r = 6 haría más sensible la detección de mutaciones asociadas con el cáncer en el plasma.

Por otro lado, con una profundidad de secuenciación de 200 veces, el número de mutaciones verdaderas asociadas con el cáncer detectadas sería aproximadamente 2.800 y 2.600, si utilizamos los criterios de ocurrencia mínima (r) de 6 y 7, respectivamente, para definir las mutaciones potenciales. Utilizando estos dos valores de *r*, los números de sitios falsos positivos serían aproximadamente 740 y 20, respectivamente. Por lo tanto, con una profundidad de secuenciación de 200 veces, el uso de un criterio más riguroso de r = 7 para definir una mutación potencial puede reducir en gran medida el número de sitios falsos positivos sin afectar significativamente de forma negativa la sensibilidad para detectar las mutaciones verdaderas asociadas con el cáncer.

D. Punto de corte dinámico para datos de secuenciación para la definición de mutaciones potenciales en plasma

La profundidad de secuenciación de cada nucleótido dentro de la región de interés sería diferente. Si aplicamos un valor de corte fijo para la ocurrencia de un cambio de nucleótido para definir una mutación potencial en plasma, los nucleótidos que están cubiertos por más lecturas de secuencia (es decir, una profundidad de secuenciación mayor) tendrían mayores probabilidades de ser falsamente marcados como que tienen variación de nucleótidos en ausencia de tal cambio en el tejido de tumor debido a los errores de secuenciación en comparación con los nucleótidos que tienen profundidades de secuenciación menores. Una realización para superar este problema es aplicar un valor de corte dinámico de *r* a diferentes posiciones de nucleótidos de acuerdo con la profundidad de secuenciación real de la posición de nucleótido particular, y de acuerdo con el límite superior deseado de la probabilidad de detectar variaciones falsas positivas.

En una realización, la tasa de falsos positivos máxima permisible se puede fijar en 1 en $1{,}5x10^8$ posiciones de nucleótidos. Con esta tasa de falsos positivos máxima permisible, el número total de sitios falsos positivos que se identifica en el genoma completo sería menos de 20. El valor de *r* para diferentes profundidades de secuenciación se puede determinar de acuerdo con las curvas mostradas en la FIG. 6B y estos puntos de corte se muestran en la Tabla 1. En otras realizaciones pueden utilizarse otras tasas de falsos positivos máximas permisibles diferentes, por ejemplo 1 de 3×10^8, 1 de 10^8 o 1 de 6×10^7, pueden ser utilizadas. El número total correspondiente de sitios falsos positivos sería menor de 10, 30 y 50, respectivamente.

Tabla 1. El número mínimo de apariciones de un cambio de nucleótidos presente en plasma para definir una posible mutación (r) para diferentes profundidades de secuenciación de la posición de nucleótido particular. La tasa máxima de falsos positivos se fija en 1 de $1{,}5x10^8$ nucleótidos.

Profundidad de secuenciación de una posición de nucleótido particular	Mínimo número de ocurrencia de un cambio de nucleótido presente en los datos de secuenciación del ADN de plasma para definir una mutación potencial (r)
<50	5
50 - 110	6
111 - 200	7
201 - 310	8
311 - 450	9
451 - 620	10

Profundidad de secuenciación de una posición de nucleótido particular	Mínimo número de ocurrencia de un cambio de nucleótido presente en los datos de secuenciación del ADN de plasma para definir una mutación potencial (r)
621 - 800	11

E. Secuenciación de enriquecimiento dirigido

Como se muestra en la FIG. 7A, una profundidad de secuenciación mayor puede resultar en una mejor sensibilidad para la detección de mutaciones asociadas con el cáncer mientras se mantiene bajo el número de sitios falsos positivos permitiendo el uso de un valor mayor de *r*. Por ejemplo, a una profundidad de secuenciación de 110 veces, se pueden detectar en el plasma 1.410 mutaciones verdaderas asociadas con el cáncer utilizando un valor de r de 6 mientras que el número de mutaciones verdaderas asociadas con el cáncer detectadas sería 2.600 cuando la profundidad de secuenciación aumenta hasta 200 veces y se aplica un valor *r* de 7. Los dos conjuntos de datos darían un número esperado de sitios falsos positivos de aproximadamente 20.

Aunque la secuenciación del genoma completo a una profundidad de 200 veces es relativamente cara en la actualidad, una posible manera para lograr una profundidad de secuenciación sería centrarse en una región de interés menor. El análisis de una región diana se puede lograr por ejemplo mediante, pero sin limitarse a, el uso de cebos de ADN o ARN para capturar regiones genómicas de interés por hibridación. Las regiones capturadas se extraen a continuación, por ejemplo, por medios magnéticos y después se someten a secuenciación. Dicha captura dirigida se puede realizar, por ejemplo, utilizando el sistema de enriquecimiento de dianas Agilent SureSelect, el sistema de enriquecimiento de dianas Roche Nimblegen y el sistema de resecuenciación de dianas Illumina. Otro procedimiento es realizar la amplificación por PCR de las regiones diana y luego realizar la secuenciación. En una realización, la región de interés es el exoma. En dicha realización, la captura de la diana de todos los exones se puede realizar en el ADN del plasma, y el ADN del plasma enriquecido para regiones exónicas puede ser secuenciado a continuación.

Además de tener mayor profundidad de secuenciación, el enfoque en regiones específicas en lugar del análisis del genoma completo reduciría significativamente el número de posiciones de nucleótidos en el espacio de búsqueda y llevaría a una reducción en el número de sitios falsos positivos dada la misma tasa de error de secuenciación.

La FIG. 7B es una gráfica 750 que muestra el número predicho de sitios falsos positivos que implican el análisis del genoma completo (WG) y todos los exones. Para cada tipo de análisis, se utilizan dos valores diferentes, 5 y 6, para *r*. Con una profundidad de secuenciación de 200 veces, si *r* = 5 se utiliza para definir las mutaciones en plasma, el número predicho de sitios falsos positivos es aproximadamente 23.000 y 230 para el genoma completo y todos los exones, respectivamente. Si *r* = 6 se utiliza para definir las mutaciones en el plasma, el número predicho de sitios falsos positivos es 750 y 7, respectivamente. Por lo tanto, el límite del número de nucleótidos en la región de interés puede reducir significativamente el número de falsos positivos en el análisis mutacional de plasma.

En la secuenciación de captura de exón o incluso captura de exoma se reduce el número de nucleótidos en el espacio de búsqueda. Por lo tanto, incluso si permitimos una tasa de falsos positivos mayor para la detección de mutaciones asociadas con el cáncer, el número absoluto de sitios falsos positivos se puede mantener a un nivel relativamente bajo. La asignación de una tasa de falsos positivos mayor permitiría un criterio menos estricto de ocurrencias mínimas (*r*) para definir una variación de nucleótidos única en el plasma que se utiliza. Esto daría lugar a una mayor sensibilidad para la detección de mutaciones verdaderas asociadas con el cáncer.

En una realización, se puede utilizar una tasa de falsos positivos máxima permisible de $1,5 \times 10^6$. Con esta tasa de falsos positivos, el número total de sitios falsos positivos dentro de los exones diana sería sólo 20. Los valores de *r* para diferentes profundidades de secuenciación utilizando una tasa de falsos positivos máxima permisible de $1,5 \times 10^6$ se muestran en la Tabla 2. En otras realizaciones se pueden utilizar otras tasas de falsos positivos máximas permisibles diferentes, por ejemplo 1 de 3×10^6, 1 de 10^6 o 1 de 6×10^5. El número total correspondiente de sitios falsos positivos sería menor de 10, 30 y 50, respectivamente. En una realización, diferentes clases de mutaciones se pueden atribuir a diferentes ponderaciones, como se describió anteriormente.

Tabla 2. El número mínimo de aparición de un cambio de nucleótido presente en plasma para definir una posible mutación (r) para diferentes profundidades de secuenciación de la posición de nucleótido particular. La tasa máxima de falsos positivos se fija en 1 de $1,5 \times 10^6$ nucleótidos

Profundidad de secuenciación de una posición de nucleótido particular	Mínimo número de ocurrencia de un cambio de nucleótido presente en los datos de secuenciación del ADN de plasma para definir una mutación potencial (r)
<50	4
50 - 125	5
126 - 235	6
236 - 380	7

Profundidad de secuenciación de una posición de nucleótido particular	Mínimo número de ocurrencia de un cambio de nucleótido presente en los datos de secuenciación del ADN de plasma para definir una mutación potencial (r)
381 - 560	8
561 - 760	9

VIII. DETECCIÓN DE CÁNCER

Como se mencionó anteriormente, los recuentos de las etiquetas de secuencia en loci variantes se pueden utilizar en diversas formas para determinar el parámetro, el cual se compara con un umbral para clasificar un nivel de cáncer. La concentración fraccional de las lecturas variantes respecto a todas las lecturas en un locus o muchos loci es otro parámetro que puede ser utilizado. A continuación se presentan algunos ejemplos para calcular el parámetro y el umbral.

A. Determinación del parámetro

Si el CG es homocigoto en un locus particular para un primer alelo y en la muestra biológica se observa un alelo variante (por ejemplo, plasma), entonces la concentración fraccional se puede calcular como *2p / (p+q)*, donde *p* es el número de etiquetas de secuencia que tienen el alelo variante y *q* es el número de etiquetas de secuencia que tienen el primer alelo del CG. Esta fórmula asume que sólo uno de los haplotipos del tumor tiene la variante, que sería el caso habitual. Así, para cada locus homocigoto se puede calcular una concentración fraccional. Las concentraciones fraccionales pueden promediarse. En otra realización, el recuento de *p* puede incluir el número de etiquetas de secuencia para todos los loci, y de manera similar para el recuento *q*, para determinar la concentración fraccional. A continuación se describe un ejemplo.

Se ha analizado la detección del genoma completo de variantes de nucleótido único derivadas del tumor (SNV) en el plasma de los 4 pacientes con HCC. Se ha secuenciado el ADN de tumor y ADN de la capa leucocitaria para conseguir profundidades de 29,5 veces (rango, 27 veces a 33 veces) y 43 veces (rango, 39 veces a 46 veces) la cobertura del genoma haploide, respectivamente. Los datos de MPS del ADN de tumor y el ADN de la capa leucocitaria de cada uno de los 4 pacientes con HCC se compararon, y las SNV presentes en el ADN de tumor pero no en el ADN de la capa leucocitaria se extraen con un algoritmo de bioinformática riguroso. Este algoritmo requiere una SNV putativa que esté presente en al menos un número de umbral de fragmentos de ADN de tumor secuenciados (es decir, en una etiqueta de secuencia correspondiente) antes de que pudiera ser clasificado como una verdadera SNV. El número de umbral se determina teniendo en cuenta la profundidad de la secuenciación de un nucleótido particular y la tasa de error de secuenciación, por ejemplo, como se describe en la presente memoria.

La FIG. 8 es una tabla 800 que muestra los resultados para 4 pacientes con HCC antes y después del tratamiento, incluyendo concentraciones fraccionales de ADN derivado de tumor en el plasma de acuerdo con las realizaciones de la presente invención. El número de SNV asociadas a tumores varió de 1.334 a 3.171 en los 4 casos de HCC. Las proporciones de tales SNV que fueron detectables en el plasma se enumeran antes y después del tratamiento. Antes del tratamiento, se detectaron 15 %-94 % de las SNV asociada a tumores en plasma. Después del tratamiento, el porcentaje fue entre 1,5 %-5,5 %. Por lo tanto, el número de SNV detectados se correlaciona con un nivel de cáncer. Esto demuestra que el número de SNV se puede utilizar como un parámetro para clasificar un nivel de cáncer.

Las concentraciones fraccionales de ADN derivado de tumor en plasma se determinaron por los recuentos fraccionados del mutante con respecto a las secuencias totales (es decir, de tipo mutante más silvestre). La fórmula es 2p/(p+q), donde los 2 que corresponden a sólo un haplotipo se mutaron en el tumor. Estas concentraciones fraccionales se correlacionaron bien con aquellas determinadas con el análisis de pérdida alélica agregada del genoma completo (GAAL) (Chan KC et al. Clin Chem 2013; 59: 211-24) y se redujeron después de la cirugía. Por lo tanto, también se muestra la concentración fraccional como un parámetro útil para la determinación de un nivel de cáncer.

La concentración fraccional del análisis de SNV puede comportar una carga de tumor. Un paciente con cáncer con una carga de tumor superior (por ejemplo, una concentración fraccional mayor deducida) tendrá una mayor frecuencia de mutaciones somáticas que uno con una carga de tumor inferior. Por lo tanto, las realizaciones también pueden utilizarse para el pronóstico. En general, los pacientes con cáncer con cargas tumorales mayores tienen peor pronóstico que aquellos con cargas tumorales inferiores. El primer grupo tendría así una mayor probabilidad de morir a causa de la enfermedad. En algunas realizaciones, si la concentración absoluta de ADN en una muestra biológica, por ejemplo plasma, se puede determinar (por ejemplo, utilizando PCR en tiempo real o fluorometría), entonces se puede determinar la concentración absoluta de aberraciones genéticas asociadas al tumor y utilizar para la detección clínica y/o el control y/o pronóstico.

B. Determinación de umbral

La Tabla 800 se puede utilizar para determinar un umbral. Como se mencionó anteriormente, el número de SNV y una concentración fraccional determinada por análisis de SNV se correlacionan con un nivel de cáncer. El umbral se puede determinar en una base individual. Por ejemplo, el valor de pre-tratamiento se puede utilizar para determinar el umbral. En diversas aplicaciones, el umbral podría ser un cambio relativo de un valor absoluto respecto al valor del pre-tratamiento. Un umbral adecuado podría ser una reducción en el número de SNV o concentración fraccional en un 50 %. Dicho umbral podría proporcionar una clasificación de un nivel menor de cáncer para cada uno de los casos en la Tabla 800. Hay que observar que tal umbral puede depender de la profundidad de secuenciación.

En una realización, un umbral podría ser utilizado en las muestras, y puede o no tener en cuenta los valores de pre-tratamiento para el parámetro. Por ejemplo, un umbral de 100 SNV podría ser utilizado para clasificar al sujeto como que no tiene cáncer o que tiene un bajo nivel de cáncer. Este umbral de 100 SNV es satisfecho por cada uno de los cuatro casos en la tabla 800. Si la concentración fraccional se utilizó como el parámetro, un umbral de 1,0 % clasificaría HCC1-HCC3 como prácticamente nivel cero de cáncer, y un segundo umbral de 1,5 % clasificaría HCC4 como un nivel bajo de cáncer. Por lo tanto, se puede utilizar más de un umbral para obtener más de dos clasificaciones.

Para ilustrar otros posibles umbrales, se analizó el plasma de los controles sanos para las SNV asociadas a tumores. Se pueden realizar numerosas mediciones de sujetos sanos para determinar una rango de cuántas variaciones se espera de la muestra biológica respecto al genoma constitucional.

La FIG. 9 es una tabla 900 que muestra la detección de las SNV asociadas a HCC en 16 sujetos de control sanos de acuerdo con las realizaciones de la presente invención. La Tabla 900 se puede utilizar para estimar la especificidad de un enfoque de análisis de SNV. Los 16 controles sanos aparecen como diferentes filas. Las columnas muestran las SNV detectadas para los pacientes con HCC específicos, y muestran el número de lecturas de secuencia en loci variantes que tienen el alelo variante y el número de lecturas de secuencia con el alelo de tipo silvestre (es decir, el alelo del CG). Por ejemplo, para HCC1, el control C01 tuvo 40 lecturas variantes en dichos loci variantes, pero 31.261 lecturas del alelo de tipo silvestre. La última columna muestra la concentración fraccional total a través de todas las SNV para los pacientes HCC1. Como las SNV asociadas a HCC fueron específicas para los pacientes con HCC, la presencia de las SNV asociados a HCC representan falsos positivos. Si un valor de corte, como se describe en la presente memoria, se aplica a estas variantes de la secuencia aparentes, todos estos falsos positivos, serían filtrados a distancia.

La presencia de un pequeño número de estas mutaciones asociadas con el tumor putativo en el plasma de los 16 controles sanos representa el “ruido estocástico” de este método y fue probablemente debido a errores de secuenciación. La concentración media fraccional estimada de este tipo de ruido fue de 0,38 %. Estos valores muestran un rango de sujetos sanos. Por lo tanto, un valor umbral para una clasificación de nivel cero de cáncer para HCC podría ser de aproximadamente 0,5 %, ya que la mayor concentración fraccional fue de 0,43 %. Por consiguiente, si todas las células de cáncer se eliminan de un paciente con HCC, se esperarían estas concentraciones fraccionales bajas.

Haciendo referencia de nuevo a la tabla 800, si 0,5 % se utilizó como un umbral de nivel cero de cáncer, entonces los datos de plasma post-tratamiento para HCC1 y HCC3 serían determinados por tener nivel cero basado en el análisis SNV. HCC2 podría ser clasificado como un nivel por encima de cero. HCC4 también podría ser clasificado como un nivel por encima de cero, o algún nivel superior, pero todavía un nivel relativamente bajo en comparación con las muestras de pre-tratamiento.

En una realización donde el parámetro corresponde con el número de loci variantes, el umbral podría ser cero (es decir, un locus variante podría indicar un nivel de cáncer de no cero). Sin embargo, con muchos ajustes (por ejemplo, de la profundidad), el umbral podría ser más alto, por ejemplo, un valor absoluto de 5 o 10. En una aplicación donde una persona es controlada después del tratamiento, el umbral puede ser un cierto porcentaje de SNV (identificado mediante el análisis de los tumores directamente) que aparece en la muestra. Si el valor de corte para el número de lecturas de variante requerida en un locus era lo suficientemente grande, tener sólo un loci variante podría ser indicativo de un nivel distinto de no cero de cáncer.

Por lo tanto, el análisis cuantitativo de las variaciones (por ejemplo, variaciones de nucleótidos únicos) en el ADN de una muestra biológica (por ejemplo, plasma) se puede utilizar para el diagnóstico, control y pronóstico de cáncer. Para la detección de cáncer, el número de variaciones de nucleótido único detectado en el plasma de un sujeto probado se puede comparar con aquel de un grupo de sujetos sanos. En los sujetos sanos, las variaciones de nucleótido único aparentes en el plasma pueden ser debido a errores de secuenciación, mutaciones no clonales de las células sanguíneas y otros órganos. Se ha mostrado que las células en sujetos sanos normales podrían llevar a un pequeño número de mutaciones (Conrad DF et al. Nat Genet 2011; 43:712-4), como se muestra en la Tabla 900. Por lo tanto, el número total de variaciones de nucleótidos únicos aparentes en el plasma de un grupo de sujetos aparentemente sanos puede ser utilizado como un rango de referencia para determinar si el paciente analizado tiene un número anormalmente alto de variaciones de nucleótido único en el plasma correspondientes a un nivel de no

cero de cáncer.

Los sujetos sanos utilizados para establecer el rango de referencia pueden ser equiparados al sujeto analizado en términos de edad y sexo. En un estudio anterior, se ha demostrado que el número de mutaciones en las células somáticas aumentaría con la edad (Cheung NK et al, JAMA 2012; 307: 1062-71). Por lo tanto, a medida que envejecemos, sería “normal” acumular clones de células, a pesar de que son relativamente benignas la mayoría de las veces, o se necesitaría mucho tiempo para convertirse en clínicamente significativas. En una realización, los niveles de referencia pueden ser generados por diferentes grupos de sujetos, por ejemplo diferente edad, sexo, etnia y otros parámetros (por ejemplo, historial de tabaquismo, estado de hepatitis, alcohol, historial de consumo de drogas).

El rango de referencia puede variar basándose en el valor de corte utilizado (es decir, el número de etiquetas de secuencia variantes requeridas en un locus), así como también la tasa de falsos positivos asumida y otras variables (por ejemplo, edad). Por lo tanto, el rango de referencia se puede determinar para un conjunto particular de uno o más criterios, y los mismos criterios se podrían utilizar para determinar un parámetro para una muestra. A continuación, el parámetro se puede comparar con los rangos de referencia, ya que ambos se determinaron utilizando los mismos criterios.

Como se ha mencionado anteriormente, las realizaciones pueden utilizar varios umbrales para determinar un nivel de cáncer. Por ejemplo, un primer nivel podría determinar que no hay signos de cáncer para los parámetros por debajo del umbral, y por lo menos un primer nivel de cáncer, que podría ser un nivel pre-neoplásico. Otros niveles podrían corresponder a diferentes etapas del cáncer.

C. Dependencia de las variables experimentales

La profundidad de secuenciación puede ser importante para establecer el umbral mínimo de detección del genoma de la minoría (por ejemplo, tumor). Por ejemplo, si se utiliza una profundidad de secuenciación de genomas haploides, entonces la concentración mínima de ADN de tumor que se podría detectar incluso con una tecnología de secuenciación sin ningún error es 1/5, es decir, 20 %. Por otro lado, si se utiliza una profundidad de secuenciación de 100 genomas haploides, entonces podría disminuir hasta el 2 %. Este análisis se refiere al escenario donde solamente un locus de mutación está siendo analizando. Sin embargo, cuando se analizan más loci de mutación, la concentración mínima de ADN de tumor puede ser más baja y está gobernada por una función de probabilidad binomial. Por ejemplo, si la profundidad de la secuenciación es 10 veces y la concentración fraccional de ADN de tumor es de 20 %, entonces la probabilidad de detectar la mutación es del 10 %. Sin embargo, si tenemos 10 mutaciones, entonces la probabilidad de detectar al menos una mutación sería de $1 - (1 - 10\ \%)^{10} = 65\ \%$.

Existen varios efectos relacionados con el aumento de la profundidad de secuenciación. Cuanto mayor sea la profundidad de la secuenciación, más errores de secuenciación se observarán, véanse las FIG. 4 y 5. Sin embargo, con una profundidad mayor de secuenciación, se pueden diferenciar más fácilmente los errores de secuenciación de mutaciones debido a la expansión clonal de una subpoblación de células (por ejemplo, células de cáncer) debido a que los errores de secuenciación se producirán al azar en el genoma, pero se producirían las mutaciones en la misma localización para la población dada de células.

Cuanto mayor sea la profundidad de secuenciación, más mutaciones de las “células sanas” se identificarán. Sin embargo, cuando no hay expansión clonal de estas células sanas y sus perfiles mutacionales son diferentes, entonces las mutaciones en estas células sanas pueden diferenciarse de las mutaciones por sus frecuencias de ocurrencia en el plasma (por ejemplo, usando un punto de corte N para un número requerido de lecturas que exhiben la mutación, tales como tener N igual a 2, 3, 4, 5, o mayor).

Como se mencionó anteriormente, el umbral puede depender de una cantidad de mutaciones en las células sanas que se expandieron clonalmente, y por lo tanto no pueden ser filtradas a través de otros mecanismos. Esta variación que se podría esperar se puede obtener mediante el análisis de los sujetos sanos. A medida que la expansión clonal se produce con el tiempo, la edad del paciente puede afectar a una variación que se ve en los sujetos sanos, y por lo tanto el umbral puede ser dependiente de la edad.

D. Combinación con enfoques dirigidos

En algunas realizaciones, una secuenciación aleatoria se puede utilizar en combinación con enfoques dirigidos. Por ejemplo, se puede realizar la secuenciación aleatoria de una muestra de plasma en la presentación de un paciente con cáncer. Los datos de secuenciación del ADN del plasma pueden ser analizados en cuanto a las aberraciones del número de copias y SNV. Las regiones que muestran aberraciones (por ejemplo, amplificación/deleción o alta densidad de SNV) pueden ser seleccionadas con fines de control seriado. El control se puede realizar durante un período de tiempo, o se hace inmediatamente después de la secuenciación aleatoria, efectivamente como un solo procedimiento. Para el análisis dirigido se han utilizado con éxito los enfoques de captura basados en la hibridación de fase en solución para enriquecer el ADN del plasma para el diagnóstico prenatal no invasivo (Liao GJ et al. Clin Chem 2011; 57:92-101). Tales técnicas se mencionaron anteriormente. Así, los enfoques aleatorios y objetivos

pueden ser utilizados en combinación para la detección y control del cáncer.

Por lo tanto, se podría llevar a cabo la secuenciación selectiva de los loci que se determinan que pueden ser potencialmente mutados utilizando el procedimiento no dirigido del genoma completo mencionado anteriormente. Tal secuenciación dirigida podría llevarse a cabo utilizando técnicas de hibridación en solución o de fase sólida (por ejemplo, utilizando el sistema de resecuenciación dirigida Agilent SureSelect, NimbleGen Sequence Capture, o Illumina) seguido por secuenciación masivamente paralela. Otro enfoque es realizar el sistema de amplificación (por ejemplo, basado en la PCR) para la secuenciación dirigida (Forshew T et al. Sci Transl Med 2012; 4: 135ra68).

IX. CONCENTRACIÓN FRACCIONAL

La concentración fraccional de ADN de tumor se puede utilizar para determinar el valor de corte para el número requerido de variaciones en un locus antes de que el locus sea identificado como una mutación. Por ejemplo, si la concentración fraccional era conocida por ser relativamente alta, entonces podría utilizarse un punto de corte alto para filtrar más falsos positivos, ya que se sabe que un número relativamente alto de las lecturas variantes debe existir para las SNV verdaderas. Por otro lado, si la concentración fraccional era baja, entonces podría ser necesario un punto de corte más bajo, de modo que algunas SNV no se pierdan. En este caso, la concentración fraccional sería determinada por un método diferente que el análisis de SNV, donde se utiliza como un parámetro.

Se pueden utilizar varias técnicas para determinar la concentración fraccional, algunas de las cuales se describen en la presente memoria. Estas técnicas se pueden utilizar para determinar la concentración fraccional de ADN derivado de tumor en una mezcla, por ejemplo, una muestra de biopsia que contiene una mezcla de células tumorales y células no malignas o una muestra de plasma de un paciente con cáncer que contiene el ADN liberado de las células tumorales y ADN liberado de células no malignas.

A. GAAL

La pérdida alélica agregada de genoma completo (GAAL) analiza loci que han perdido heterocigosidad (Chan KC et al. Clin Chem 2013; 59:211-24). En un sitio del genoma constitucional que es heterocigoto, un tumor a menudo tiene un locus que tiene una deleción de uno de los alelos. Así, las lecturas de secuencia para un locus mostrarán más de un alelo que otro, donde la diferencia es proporcional a la concentración fraccional de ADN de tumor en la muestra. Un ejemplo de tal cálculo se da a continuación.

El ADN extraído de la capa leucocitaria y los tejidos tumorales de los pacientes con HCC se genotipificaron con el sistema Affymetrix Genome-Wide Human SNP Array 6.0. Los datos de los microchips fueron procesados con el análisis de genotipificación Affymetrix Genotyping Console versión 4.1. y la detección de polimorfismo de nucléotido único (SNP) se realizó con el algoritmo Birdseed v2. Los datos de genotipificación para la capa leucocitaria y los tejidos tumorales fueron utilizados para la identificación de pérdida de regiones de heterocigosidad (LOH) y para realizar el análisis del número de copias. El análisis del número de copias se realizó con Genotyping Console con los parámetros por defecto de Affymetrix y con un tamaño de segmento genómico mínimo de 100 pb y un mínimo de 5 marcadores genéticos dentro del segmento.

Las regiones con LOH fueron identificadas como regiones que tienen 1 copia en el tejido tumoral y 2 copias en la capa leucocitaria, siendo los SNP dentro de estas regiones heterocigotos en la capa leucocitaria pero homocigotos en el tejido tumoral. Para una región genómica que exhibe LOH en un tejido tumoral, los alelos de SNP que estaban presentes en la capa leucocitaria, pero estaban ausentes del o con intensidad reducida en los tejidos tumorales fueron considerados como los alelos en el segmento suprimido de la región cromosómica. Los alelos que estaban presentes tanto en la capa leucocitaria como en el tejido del tumor fueron considerados como que habían sido derivados del segmento no suprimido de la región cromosómica. Para todas las regiones cromosómicas con una pérdida de copia única en el tumor, se contó el número total de lecturas de secuencias que llevan los alelos delecionados y los alelos no delecionados. Se utilizó la diferencia de estos dos valores para inferir la concentración fraccional de ADN derivado de tumor (F_{GAAL}) en la muestra utilizando la siguiente ecuación:

$$F_{GAAL} = \frac{N_{no\ -del} - N_{del}}{N_{no\ -del}}$$

donde $N_{no\ del}$ representa el número total de lecturas de secuencia que lleva los alelos no delecionados y N_{del} representa el número total de lecturas de secuencia que lleva los alelos delecionados.

B. Estimación utilizando representación genómica

Un problema con la técnica GAAL es que se identifican los loci particulares (es decir, los que exhiben LOH) y sólo se utilizan lecturas de secuencia que alinean tales loci. Tal requisito puede agregar etapas adicionales, y por lo tanto costos. Ahora se describe una realización que utiliza sólo el número de copias, por ejemplo, una densidad de lectura de secuencia.

Las aberraciones cromosómicas, por ejemplo, deleciones y amplificaciones se observan con frecuencia en los genomas del cáncer. Las aberraciones cromosómicas observadas en tejidos de cáncer generalmente involucran regiones subcromosómicas y estas aberraciones pueden ser más cortas que 1 Mb. Y, las aberraciones cromosómicas asociadas con el cáncer son heterogéneas en diferentes pacientes, y por lo tanto diferentes regiones pueden verse afectadas en diferentes pacientes. Tampoco es raro que las decenas, cientos o incluso miles de aberraciones del número de copias se encuentren en un genoma del cáncer. Todos estos factores hacen difícil determinar la concentración de ADN del tumor.

Las realizaciones implican el análisis de los cambios cuantitativos que resultan de las aberraciones cromosómicas asociadas con el tumor. En una realización, las muestras de ADN que contienen ADN derivado de células de cáncer y las células normales son secuenciadas utilizando secuenciación masivamente paralela, por ejemplo, mediante la plataforma de secuenciación Illumina HiSeq2 000. El ADN derivado puede ser ADN libre de células en el plasma u otra muestra biológica adecuada.

Las regiones cromosómicas que se amplifican en los tejidos tumorales aumentarían la probabilidad de ser secuenciadas y las regiones que se suprimen en los tejidos tumorales reducirían la probabilidad de ser secuenciadas. Por consiguiente, la densidad de las lecturas de secuencia que se alinean con las regiones amplificadas se incrementaría y las que alinean las regiones suprimidas se reduciría. El grado de variación es proporcional a la concentración fraccional del ADN derivado de tumor en la mezcla de ADN. Cuanto mayor sea la proporción de ADN del tejido tumoral, mayor sería el cambio causado por las aberraciones cromosómicas.

1. Estimación de la muestra con alta concentración de tumor

Se extrajo el ADN de los tejidos tumorales de cuatro pacientes con carcinoma hepatocelular. El ADN se fragmentó usando el sistema de sonicación de ADN Covaria y se secuenció utilizando la plataforma Illumina HiSeq2000 como se describe (Chan KC et al. Clin Chem 2013; 59:211-24). Las lecturas de secuencia fueron alineadas al genoma de referencia humano (hgl8). A continuación, el genoma se dividió en agrupaciones (regiones) de 1 Mb y la densidad de lectura de secuencia se calculó para cada agrupación después del ajuste por desviaciones respecto al GC como se describe (Chen EZ et al. PLoS One. 2011;6:e21791).

Después de que las lecturas de secuencia se hayan alineado con un genoma de referencia, se puede calcular una densidad de lectura de secuencia para varias regiones. En una realización, la densidad de lectura de secuencia es una proporción determinada como el número de lecturas asignadas a una agrupación particular (por ejemplo, región de 1 Mb) dividido por las lecturas de secuencia total que pueden alinearse con el genoma de referencia (por ejemplo, a una posición única en el genoma de referencia). Se espera que las agrupaciones que se superponen con las regiones cromosómicas amplificadas en el tejido tumoral tengan una mayor densidad de lectura de secuencia que aquellos de las agrupaciones sin tales superposiciones. Por otra parte, se espera que las agrupaciones que se superponen con las regiones cromosómicas que se suprimen tengan densidades de lectura de secuencia inferiores que aquellas sin tales superposiciones. La magnitud de la diferencia en las densidades de lecturas de secuencia entre las regiones con y sin aberraciones cromosómicas se ve afectada principalmente por la proporción de ADN derivado de tumor en la muestra y el grado de amplificación/deleción en las células tumorales.

Se pueden utilizar varios modelos estadísticos para identificar las agrupaciones que tienen densidades de lectura de secuencia correspondientes a los diferentes tipos de aberraciones cromosómicas. En una realización, se puede utilizar un modelo de mezcla normal (McLachlan G and Peel D. Multvariate normal mixtures. In finite mixture models 2004: p81-116. John Wiley & Sons Press). También se pueden utilizar otros modelos estadísticos, por ejemplo el modelo de mezcla binomial y el modelo de regresión de Poisson (McLachlan G y Peel D. Mixtures with non-normal components, Finite mixture models 2004: p135-174. John Wiley & Sons Press).

La densidad de lectura de secuencia para una agrupación se puede normalizar utilizando la densidad de lectura de secuencia de la misma agrupación como se determina a partir de la secuenciación del ADN de la capa leucocitaria. Las densidades de lectura de secuencia de diferentes agrupaciones pueden estar afectadas por el contexto de la secuencia de una región cromosómica particular, y por lo tanto la normalización puede ayudar a identificar con más precisión regiones que muestran la aberración. Por ejemplo, la posibilidad de cartografía (que se refiere a la probabilidad de la alineación de una secuencia de nuevo con su posición original) de diferentes regiones cromosómicas puede ser diferente. Además, el polimorfismo del número de copias (es decir, variaciones en el número de copias) también afectaría a las densidades de lectura de secuencia de las agrupaciones. Por lo tanto, la normalización con el ADN de la capa leucocitaria potencialmente puede reducir al mínimo las variaciones asociadas con la diferencia en el contexto de secuencia entre diferentes regiones cromosómicas.

La FIG. 10A muestra una gráfica de distribución 1000 de las densidades de lectura de secuencia de la muestra de tumor de un paciente con HCC de acuerdo con las realizaciones de la presente invención. El tejido tumoral se obtuvo después de la resección quirúrgica del paciente con HCC. El eje x representa el $\log_2$ de la relación (R) de la densidad de lecturas de secuencia entre el tejido tumoral y la capa leucocitaria del paciente. El eje y representa el número de agrupaciones.

Los picos pueden ser ajustados a la curva de distribución para representar las regiones con deleción, amplificación y sin aberraciones cromosómicas utilizando el modelo de mezcla normal. En una realización, el número de picos puede ser determinado por el criterio de información de Akaike (AIC) a través de diferentes valores plausibles. El pico central con un $\log_2 R = 0$ (es decir, R = 1) representa las regiones sin ninguna aberración cromosómica. El pico de la izquierda (respecto al central) representa regiones con pérdida de una copia. El pico de la derecha (respecto al central) representa las regiones con amplificación de una copia.

La concentración fraccional de ADN derivado de tumor puede reflejarse por la distancia entre los picos que representan las regiones amplificadas y suprimidas. Cuanto mayor es la distancia, mayor será la concentración fraccional de ADN derivado de tumor en la muestra. La concentración fraccional de ADN derivado de tumor en la muestra puede ser determinada por este procedimiento de representación genómica, denominado como F_{GR} utilizando la siguiente ecuación: $F_{GR} = R_{derecho} - R_{izquierdo}$, donde $R_{derecho}$ es el valor *R* del pico derecho y $R_{izquierdo}$ es el valor R del pico izquierdo. La mayor diferencia sería 1, que corresponde a 100 %. La concentración fraccional de ADN derivado de tumor en la muestra de tumor obtenida del paciente con HCC se estima que es 66 %, donde los valores de $R_{derecho}$ y $R_{izquierdo}$ son 1.376 y 0.712, respectivamente.

Para verificar este resultado también se utilizó otro método que utiliza el análisis de pérdida de alelo agregado de genoma completo (GAAL) para determinar de forma independiente la concentración fraccional de proporción de ADN de tumor (Chan KC et al. Clin Chem 2013; 59: 211-24). La Tabla 3 muestra las concentraciones fraccionales de ADN derivado de tumor en los tejidos de tumor de los cuatro pacientes con HCC utilizando los enfoques de representación genómica (F_{GR}) y GAAL (F_{GAAL}). Los valores determinados por estos dos enfoques diferentes concuerdan bien entre sí.

La Tabla 3 muestra la concentración fraccional determinada por GAAL y representación genómica (GR).

Tumor HCC	F_{GAAL}	F_{GR}
1	60,0%	66,5%
2	60,0%	61,4%
3	58,0%	58,9%
4	45,7%	42,2%

2. Estimación en la muestra con baja concentración de tumor

El análisis anterior ha mostrado que nuestro método de representación genómica se puede utilizar para medir la concentración fraccional de ADN de tumor cuando más de 50 % de la muestra de ADN deriva del tumor, es decir, cuando el ADN de tumor es una proporción mayoritaria. En el análisis anterior, hemos demostrado que este método también se puede aplicar a muestras en las que el ADN derivado de tumor representa una proporción menor (es decir, por debajo de 50 %). Las muestras que pueden contener una proporción menor de ADN de tumor incluyen, pero no se limitan a, la sangre, plasma, suero, orina, fluido pleural, líquido cefalorraquídeo, lágrimas, saliva, fluido ascítico y heces de pacientes con cáncer. En algunas muestras, la concentración fraccional de ADN derivado de tumor puede ser 49 %, 40 %, 30 %, 20 %, 10 %, 5 %, 2 %, 1 %, 0,5 %, 0,1 % o menor.

Para tales muestras, los picos de densidad de lectura de secuencia que representan las regiones con amplificación y deleción pueden no ser tan obvias como en las muestras que contienen una concentración relativamente alta de ADN derivado de tumor como se ilustra anteriormente. En una realización, las regiones con aberraciones cromosómicas en las células de cáncer se pueden identificar comparando con las muestras de referencia que se sabe que no contienen ADN de cáncer. Por ejemplo, el plasma de sujetos sin un cáncer se puede utilizar como referencia para determinar el rango normativo de las densidades de lectura de secuencia para las regiones cromosómicas. La densidad de lectura de secuencia del sujeto analizado puede ser comparada con el valor del grupo de referencia. En una realización se puede determinar la media y la desviación estándar (SD) de la densidad de lectura de secuencia. Para cada agrupación, la densidad de lectura de secuencia del sujeto analizado se compara con la media del grupo de referencia para determinar la puntuación z utilizando la siguiente fórmula:

$$puntuación\ z = \frac{(GR_{prueba} - \overline{GR}_{ref})}{SD_{ref}},$$

donde GR_{prueba} representa la densidad de lectura de secuencia del paciente con cáncer; $\overline{GR}_{ref}$ representa la densidad de lectura de secuencia media de los sujetos de referencia y SD_{ref} representa la SD de las densidades de lectura de secuencia para los sujetos de referencia.

Las regiones con puntuación z <-3 significa subpresentación significativa de la densidad de lectura de secuencia para una agrupación particular en el paciente con cáncer que sugiere la presencia de una deleción en el tejido de tumor. Las regiones con puntuación z > 3 significa sobrepresentación significativa de la densidad de lectura de secuencia de una agrupación particular en el paciente con cáncer que sugiere la presencia de una amplificación en el tejido de tumor.

A continuación, se puede construir la distribución de las puntuaciones z de todas las agrupaciones para identificar regiones con diferentes números de ganancia y pérdida de copias, por ejemplo, deleción de 1 o 2 copias de un cromosoma; y amplificación, dando como resultado 1, 2, 3 y 4 copias adicionales de un cromosoma. En algunos casos, más de un cromosoma o más de una región de un cromosoma pueden estar involucrados.

La FIG. 10B muestra una gráfica de distribución 1050 de las puntuaciones z para todas las agrupaciones en el plasma de un paciente con HCC de acuerdo con las realizaciones de la presente invención. Los picos (de izquierda a derecha), que representan pérdida de 1 copia, ningún cambio de copia, ganancia de 1 copia y ganancia de 2 copias se ajustan a la distribución de la puntuación z. Las regiones con diferentes tipos de aberraciones cromosómicas se pueden identificar a continuación, por ejemplo, utilizando el modelo de mezcla normal como se describió anteriormente.

La concentración fraccional de ADN de cáncer en la muestra (F) se puede inferir después a partir de las densidades de lectura de secuencia de las agrupaciones que exhiben ganancia de una copia o pérdida de una copia. La concentración fraccional determinada para una agrupación particular, se puede calcular como $F = \frac{\left| GR_{prueba} - \overline{GR}_{ref} \right| \times 2}{GR_{ref}} \times 100\%$. Esto también se puede expresar como: $F = \frac{\left| puntuación\ z \times SD_{ref} \right|}{\overline{GR}_{ref}} \times 2$,

que puede reescribirse como: F = |puntuación z\ x *CV* x 2, donde CV es el coeficiente de variación para la medición de la densidad de lectura de secuencia de los sujetos de referencia; y $CV = \frac{SD_{ref}}{\overline{GR}_{ref}}$.

En una realización, los resultados de las agrupaciones se combinan. Por ejemplo, las puntuaciones z de las agrupaciones que muestran una ganancia de 1 copia se pueden promediar o se promedian los valores F resultantes. En otra aplicación, el valor de la puntuación z utilizado para inferir F se determina mediante un modelo estadístico y está representado por los picos mostrados en la FIG. 10B y FIG. 11. Por ejemplo, la puntuación z del pico de la derecha se puede utilizar para determinar la concentración fraccional para las regiones que exhiben ganancia de 1 copia.

En otra realización, todas las agrupaciones con puntuación z <-3 y puntuación z > 3 se pueden atribuir a las regiones con pérdida de copia única y ganancia de copia única, respectivamente, debido a que estos dos tipos de aberraciones cromosómicas son los más comunes. Esta aproximación es más útil cuando el número de agrupaciones con aberraciones cromosómicas es relativamente pequeño y el ajuste de la distribución normal puede no ser exacto.

La FIG. 11 muestra una gráfica de distribución 1100 de puntuaciones z para el plasma de un paciente con HCC de acuerdo con las realizaciones de la presente invención. Aunque el número de agrupaciones superpuestas con aberraciones cromosómicas es relativamente pequeño, todas las agrupaciones con puntuación z <-3 y puntuación z > 3 se ajustan a las distribuciones normales de la pérdida de copia única y ganancia de copia única, respectivamente.

Las concentraciones fraccionales de ADN derivado de tumor en plasma de los cuatro pacientes con HCC se determinaron utilizando análisis de GAAL y este enfoque basado en GR. Los resultados se muestran en la Tabla 4. Como puede verse, la representación fraccional deducida se correlaciona bien entre el análisis de GAAL y el análisis de GR.

Tabla 4. Concentración fraccional de ADN derivado de tumor en plasma deducida por el análisis de aberraciones cromosómicas.

	Concentración fraccional de ADN derivado de tumor en plasma	
Muestras	Análisis GAAL	Análisis GR
caso11	4,3 %	4,5 %
caso13	5 %	5,5 %
caso 23	52 %	62 %
caso 27	7,6 %	6,1 %

C. Método de determinación de la concentración fraccional

La FIG. 12 es un diagrama de flujo de un método 1200 para determinar una concentración fraccional de ADN de tumor en una muestra biológica incluyendo ADN libre de células de acuerdo con las realizaciones de la presente invención. El método 1200 puede realizarse a través de diversas realizaciones, incluyendo realizaciones descritas anteriormente.

En el bloque 1210, una o más etiquetas de secuencia son recibidas para cada una de una pluralidad de fragmentos de ADN en la muestra biológica. El bloque 1210 puede realizarse como se describe en la presente memoria para otros métodos. Por ejemplo, un extremo de un fragmento de ADN puede ser secuenciado de una muestra de plasma. En otra realización, ambos extremos de un fragmento de ADN pueden ser secuenciados, permitiendo de este modo la estimación de una longitud del fragmento.

En el bloque 1220, las posiciones genómicas están determinadas por las etiquetas de secuencia. Las posiciones genómicas se pueden determinar, por ejemplo, como se describe en la presente memoria mediante la alineación de las etiquetas de secuencia con un genoma de referencia. Si ambos extremos de un fragmento son secuenciados, entonces las etiquetas apareadas pueden estar alineadas como un par restringiendo la distancia entre las dos etiquetas a menos de una distancia especificada, por ejemplo, 500 o 1.000 bases.

En el bloque 1230, para cada una de una pluralidad de regiones genómicas se determina una cantidad respectiva de fragmentos de ADN dentro de la región genómica a partir de las etiquetas de secuencia que tienen una posición genómica dentro de la región genómica. Las regiones genómicas pueden ser agrupaciones no superpuestas de longitud igual en el genoma de referencia. En una realización, se puede contar un número de etiquetas que se alinean con una agrupación. Por lo tanto, cada agrupación puede tener un número correspondiente de etiquetas alineadas. Se puede calcular un histograma que ilustre la frecuencia de las agrupaciones que tienen un cierto número de etiquetas alineadas. El método 1200 se puede realizar para regiones genómicas que tienen cada una la misma longitud (por ejemplo, agrupaciones de 1 Mb), donde las regiones no están superpuestas. En otras realizaciones se pueden utilizar diferentes longitudes, que pueden ser tenidas en cuenta y las regiones se pueden superponer.

En el bloque 124 0, la cantidad respectiva se normaliza para obtener una densidad respectiva. En una realización, la normalización de la cantidad respectiva para obtener una densidad respectiva incluye el uso de un mismo número total de etiquetas de referencia alineadas para determinar la respectiva densidad y la densidad de referencia. En otra realización, la cantidad respectiva se puede dividir por un número total de etiquetas de referencia alineadas.

En el bloque 1250, la densidad respectiva se compara con una densidad de referencia para identificar si la región genómica exhibe una pérdida de 1 copia o una ganancia de 1 copia. En una realización, la diferencia se calcula entre la densidad respectiva y la densidad de referencia (por ejemplo, como parte de la determinación de una puntuación z) y se compara con un valor de corte. En diversas realizaciones, la densidad de referencia puede obtenerse de una muestra de células sanas (por ejemplo, de la capa leucocitaria) o de las propias cantidades respectivas (por ejemplo, tomando un valor mediano o promedio, suponiendo que la mayoría de las regiones no exhiben una pérdida o una ganancia).

En el bloque 1260, se calcula una primera densidad de una o más densidades respectivas identificadas que exhiben una pérdida de 1 copia o de una o más densidades respectivas identificadas que exhiben una ganancia de 1 copia. La primera densidad puede corresponder tan solo a una región genómica, o puede ser determinada a partir de las densidades de múltiples regiones genómicas. Por ejemplo, la primera densidad puede calcularse a partir de las densidades respectivas que tienen una pérdida de 1 copia. Las densidades respectivas proporcionan una medida de la cantidad de la diferencia de densidad resultante de la deleción de la región en un tumor, dada la concentración de tumor. Del mismo modo, si la primera densidad procede de densidades respectivas que tienen una ganancia de 1 copia, entonces se puede obtener una medida de la cantidad de diferencia de densidad resultante de la duplicación de la región en un tumor. Las secciones anteriores describen varios ejemplos de cómo las densidades de múltiples regiones se pueden utilizar para determinar una densidad promedio que se utiliza para la primera densidad.

En el bloque 1270, la concentración fraccional se calcula comparando la primera densidad con otra densidad para obtener un diferencial. El diferencial se normaliza con la densidad de referencia, lo cual puede hacerse en el bloque 1270. Por ejemplo, el diferencial puede ser normalizado con la densidad de referencia dividiendo el diferencial por la densidad de referencia. En otra realización, el diferencial puede ser normalizado en los bloques anteriores.

En una aplicación, la otra densidad es la densidad de referencia, por ejemplo, como en la sección 2 anterior. Por lo tanto, el cálculo de la concentración fraccional puede incluir multiplicar el diferencial por dos. En otra aplicación, la otra densidad es una segunda densidad calculada a partir de las densidades respectivas identificadas que exhiben una pérdida de 1 copia (donde la primera densidad se calcula utilizando densidades respectivas identificadas que exhiben una ganancia de 1 copia), por ejemplo, como se describe en la sección 1 anterior. En este caso, el diferencial normalizado se puede determinar calculando una primera relación (por ejemplo, $R_{derecho}$) de la primera densidad y la densidad de referencia y calculando una segunda relación ($R_{izquierdo}$) de la segunda densidad y la densidad de referencia, donde el diferencial está entre la primera relación y la segunda relación. Como se describió anteriormente, la identificación de región genómica que exhibe una pérdida de 1 copia o una ganancia de 1 copia se puede realizar ajustando los picos a una curva de distribución de un histograma de las densidades respectivas.

En resumen, las realizaciones pueden analizar la representación genómica de ADN de plasma en diferentes regiones cromosómicas para determinar simultáneamente si la región cromosómica se amplifica o suprime en el tejido de tumor y, si la región se amplifica o suprime, para utilizar su representación genómica para deducir la

concentración fraccional del ADN derivado de tumor. Algunas aplicaciones utilizan un modelo de mezcla normal para analizar la distribución general de la representación genómica de diferentes regiones para determinar la representación genómica asociada con diferentes tipos de aberraciones, principalmente las ganancias de 1, 2, 3 o 4 copias y las pérdidas de 1 o 2 copias.

Las realizaciones tienen varias ventajas sobre otros métodos, por ejemplo el procedimiento de pérdida alélica agregada de genoma completo (GAAL) (solicitud de patente de Estados Unidos 13/308.473; Chan KC et al. Clin Chem 2013; 59:211-24) y el análisis de mutaciones de nucleótido único asociadas con el tumor (Forshew T et al. Sci Transl Med 2012;4:136ra68). Toda la cartografía de lecturas de secuencia con regiones con aberraciones cromosómicas se puede utilizar para determinar la densidad de lectura de secuencia de la región y, por lo tanto, son informativos con respecto a la concentración fraccional de ADN de tumor. Por otro lado, en el análisis de GAAL, sólo serían informativas las lecturas de secuencia que cubren los nucleótidos únicos que son heterocigotos en el individuo y localizados dentro de una región cromosómica con la ganancia o pérdida de cromosoma. Del mismo modo, para el análisis de mutaciones asociadas con el cáncer, sólo las lecturas de secuencia que cubren las mutaciones serían útiles para la deducción de la concentración de ADN de tumor. Por lo tanto, las realizaciones pueden permitir un uso más rentable de los datos de secuenciación, ya que lecturas de secuenciación relativamente menores pueden ser necesarias para lograr el mismo grado de precisión en la estimación de la concentración fraccional de ADN derivado de tumor en comparación con otros enfoques.

X. METODOLOGIAS ALTERNATIVAS

Aparte de utilizar el número de veces que se observa una mutación en una etiqueta de secuencia como criterio para identificar un locus que es una mutación verdadera (ajustando de ese modo el valor predictivo positivo), se podrían emplear otras técnicas en lugar de o además de utilizar un valor de corte para proporcionar un valor predictivo mayor en la identificación de una mutación cancerosa. Por ejemplo, se podrían utilizar filtros de bioinformática de diferente rigurosidad al procesar los datos de secuenciación, por ejemplo, teniendo en cuenta la puntuación de calidad de un nucleótido secuenciado. En una realización, se podrían utilizar secuenciadores de ADN y procesos químicos de secuenciación con diferentes perfiles de error de secuenciación. Los secuenciadores y los procesos químicos con tasas de error de secuenciación menores darían unos valores predictivos positivos mayores. También se puede utilizar la secuenciación repetida del mismo fragmento de ADN para aumentar la precisión de secuenciación. Una estrategia posible es la estrategia de secuenciación de consenso circular de Pacific Biosciences.

En otra realización, se podría incorporar información del tamaño de los fragmentos secuenciados en la interpretación de los datos. Ya que el ADN derivado de tumor es más corto que el ADN derivado de tumor en plasma (véase Solicitud de Patente de Estados Unidos N.º 13/308.473), el valor predictivo positivo de un fragmento de ADN de plasma más corto que contiene una mutación potencial derivada de tumor será mayor que el de un fragmento de ADN de plasma más largo. Los datos del tamaño se obtendrán fácilmente si se realiza la secuenciación de extremo apareado del ADN de plasma. Como alternativa, se podrían utilizar secuenciadores de ADN con longitudes de lectura largas, produciendo de este modo la longitud completa de un fragmento de ADN de plasma. También se podría realizar el fraccionamiento de tamaño de la muestra de ADN de plasma antes de la secuenciación del ADN. Los ejemplos de métodos que se podrían utilizar para el fraccionamiento de tamaño incluyen electroforesis en gel, el uso del procedimiento de microfluído (por ejemplo, el sistema de Caliper LabChip XT) y columnas de centrifugación de exclusión molecular.

En otra realización más, se esperaría que la concentración fraccional de mutaciones asociadas con el tumor en plasma en un paciente con cáncer no hematológico aumente si uno se centra en los fragmentos de ADN más cortos en el plasma. En una aplicación, se puede comparar la concentración fraccional de mutaciones asociadas con el tumor en plasma en fragmentos de ADN de dos o más distribuciones de tamaño diferentes. Un paciente con un cáncer no hematológico tendrá concentraciones fraccionales mayores de mutaciones asociadas con el tumor en los fragmentos más cortos en comparación con los fragmentos más grandes.

En algunas realizaciones, se podrían combinar los resultados de secuenciación de dos o más alícuotas de la misma muestra de sangre, o de dos o más muestras de sangre tomadas en las mismas ocasiones o en diferentes ocasiones. Las mutaciones potenciales observadas en más de una alícuota o muestras tendrían un valor predictivo positivo mayor de mutaciones asociadas con el tumor. El valor predictivo positivo aumentaría con el número de muestras que muestran tal mutación. Las mutaciones potenciales que están presentes en muestras de plasma tomadas en diferentes puntos temporales pueden ser consideradas como mutaciones potenciales.

XI. EJEMPLOS

Los siguientes son ejemplos de técnicas y datos, que no deben considerarse limitativos en las realizaciones de la presente invención.

A. Materiales y Métodos

En cuanto a la recolección de muestras, se reclutaron pacientes con carcinoma hepatocelular (HCC), portadores de hepatitis B crónica, y un paciente con cáncer de mama y ovario sincrónico. Todos los pacientes con CHC tenían enfermedad en estadio A1 según el sistema Barcelona-Clinic Liver Cancer. Las muestras de sangre periférica de todos los participantes se recolectaron en tubos que contienen EDTA. Los tejidos de tumor de los pacientes con HCC se obtuvieron durante sus cirugías de resección del cáncer.

Las muestras de sangre periférica se centrifugaron 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 y luego se almacenó a 80 °C. Las moléculas de ADN libre de células de 4,8 ml de plasma se extrajeron según el protocolo de fluido sanguíneo y corporal del QIAamp DSP DNABlood Mini Kit (Qiagen). El ADN de plasma se concentró con un concentrador SpeedVac (Savant DNA120; Thermo Scientific) en un volumen final de 4 µl por caso para su posterior preparación de la biblioteca de secuenciación de ADN.

El ADN genómico se extrajo de muestras de capa leucocitaria de los pacientes de acuerdo con el protocolo de fluido sanguíneo y corporal del QIAamp DSP DNA Blood Mini Kit. Se extrajo el ADN de tejidos de tumor con el QIAamp DNA Mini Kit (Qiagen).

Las bibliotecas de secuenciación de las muestras de ADN genómico se construyeron con el Paired-End Sample Preparation Kit (Illumina) según las instrucciones del fabricante. Resumiendo, en primer lugar se cizallaron con un ultrasonicador focalizado Covaris S220 1-5 microgramos de ADN genómico en fragmentos de 200 pb. Después, las moléculas de ADN fueron reparadas en sus extremos con ADN polimerasa de T4 y polimerasa de Klenow; la polinucleótido quinasa de T4 se utilizó a continuación para fosforilar los extremos 5'. Se creó un saliente 3' con un fragmento de Klenow deficiente de exonucleasa 3' a 5'. Los oligonucleótidos adaptadores Illumina se ligaron a los extremos cohesivos. El ADN ligado al adaptador se enriqueció con una PCR de 12 ciclos. Debido a que las moléculas de ADN de plasma eran fragmentos cortos y las cantidades de ADN total en las muestras de plasma eran relativamente pequeñas, se omitieron las etapas de fragmentación y se utilizó una PCR de 15 ciclos para construir las bibliotecas de ADN de las muestras de plasma.

Se utilizó un Bioanalizador Agilent 2100 (Agilent Technologies) para verificar la calidad y el tamaño de las bibliotecas de ADN ligado al adaptador. Las bibliotecas de ADN se midieron a continuación con un KAPA Library Quantification Kit (Kapa Biosystems) de acuerdo con las instrucciones del fabricante. La biblioteca de ADN se diluyó y se hibridó con las células de flujo de secuenciación de extremo apareado. Los agrupamientos de ADN se generaron en un sistema de generación de agrupamiento cBot (Illumina) con el TruSeq PE Cluster Generation Kit v2 (Illumina), seguido por 51_2 ciclos o 76_2 ciclos de secuenciación en un sistema HiSeq 2000 (Illumina) con el Kit TruSeq SBS v2 (Illumina).

Los datos de secuenciación de extremo apareado se analizaron por medio del Programa de Alineación de Oligonucleótidos Cortos 2 (SOAP2) en el modo de extremo apareado. Para cada lectura de extremo apareado, se alinean 50 pb o 75 pb de cada extremo con el genoma humano de referencia sin enmascaramiento de repetición (hgl8). Se permitieron hasta 2 mal apareamientos de nucleótidos para la alineación de cada extremo. Las coordenadas genómicas de estas alineaciones potenciales para los 2 extremos se analizaron después para determinar si alguna combinación permitiría que los 2 extremos estuviesen alineados con el mismo cromosoma con la orientación correcta, abarcando un tamaño de inserto de menor que o igual a 600 pb, y cartografiando una ubicación única en el genoma humano de referencia. Las lecturas duplicadas fueron definidas como lecturas de extremo apareado en las que la molécula de ADN de inserto mostró ubicaciones de inicio y fin idénticas en el genoma humano; las lecturas duplicadas se eliminaron como se ha descrito previamente (Lo et al. Sci Transl Med 2010; 2:61ra91).

En algunas realizaciones, se secuenciaron las muestras de ADN constitucional y de tumor apareadas para identificar las variantes de nucleótido único asociadas con el tumor (SNV). En algunas aplicaciones, nos centramos en las SNV que ocurren en los sitios homocigotos en el ADN constitucional (en este ejemplo es el ADN de capa leucocitaria). En principio, cualquier variación de nucleótido detectada en los datos de secuenciación de los tejidos de tumor pero ausente en el ADN constitucional podría ser una mutación potencial (es decir, una SNV). Sin embargo, debido a los errores de secuenciación (0,1 % - 0,3 % de los nucleótidos secuenciados), se identificarían millones de falsos positivos en el genoma si una ocurrencia única de cualquier cambio de nucleótido en los datos de secuenciación del tejido de tumor se considera como una SNV asociada con el tumor. Una forma de reducir el número de falsos positivos sería instituir el criterio de observar múltiples ocurrencias del mismo cambio de nucleótido en los datos de secuenciación en el tejido de tumor antes de que analizar una SNV asociada con el tumor.

Debido a que la ocurrencia de errores de secuenciación es un proceso estocástico, el número de falsos positivos debido a errores de secuenciación se reduciría de forma exponencial con el aumento del número de ocurrencias necesarias para que una SNV observada sea calificada como SNV asociada con el tumor. Por otro lado, el número de falsos positivos aumentaría con el aumento de la profundidad de secuenciación. Estas relaciones se podrían predecir con funciones de distribución binomial y de Poisson. Las realizaciones pueden determinar un punto de corte

dinámico de ocurrencia para calificar una SNV observada como asociada con el tumor. Las realizaciones pueden tener en cuenta la cobertura real del nucleótido particular en los datos de secuenciación de tumor, la tasa de error de secuenciación, la tasa de falsos positivos máxima permisible, y la sensibilidad deseada para la detección de mutación.

En algunos ejemplos, establecemos criterios muy estrictos para reducir los falsos positivos. Por ejemplo, se puede requerir que una mutación esté completamente ausente en la secuenciación de ADN constitucional, y la profundidad de secuenciación para la posición de nucleótido particular tiene que ser 20 veces. En algunas aplicaciones, el punto de corte de ocurrencia logra una velocidad de detección de falsos positivos de menos de 10^{-7}. En algunos ejemplos, también se filtraron SNV que estaban dentro de las regiones centroméricas, teloméricas, y de baja complejidad para minimizar los falsos positivos debido a los artefactos de alineación. Además, también se eliminó la cartografía de SNV putativa con SNP conocidos en las 135 bases de datos

B. Resección antes y después

La FIG. 13A muestra una tabla 1300 del análisis de mutaciones en el plasma de la paciente con cáncer de ovario y cáncer de mama en el momento del diagnóstico de acuerdo con las realizaciones de la presente invención. Aquí, mostramos un ejemplo de una paciente con cáncer de ovario bilateral y un cáncer de mama. Los datos de secuenciación del plasma se compararon con los resultados de la secuenciación del ADN constitucional de la paciente (capa leucocitaria). Los cambios de nucleótido único que estaban presentes en el plasma, pero no en el ADN constitucional se consideraron como mutaciones potenciales. Los cánceres de ovario en el lado derecho e izquierdo de la paciente se muestrearon cada uno en dos sitios, lo que hace un total de cuatro muestras de tumor. Las mutaciones de tumor fueron mutaciones detectadas en los cuatro tejidos de tumor de ovario en cuatro sitios diferentes.

Se detectaron más de 3,6 millones de cambios de nucleótido único en el plasma durante al menos una vez por secuenciación. De estos cambios, sólo 2.064 también se detectaron en los tejidos de tumor, lo que da un valor de predicción positiva de 0,06 %. Utilizando el criterio de que se detectan al menos dos veces en el plasma, el número de mutaciones potenciales se redujo significativamente en un 99,5 % hasta 18.885. El número de mutaciones de tumor sólo se redujo en un 3 % hasta 2.003, y el valor de predicción positiva aumentó hasta el 11 %.

Usando los criterios de detección de al menos cinco veces en plasma, se detectaron sólo 2.572 mutaciones potenciales y entre éstas, 1.814 eran mutaciones detectadas en todos los tejidos de tumor, dando, por lo tanto, un valor predictivo positivo de 71 %. Otros criterios para determinar el número de ocurrencias (por ejemplo, 2, 3, 4, 6, 7, 8, 9, 10, etc.) se pueden utilizar para definir las mutaciones potenciales dependiendo de la sensibilidad y valor predictivo positivo requerido. Cuanto mayor sea el número de ocurrencias utilizadas como criterio, mayor será el valor predictivo positivo con una reducción en la sensibilidad.

La FIG. 13B muestra una tabla 1350 del análisis de mutaciones en el plasma de la paciente con cáncer de ovario bilateral y un cáncer de mama después de la resección del tumor de acuerdo con las realizaciones de la presente invención. Se realizó la resección quirúrgica de los cánceres de la paciente. Se obtuvo una muestra de sangre un día después de la resección de los tumores de ovario y el cáncer de mama. El ADN de plasma se secuenció a continuación. Para este ejemplo, sólo se analizaron las mutaciones de los cánceres de ovario. Más de 3 millones de mutaciones potenciales se detectaron al menos una vez en una muestra de plasma. Sin embargo, utilizando el criterio de tener al menos cinco ocurrencias, el número de mutaciones potenciales se redujo a 238. Se observó una reducción significativa en comparación con el número de mutaciones potenciales para la muestra tomada en el diagnóstico y utilizando el mismo criterio de cinco mutaciones.

En una realización, el número de cambios de nucleótido único detectados en el plasma se puede utilizar como un parámetro para la detección, control y pronóstico de un paciente con cáncer. Se puede usar diferente número de ocurrencias como criterio para lograr la sensibilidad y especificidad deseadas. Se espera que un paciente con una carga de tumor mayor y, por lo tanto, un peor pronóstico tenga una carga mutacional en el plasma mayor.

Para este análisis, se podría establecer el perfil de carga mutacional para diferentes tipos de cáncer. Con fines de control, se vería que se reduciría la carga mutacional en el plasma de un paciente que responde al tratamiento. Si el tumor ha recurrido, por ejemplo, durante una recidiva, entonces se espera que la carga mutacional aumente. Tal control permitiría controlar la eficacia de la modalidad seleccionada de tratamiento para un paciente y detectar la ocurrencia de resistencia a un tratamiento particular.

Mediante el análisis de las mutaciones específicas que se podrían ver en los resultados de la secuenciación del ADN de plasma, se podría también identificar dianas que podrían predecir la sensibilidad (por ejemplo, las mutaciones en el gen del receptor del factor de crecimiento epidérmico y la respuesta al tratamiento con inhibidor de la tirosina quinasa) y la resistencia al tratamiento dirigido particular (por ejemplo, mutaciones *KRAS* en el cáncer colorrectal y la resistencia al tratamiento con panitumumab y cetuximab), y podrían guiar la planificación de los regímenes de tratamiento.

El ejemplo anterior se refería al cáncer de ovario bilateral. También se podría realizar el mismo análisis en las mutaciones del cáncer de mama, pudiendo entonces rastrear las mutaciones de ambos tipos de cáncer en el plasma. También se puede utilizar una estrategia similar para rastrear las mutaciones de un cáncer primario y su metástasis o metástasis.

Las realizaciones serían útiles para la detección de cáncer en sujetos aparentemente sanos o en sujetos con factores de riesgo particulares (por ejemplo, historial de tabaquismo, estado viral (tales como portadores del virus de la hepatitis, sujetos infectados con virus del papiloma humano). La carga mutacional que se podría ver en el plasma de tales sujetos sería indicativa del riesgo de que el sujeto desarrolle cáncer sintomático en un plazo determinado. Por lo tanto, se esperaría que los sujetos con una mayor carga mutacional en el plasma tengan un mayor riesgo que aquellos con una carga mutacional menor. Además, el perfil temporal de tal carga mutacional en plasma también sería un potente indicador de riesgo. Por ejemplo, si un sujeto tiene una carga mutacional en plasma determinada cada año y si las cargas mutacionales están aumentando progresivamente, entonces este sujeto debe ser derivado para modalidades de detección adicionales del cáncer, por ejemplo, radiografía torácica, ecografía, tomografía computarizada, resonancia magnética o tomografía por emisión de positrones.

C. Puntos de corte dinámicos para deducir las mutaciones de secuenciación en plasma

Para este estudio se reclutaron cuatro pacientes con carcinoma hepatocelular (HCC) y una paciente con cáncer de ovario y mama. En esta última paciente, nos centramos en el análisis del cáncer de ovario. Las muestras de sangre se recolectaron de cada paciente antes y después de la resección quirúrgica de los tumores. También se recolectaron los tejidos de tumor resecados. El ADN extraído del tejido de tumor, los leucocitos de la muestra de sangre preoperatoria y las muestras de plasma pre- y post-operatorias se secuenciaron utilizando el sistema de secuenciación HiSeq2000 (Illumina). Los datos de secuenciación se alinearon con la secuencia del genoma humano de referencia (hg18) utilizando el Paquete de Análisis de Oligonucleótidos Cortos 2 (SOAP2) (Li R et al. Bioinformatics 2009; 25: 1966-1967). Las secuencias de ADN de los leucocitos fueron consideradas como secuencia de ADN constitucional para cada sujeto de estudio.

En este ejemplo, las SNM asociadas con el tumor se dedujeron primero de los datos de secuenciación de ADN de plasma y el CG sin referencia a los tejidos de tumor. A continuación, los resultados deducidos del plasma se compararon con los datos de secuenciación generados de los tejidos de tumor (como datos referencia) para determinar la exactitud de los resultados deducidos. En este sentido, los datos de referencia se obtuvieron comparando los datos de secuenciación de los tejidos de tumor y la secuencia constitucional para determinar las mutaciones en los tejidos de tumor. En este análisis, nos centramos en las posiciones de nucleótidos en las que el ADN constitucional del sujeto estudiado era homocigoto.

1. Análisis del genoma completo no dirigido

Las profundidades de secuenciación para los leucocitos, los tejidos de tumor y el ADN de plasma de cada paciente se muestran en la Tabla 5.

Tabla 5. Mediana de las profundidades de secuenciación de diferentes muestras de los cuatro casos de CHC.

Caso	Mediana de la profundidad de secuenciación (veces)			
	Leucocitos	Tejido tumoral	Plasma preoperatorio	Plasma postoperatorio
HCC1	39	29	23	24
HCC2	39	29	25	28
HCC3	46	33	18	21
HCC4	46	27	20	23
Paciente con cáncer de ovario	44	53	37	28

Los puntos de corte dinámicos para las ocurrencias mínimas para la definición de las mutaciones de plasma (*r*) como se muestra en la tabla 1 se utilizan para identificar las mutaciones en el plasma de cada paciente. Como la profundidad de secuenciación de cada locus puede variar, el punto de corte puede variar, lo que proporciona eficazmente una dependencia del punto de corte en el número total de lecturas para un locus. Por ejemplo, aunque la mediana de la profundidad es menor de 50 (Tabla 5), la profundidad de secuenciación de loci individuales puede variar mucho y ser cubierta > 100 veces.

Además de los errores de secuenciación, otra fuente de error serían los errores de alineación. Para minimizar este tipo de errores, la lecturas de secuenciación que llevan una mutación fueron realineadas con el genoma de referencia utilizando el programa de alineación Bowtie (Langmead B et al. Genome Biol 2009, 10:R25). Sólo las lecturas que podían ser alineadas con una posición única del genoma de referencia por SOAP2 y Bowtie se utilizaron para el análisis aguas abajo para las mutaciones de plasma. También se podrían utilizar otras combinaciones de paquetes de software de alineación basados en diferentes algoritmos.

Para minimizar adicionalmente los errores de secuenciación y alineación en los datos de secuenciación reales, se aplicaron dos algoritmos de filtración adicionales para determinar las posiciones de nucleótidos que mostraron variaciones de nucleótido único en las lecturas de secuencia: (1) ≥ 70 % de las lecturas de secuencia que llevan la mutación podrían ser realineadas con la misma coordenada genómica utilizando Bowtie con calidad de cartografía ≥ Q20 (es decir, probabilidad de mala alineación < 1 %); (2) ≥ 70 % de las lecturas de secuencia que llevan la mutación no estaban dentro de las 5 pb de ambos extremos (es decir, extremos 5' y 3') de la lecturas de secuencia. Esta regla de filtración fue implantada porque los errores de secuenciación eran más prevalentes en ambos extremos de una lectura de secuencia.

También se investigaron los factores que afectan a la deducción de un tumor sin previo conocimiento del genoma de tumor. Tal parámetro fue la concentración fraccional de ADN derivado de tumor en plasma. Este parámetro podría ser considerado como otro parámetro de referencia y se dedujo como una referencia conociendo previamente el genoma de tumor utilizando GAAL.

La Tabla 6 muestra variaciones de nucleótido detectadas en el plasma antes y durante el tratamiento. Para HCC1, sin conocimiento previo del genoma del tumor, se detectaron un total de 961 variaciones de nucleótido único. Entre estas variaciones de nucleótido detectadas en el plasma, 828 eran mutaciones asociadas con el cáncer. Después de la resección quirúrgica de HCC, el número total de variaciones de nucleótido se redujo a 43 y ninguna de ellas fueron mutaciones asociadas con el cáncer.

Con fines de referencia, la concentración fraccional de ADN derivado de tumor en la muestra de plasma post-operatorio fue 53 % y se dedujo con el conocimiento previo del genoma del tumor. Para HCC2, HCC3 y HCC4, sin conocimiento previo de los genomas del tumor, los números de variaciones de nucleótido único en el plasma se dedujeron, variando desde 27 hasta 32 para las muestras de plasma preoperatorio. Estos resultados son compatibles con la predicción matemática de que, con una profundidad de secuenciación de aproximadamente 20 veces, un porcentaje muy bajo de mutaciones asociadas con el cáncer podría ser detectado en el plasma y la mayoría de las variaciones de secuencia detectadas en el plasma fueron debido a errores de secuenciación. Después de la resección del tumor, no hubo cambio significativo en el número de variaciones de la secuencia detectadas. Con fines de referencia, se dedujo que las concentraciones fraccionales de ADN derivado de tumor en plasma variaban desde 2,1 % hasta 5 % y se dedujeron con el conocimiento previo de los genomas del tumor.

Tabla 6. Variaciones de nucleótido detectadas en el plasma.

	Plasma preoperatorio			Plasma postoperatorio		
	Concentración fraccional de ADN derivado de tumor	N.º total de variaciones de nucleótido único	N.º de mutaciones asociadas con el cáncer identificadas	Concentración fraccional de ADN derivado de tumor	N.º total de variaciones de nucleótido único	N.º de mutaciones asociadas con el cáncer identificadas
HCC1	53%	961	828	0,4%	43	0
HCC2	5%	32	0	0,6%	49	0
HCC3	2,1%	29	0	0,2%	32	0
HCC4	2,6%	27	0	1,3%	35	1
Paciente con cáncer de ovario (y de mama)	46%	1718	1502	0,2%	2	0

2. Enriquecimiento de objetivo de los exones

Como se ha descrito anteriormente, el aumento de la profundidad de secuenciación para la región de interés puede aumentar tanto la sensibilidad como la especificidad para identificar mutaciones asociadas con el cáncer en plasma y, por lo tanto, aumentando la potencia de discriminación entre los pacientes con cáncer y sujetos sin cáncer. Aunque el aumento de la profundidad de secuenciación del genoma completo es todavía muy costoso, una alternativa es enriquecer ciertas regiones para la secuenciación. En una realización, los exones seleccionados o de hecho todo el exoma pueden ser enriquecidos de forma dirigida para la secuenciación. Este procedimiento puede aumentar significativamente la profundidad de secuenciación de la región diana sin aumentar la cantidad total de lecturas de secuencia.

Las bibliotecas de secuenciación del ADN de plasma de los pacientes con HCC y el paciente con cáncer de ovario (y mama) fueron capturadas utilizando el kit Agilent SureSelect All Exon para el enriquecimiento de diana del exoma. Las bibliotecas de secuenciación enriquecidas con exón se secuenciaron después utilizando el sistema de secuenciación HiSeq 2000. Las lecturas de secuencia se alinearon con el genoma de referencia humano (hgl8). Después de la alineación, las lecturas de secuencia cartografiadas únicamente para los exones se analizaron para determinar las variaciones de nucleótido único. Para la identificación de las variaciones de nucleótido único en

plasma para el análisis de captura de exoma, se utilizan los valores de corte dinámico que se muestran en la tabla 2.

La FIG. 14A es una tabla 1400 que muestra la detección de variaciones de nucleótido único en ADN de plasma para HCC1. Sin conocimiento previo del genoma del tumor, se dedujo de los datos de secuenciación de diana un total de 57 variaciones de nucleótido único en plasma. En la validación posterior de los datos de secuenciación obtenidos de los tejidos de tumor, 55 resultaron ser mutaciones verdaderas asociadas con el tumor. Como se ha discutido antes, la concentración fraccional de ADN derivado de tumor en plasma pre-operatorio fue 53 %. Después de la resección del tumor, no se detectaron variaciones de nucleótido único en los datos de secuenciación de diana obtenidos del plasma. Estos resultados indican que el análisis cuantitativo del número de variaciones de nucleótido único en plasma se puede utilizar para controlar la progresión de la enfermedad de los pacientes con cáncer.

La FIG. 14B es una tabla 1450 que muestra la detección de variaciones de nucleótido único en ADN de plasma para HCC2. Sin el conocimiento previo del genoma del tumor, se dedujo de los datos de secuenciación de diana en el plasma un total de 18 variaciones de nucleótido único. Todas estas mutaciones se encontraron en los tejidos de tumor. Como se ha descrito anteriormente, la concentración fraccional de ADN derivado de tumor en plasma pre-operatorio fue 5 %. Después de la resección del tumor, no se detectaron variaciones de nucleótido único en el plasma. En comparación con HCC1 que tuvo una concentración fraccional mayor de ADN derivado de tumor en plasma, se detectaron pocas variaciones de nucleótido único en el plasma del caso HCC2. Estos resultados sugieren que el número de variaciones de nucleótido único en plasma se puede utilizar como un parámetro para reflejar la concentración fraccional de ADN derivado de tumor en plasma y, por lo tanto, la carga de tumor en el paciente, ya que se ha demostrado que la concentración de ADN derivado de tumor en plasma se correlaciona positivamente con la carga de tumor (Chan KC et al. Clin Chem 2005; 51:2192-5).

La FIG. 15A es una tabla 1500 que muestra la detección de variaciones de nucleótido único en el ADN de plasma para HCC3. Sin conocimiento previo del genoma del tumor, no se observó a partir de los datos de secuenciación de diana ninguna variación de nucleótido único ni en las muestras de plasma de pre-resección ni post-resección. Esto es probable que sea debido a una concentración fraccional relativamente baja (2,1 %) de ADN derivado de tumor en plasma en este paciente. El aumento adicional de la profundidad de secuenciación se predice para mejorar la sensibilidad para la detección de mutaciones asociadas con el cáncer en los casos con baja concentración fraccional de ADN derivado de tumor.

La FIG. 15B es una tabla 1550 que muestra la detección de variaciones de nucleótido único en el ADN de plasma para HCC4. Sin el conocimiento previo del genoma del tumor, se dedujo a partir de los datos de secuenciación de diana del plasma un total de 3 variaciones de nucleótido único. Todas estas mutaciones se encontraron en los tejidos de tumor. En comparación con HCC1 y HCC2 que tenían mayores concentraciones fraccionales de ADN derivado de tumor en plasma, se detectaron pocas variaciones de nucleótido único en el plasma del caso HCC4, el cual tenía un ADN de tumor fraccional en plasma de 2,6 %. Estos resultados sugieren que el número de variaciones de nucleótido único en plasma se puede utilizar como un parámetro para reflejar la concentración fraccional de ADN derivado de tumor en plasma y la carga tumoral en un paciente.

La FIG. 16 es una tabla 1600 que muestra la detección de variaciones de nucleótido único en el ADN de plasma para la paciente con cáncer de ovario (y mama). Sin conocimiento previo del genoma tumoral, se dedujo a partir de los datos de secuenciación de diana del plasma un total de 64 variaciones de nucleótido único. Entre estas, 59 se detectaron en los tejidos de tumor de ovario. La concentración fraccional estimada de ADN derivado de tumor de ovario en el plasma fue 46 %. Se detectó una reducción significativa en el número total de variaciones de nucleótido único en plasma después de la resección del cáncer de ovario.

Además de la utilización del sistema de enriquecimiento de diana SureSelect (Agilent), también se utilizó el sistema de enriquecimiento de diana Nimblegen SeqCap EZ Exome+UTR (Roche) para enriquecer las secuencias de exones para la secuenciación. El sistema Nimblegen SeqCap cubre las regiones de exón del genoma, así como la región no traducida 5' y 3'. Se analizaron las muestras de plasma pretratamiento de los cuatro pacientes con HCC, dos sujetos de control sanos y dos portadores de hepatitis B crónica sin cáncer (Tabla 7). En otras realizaciones se pueden utilizar otros sistemas de enriquecimiento de diana, incluyendo, pero sin limitarse a aquellos que utilizan la hibridación en fase de solución o en fase sólida.

Tabla 7. Resultados de secuenciación del exoma para los cuatro pacientes con HCC (HCC1-4) utilizando el sistema de enriquecimiento de diana Nimblegen SeqCap EZ Exome+UTR de la captura de secuencia. El análisis de secuenciación del plasma pre-tratamiento de HCC3 fue sub-óptimo debido a un mayor porcentaje de lecturas duplicadas de PCR.

	Plasma pretratamiento			Plasma postratamiento	
	Concentración fraccional de ADN derivado de tumor en plasma por análisis de GAAL	N.º de variación de secuencia detectada en plasma que cumple los puntos de corte dinámicos	N.º de variación de secuencia que se superpone con las mutaciones detectadas en el tejido de tumor correspondiente	N.º de variación de secuencia detectada en plasma que cumple los puntos de corte dinámicos	N.º de variación de secuencia que se superpone con las mutaciones detectadas en el tejido de tumor correspondiente
HCC1	53%	69	64	1	1
HCC2	5%	51	47	3	0
HCC3	2,1%	0	0	1	0
HCC4	2,6%	8	7	0	0

En los dos portadores de hepatitis B crónica y los dos sujetos de control sanos, se detectaron una o menos variaciones de nucleótido único que cumplían los criterios de punto de corte dinámico (Tabla 8). En tres de los cuatro pacientes con HCC, el número de variaciones de secuencia detectadas en el plasma que cumplía con el requisito de punto de corte dinámico fue al menos 8. En HCC3, no se detectó SNV que cumpliese con el punto de corte dinámico. En esta muestra, hubo una alta proporción de lectura duplicada de PCR en las lecturas secuenciadas, lo que conduce a un menor número de lecturas secuenciadas no duplicadas. Se observó una reducción marcada de SNV detectables en plasma después de la resección quirúrgica del tumor.

Tabla 8. Resultados de secuenciación del exoma de 2 portadores de hepatitis B crónica (HBV1 y HBV2) y 2 sujetos de control sanos (Ctrll y Ctrl2) utilizando el sistema de enriquecimiento de diana Nimblegen SeqCap EZ Exome+UTR para la captura de secuencia.

	N.º de variación de secuencia detectada en plasma que cumple con los puntos de corte dinámicos
HBV1	0
HBV2	1
Ctrl1	1
Ctrl2	1

XII. HETEROGENEIDAD DEL TUMOR

La cuantificación de mutaciones de nucleótido único en una muestra biológica (por ejemplo, plasma/suero) también es útil para el análisis de la heterogeneidad del tumor, tanto la heterogeneidad intra-tumoral como inter-tumoral. La heterogeneidad intra-tumoral se refiere a la existencia de múltiples clones de células tumorales dentro del mismo tumor. La heterogeneidad inter-tumoral se refiere a la existencia de múltiples clones de células tumorales para dos o más tumores del mismo tipo histológico, pero presentes en diferentes sitios (ya sea en los mismos órganos, o en diferentes órganos). En ciertos tipos de tumores, la existencia de heterogeneidad de tumor es un indicador de mal pronóstico (Yoon HH et al. J Clin Oncol 2012; 30: 3932 a 3938; Merlo LMF et al. Cancer Prev Res 2010; 3: 1388-1397). En ciertos tipos de tumores, cuanto mayor sea el grado de heterogeneidad del tumor, mayor sería la posibilidad de progresión del tumor o el desarrollo de clones resistentes después del tratamiento dirigido.

Aunque se cree que los cánceres surgen de la expansión clonal de una célula tumoral, el crecimiento y la evolución de un cáncer darían lugar a la acumulación de nuevas y diferentes mutaciones en diferentes partes de un cáncer. Por ejemplo, cuando un paciente con cáncer desarrolla metástasis, el tumor situado en el órgano original y el tumor metastásico compartirían un número de mutaciones. Sin embargo, las células de cáncer de los dos sitios también serían portadoras de un conjunto único de mutaciones que están ausentes en el otro sitio del tumor. Se espera que las mutaciones que son compartidas por los dos sitios estén presentes a concentraciones mayores que aquellas mutaciones que sólo se observan en un sitio de tumor.

A. Ejemplo

Se analizó el plasma sanguíneo de una paciente que tenía cáncer de ovario bilateral y cáncer de mama. Ambos tumores de ovario eran del tipo adenocarcinoma seroso. El de la izquierda medía 6 cm y el de la derecha medía 12 cm en la dimensión más larga. También hubo múltiples lesiones metastásicas en el colon y el omentum. El ADN extraído de los leucocitos se secuenció utilizando la plataforma de secuenciación por síntesis de Illumina hasta un promedio de cobertura de genoma haploide de 44 veces. Las localizaciones de nucleótidos que muestran sólo un alelo, es decir homocigotos, fueron analizadas adicionalmente para determinar mutaciones de nucleótido único en

plasma.

El ADN se extrajo de cuatro sitios diferentes de los tumores izquierdo y derecho y se secuenció utilizando la plataforma de secuenciación de Illumina. Dos sitios (sitios A y B) eran del tumor derecho y los otros dos sitios (sitios C y D) eran del tumor izquierdo. Los sitios A y B tenían un tamaño de aproximadamente 4 cm. La distancia entre los sitios C y D también era de aproximadamente 4 cm. Las muestras de plasma se recogieron de la paciente antes y después de la resección quirúrgica de los tumores de ovario. El ADN se extrajo a continuación del plasma de la paciente. La profundidad de secuenciación del tumor de los sitios A, B, C y D, así como las muestras de plasma, se muestran en la tabla 9.

Tabla 9. Profundidad de secuenciación del tumor de los sitios A, B, C y D.

Muestra	N.º lecturas de secuenciación sin procesar	N.º de lecturas alineadas	Veces de cobertura de genoma haploide
ADN constitucional de capa leucocitaria	1.091.250.072	876.269.922	43,81
Tumor de ovario derecho (sitio A)	1.374.495.256	1.067.277.229	53,36
Tumor de ovario derecho (sitio B)	934.518.588	803.007.464	40,15
Tumor de ovario izquierdo (sitio C)	1.313.051.122	1.036.643.946	51,83
Tumor de ovario izquierdo (sitio D)	1.159.091.833	974.823.207	48,74
Muestra de plasma recogida antes de la cirugía	988.697.457	741.982.535	37,10
Muestra de plasma recogida después de la cirugía	957.295.879	564.623.127	28,23

En el ejemplo actual, para definir una mutación única de nucleótido único asociada con el tumor, la localización de los nucleótidos es secuenciada al menos 20 veces en el tejido de tumor y 30 veces en el ADN constitucional. En otras realizaciones se pueden usar otras profundidades de secuenciación, por ejemplo, 35, 40, 45, 50, 60, 70, 80, 90, 100 y > 100 veces. La reducción de los costos de secuenciación permitiría que se realizaran profundidades mayores mucho más fácilmente. La posición de nucleótidos es homocigota en el ADN constitucional mientras que se observa un cambio de nucleótido en el tejido de tumor. El criterio para la ocurrencia del cambio de nucleótido en el tejido de tumor depende de la profundidad de secuenciación total de la posición de nucleótido particular en el tejido de tumor. Para la cobertura de nucleótidos de 20 a 30 veces, la ocurrencia de cambio de nucleótido (valor de corte) es al menos cinco veces. Para la cobertura de 31 a 50 veces, la ocurrencia del cambio de nucleótido es al menos seis veces. Para la cobertura de 51 a 70 veces, la ocurrencia necesaria es al menos siete veces. Estos criterios se derivan de la predicción de la sensibilidad de la detección de las mutaciones verdaderas y el número esperado de loci falsos positivos utilizando la distribución de Poisson.

La FIG. 17 es una tabla 1700 que muestra las sensibilidades predichas de diferentes requisitos de ocurrencia y profundidades de secuenciación. La sensibilidad correspondería con el número de mutaciones verdaderas detectadas a una profundidad de veces particular, utilizando un punto de corte particular. Cuanto mayor es la profundidad de secuenciación, más probable es que se detecte una mutación con un punto de corte dado, ya que se obtendrán más lecturas de secuencia de mutación. Para puntos de corte mayores, es menos probable que se detecte un mutante, ya que los criterios son más estrictos.

La FIG. 18 es una tabla 1800 que muestra los números predichos de loci falsos positivos para diferentes puntos de corte y diferentes profundidades de secuenciación. El número de falsos positivos aumenta al aumentar la profundidad de secuenciación, ya que se obtienen más lecturas de secuencia. Sin embargo, no se predicen falsos positivos para un punto de corte de cinco o más, incluso hasta una profundidad de secuenciación de 70. En otras realizaciones se pueden usar diferentes criterios de ocurrencia para lograr la sensibilidad y especificidad deseadas.

La FIG. 19 muestra un diagrama de árbol que ilustra el número de mutaciones detectadas en los diferentes sitios de tumor. Las mutaciones se determinaron mediante la secuenciación directa de los tumores. El sitio A tiene 71 mutaciones que son específicas de ese tumor, y el sitio B tiene 122 mutaciones específicas del sitio, a pesar de que estaban sólo a 4 cm de distancia. Se observaron 10 mutaciones en ambos sitios A y B. El sitio C tiene 168 mutaciones que son específicas de ese tumor, y el sitio D tiene 248 mutaciones específicas del sitio, a pesar de que estaban sólo a 4 cm de distancia. Se observaron 12 mutaciones en ambos sitios C y D. Existe una heterogeneidad significativa en los perfiles mutacionales para los diferentes sitios de tumor. Por ejemplo, sólo se detectaron 248 mutaciones en el sitio del tumor D pero no se detectaron en los otros tres sitios del tumor. Un total de 2.129

mutaciones se observaron en todos los sitios. Por lo tanto, muchas mutaciones eran compartidas entre los diferentes tumores. Por lo tanto, había siete grupos de SNV. No hubo diferencias observables entre estas cuatro regiones en términos de aberraciones del número de copias.

La FIG. 20 es una tabla 2000 que muestra el número de fragmentos que llevan las mutaciones derivadas de tumor en la muestra de plasma pre-tratamiento y post-tratamiento. También se muestran las concentraciones fraccionales inferidas del ADN derivado de tumor que llevan las mutaciones respectivas. La categoría de mutación se refiere a los sitios del tumor donde se detectaron las mutaciones. Por ejemplo, las mutaciones de categoría A se refieren a las mutaciones que están sólo presentes en el sitio A mientras que las mutaciones de categoría ABCD se refieren a las mutaciones presentes en los cuatro sitios de tumor.

De las 2.129 mutaciones que estaban presentes en los cuatro sitios de tumor, 2.105 (98,9 %) fueron detectables en al menos un fragmento de ADN de plasma. Por otro lado, para las 609 mutaciones que estaban presentes en sólo uno de los cuatro sitios de tumor, sólo 77 (12,6 %) fueron detectables en al menos un fragmento de ADN de plasma. Por lo tanto, la cuantificación de mutaciones de nucleótido único en el plasma se puede utilizar para reflejar la abundancia relativa de estas mutaciones en los tejidos de tumor. Esta información sería útil para el estudio de la heterogeneidad del cáncer. En este ejemplo, se consideró una mutación potencial cuando se observaba una vez en los datos de secuenciación.

Las concentraciones fraccionales de ADN de tumor circulante se determinaron con cada grupo de SNV. Las concentraciones fraccionales de ADN de tumor en el plasma antes de la cirugía y después de la cirugía, como se determina por SNV compartidas por las 4 regiones (es decir, grupo ABCD), fueron 46 % y 0,18 %, respectivamente. Estos últimos porcentajes se correlacionan bien con los obtenidos en el análisis de GAAL, 46 % y 0,66 %. Las mutaciones compartidas por las 4 regiones (es decir, grupo ABCD) aportaron la mayor contribución fraccional de ADN derivado de tumor al plasma.

Las concentraciones fraccionales de ADN derivado de tumor en plasma preoperatorio determinadas con SNV de los grupos AB y CD fueron 9,5 % y 1,1 %, respectivamente. Estas concentraciones fueron consistentes con los tamaños relativos de los tumores de ovario derecho e izquierdo. Las concentraciones fraccionales de ADN derivado de tumor determinadas con las SNV de región única (es decir, aquellas en los grupos A, B, C y D) fueron generalmente bajas. Estos datos sugieren que para una medición exacta de la carga tumoral total en un paciente con cáncer, el uso de un enfoque de disparo de pistola del genoma completo podría proporcionar una imagen más representativa, en comparación con el enfoque más tradicional de seleccionar las mutaciones asociadas con el tumor específicas. En el último enfoque, si sólo un subconjunto de las células tumorales posee las mutaciones dirigidas, se podría perder información importante con respecto a la inminente recidiva o progresión de la enfermedad causada por las células tumorales que no poseen las mutaciones seleccionadas, o se podría perder la aparición de un clon resistente al tratamiento.

La FIG. 21 es una gráfica 2100 que muestra las distribuciones de ocurrencia en el plasma de las mutaciones detectadas en un sitio de tumor único y mutaciones detectadas en los cuatro sitios de tumor. La gráfica de barras 2100 muestra los datos para dos tipos de mutación: (1) mutaciones detectadas en un sólo sitio y (2) mutaciones detectadas en los cuatro sitios de tumor. El eje horizontal es el número de veces que se detecta una mutación en el plasma. El eje vertical muestra el porcentaje de mutaciones que corresponde a un valor particular en el eje horizontal. Por ejemplo, aproximadamente 88 % de las mutaciones de tipo (1) mostradas sólo una vez en el plasma. Como se puede ver, las mutaciones que se observaron en un sitio, se detectaron la mayoría de las veces una vez, y no más de cuatro veces. Las mutaciones presentes en un sitio de tumor único se detectaron con mucha menos frecuencia en el plasma en comparación con las mutaciones presentes en los cuatro sitios de tumor.

Una aplicación de esta tecnología sería permitir a los médicos estimar la carga de las células tumorales que llevan las diferentes clases de mutaciones. Una proporción de estas mutaciones podría ser potencialmente tratable con fármacos específicos. Se esperaría que los agentes dirigidos a las mutaciones presentes en una mayor proporción de células tumorales tengan unos efectos terapéuticos más prominentes.

La FIG. 22 es una gráfica 2200 que muestra la distribución predicha de ocurrencia en el plasma de las mutaciones procedentes de un tumor heterogéneo. El tumor contiene dos grupos de mutaciones. Un grupo de mutaciones está presente en todas las células tumorales y el otro grupo de mutaciones sólo está presente en ¼ de las células tumorales, basándose en la aproximación de que dos sitios son representativos de cada tumor de ovario. La concentración fraccional total de ADN derivado de tumor en plasma se supone que es 40 %. Se supone que la muestra de plasma es secuenciada a una profundidad promedio de 50 veces por posición del nucleótido. De acuerdo con esta distribución predicha de ocurrencia en el plasma, las mutaciones que están presentes en todos los tejidos de tumor pueden ser diferenciadas de las mutaciones sólo presentes en ¼ de las células tumorales por su ocurrencia en el plasma. Por ejemplo, la ocurrencia de 6 veces puede ser utilizada como un punto de corte. Para las mutaciones presentes en todas las células tumorales, el 92,3 % de las mutaciones estarían presentes en el plasma al menos 6 veces. Por el contrario, para las mutaciones que están presentes en ¼ de las células tumorales, sólo 12,4 % de las mutaciones estarían presentes en el plasma al menos 6 veces.

La FIG. 23 es una tabla 2300 que demuestra la especificidad de las realizaciones para 16 sujetos de control sanos. Sus muestras de ADN de plasma se secuenciaron a una cobertura media de 30 veces. La detección de las mutaciones presentes en el plasma de la paciente con cáncer de ovario anterior se realizó en las muestras de plasma de estos sujetos sanos. Las mutaciones presentes en el tumor de la paciente con cáncer de ovario se detectaron con muy poca frecuencia en los datos de secuenciación del plasma de los sujetos de control sanos y ninguna de la categoría de mutaciones tuvo una concentración fraccional aparente de > 1 %. Estos resultados muestran que este método de detección es altamente específico.

B. Método

La FIG. 24 es un diagrama de flujo de un método 2400 para el análisis de una heterogeneidad de uno o más tumores de un sujeto de acuerdo con las realizaciones de la presente invención. Ciertas etapas del método 2400 se pueden realizar como se describe en la presente memoria,

En el bloque 2410, se obtiene un genoma constitucional del sujeto. En el bloque 2420, una o más etiquetas de secuencia se reciben para cada una de una pluralidad de fragmentos de ADN en una muestra biológica del sujeto, donde la muestra biológica incluye ADN libre de células. En el bloque 2430, las posiciones genómicas se determinan por las etiquetas de secuencia. En el bloque 2440, las etiquetas de secuencia se comparan con el genoma constitucional para determinar un primer número de primeros loci. En cada uno de los primeros loci, un número de las etiquetas de secuencia que tienen una variante de secuencia respecto al genoma constitucional está por encima de un valor de corte, donde el valor de corte es mayor que uno.

En el bloque 2450, una medida de la heterogeneidad de uno o más tumores se calcula basándose en los primeros números respectivos del conjunto de primeras ubicaciones genómicas. En un aspecto, las medidas pueden proporcionar un valor que representa un número de mutaciones que son compartidas por tumores respecto a un número de mutaciones que no son compartidas por los tumores. Aquí, varios tumores pueden existir como un objeto único, con diferentes tumores dentro del objeto, que pueden representar lo que normalmente se llama heterogeneidad intra-tumoral. La medida también puede referirse a si algunas mutaciones están en uno o unos cuantos tumores en comparación con las mutaciones que están en muchos o la mayoría de los tumores. Se puede calcular más de una medida de heterogeneidad.

En el bloque 2460, la medida de la heterogeneidad se puede comparar con un valor umbral para determinar una clasificación de un nivel de heterogeneidad. Una o más medidas se pueden utilizar de varias maneras. Por ejemplo, una o más medidas de heterogeneidad se pueden utilizar para predecir la posibilidad de progresión del tumor. En algunos tumores, cuanta más heterogeneidad mayor es la posibilidad de progresión y mayor es la posibilidad de aparición de un clon resistente después del tratamiento (por ejemplo tratamiento dirigido).

C. Medidas de heterogeneidad del tumor

Un ejemplo de una medida de heterogeneidad es el número de 'bandas de concentración' de los diferentes grupos de mutaciones en el plasma. Por ejemplo, si hay dos clones de tumor predominantes dentro de un paciente, y si estos clones están presentes en diferentes concentraciones, entonces se esperaría ver dos mutaciones diferentes con diferentes concentraciones en el plasma. Estos diferentes valores se pueden calcular mediante la determinación de la concentración fraccional para diferentes conjuntos de mutaciones, donde cada conjunto corresponde a uno de los tumores.

Cada una de estas concentraciones se puede denominar una 'banda de concentración' o 'clase de concentración'. Si un paciente tiene más clones, entonces se verán más bandas/clases de concentración. Por lo tanto, cuantas más bandas, más heterogénea. El número de bandas de concentración se puede ver trazando las concentraciones fraccionales para diversas mutaciones. Se puede crear un histograma para las distintas concentraciones, donde diferentes picos corresponden a diferentes tumores (o diferentes clones de un tumor). Un pico grande corresponderá probablemente a mutaciones que son compartidas por todos o algunos tumores (o clones de un tumor). Estos picos pueden ser analizados para determinar qué picos más pequeños se combinan para determinar un pico más grande. Se puede usar un procedimiento de ajuste, por ejemplo, similar al procedimiento apropiado para las FIGS. 10B y 11.

En una aplicación, el histograma es una gráfica donde el eje Y es la cantidad (por ejemplo, número o proporción) de loci y el eje x es la concentración fraccional. Las mutaciones que son compartidas por todos o algunos tumores darían como resultado una concentración fraccional mayor. El tamaño de pico representaría la cantidad de loci que dan lugar a una concentración fraccional particular. El tamaño relativo de los picos a concentración baja y alta reflejaría el grado de heterogeneidad de los tumores (o clones de un tumor). Un pico mayor a la concentración alta refleja que la mayoría de las mutaciones son compartidas por la mayoría o todos los tumores (o clones de un tumor) e indica un menor grado de heterogeneidad del tumor. Si el pico a la concentración baja es mayor, entonces la mayoría de las mutaciones son compartidas por unos cuantos tumores (o unos cuantos clones de un tumor). Esto indicaría un mayor grado de heterogeneidad del tumor.

Cuantos más picos existan, más mutaciones específicas del sitio hay. Cada pico puede corresponder a un conjunto diferente de mutaciones, donde el conjunto de mutaciones son de un subconjunto de los tumores (por ejemplo, sólo uno o dos tumores - como se ilustra anteriormente). Para el ejemplo de la FIG. 19, puede haber un total de 7 picos, teniendo los 4 picos de un sitio único probablemente la concentración más pequeña (dependiendo del tamaño relativo de los tumores), dos picos para los sitios AB y sitios CD, y un pico para las mutaciones compartidas por todos los sitios.

La ubicación de los picos también puede proporcionar un tamaño relativo de los tumores. Una mayor concentración se correlacionaría con un tumor más grande, ya que un tumor más grande liberaría más ADN de tumor en la muestra, por ejemplo, en plasma. Por lo tanto, se podría estimar la carga de células tumorales que llevan las diferentes clases de mutaciones.

Otro ejemplo de una medida de heterogeneidad es la proporción de sitios de mutación que tienen relativamente pocas lecturas de variantes (por ejemplo, 4, 5, o 6) en comparación con la proporción de lecturas de mutación que tienen lecturas de variante relativamente altas (por ejemplo, 9-13). Haciendo referencia de nuevo a la FIG. 22, se puede ver que las mutaciones de sitio específico tienen menos lecturas de variante (lo cual también tiene como resultado una concentración fraccional menor). Las mutaciones compartidas tienen más lecturas de variante (lo cual también tiene como resultado una concentración fraccional más grande). Una relación de una primera proporción a 6 (recuento menor) dividida por una segunda proporción a 10 (recuento mayor) lleva una medida de heterogeneidad. Si la relación es pequeña, entonces hay pocas mutaciones que son específicas del sitio, y por lo tanto el nivel de heterogeneidad es bajo. Si la relación es grande (o al menos más grande que los valores calibrados de las muestras conocidas), entonces el nivel de heterogeneidad es mayor.

D. Determinación de umbrales

Los valores de umbral se pueden determinar de sujetos cuyos tumores son sometidos a biopsia (por ejemplo, como se describe anteriormente) para determinar directamente un nivel de heterogeneidad. El nivel se puede definir de varias maneras, tales como relaciones entre mutaciones específicas del sitio y mutaciones compartidas. Las muestras biológicas (por ejemplo, muestras de plasma) se pueden analizar a continuación para determinar las medidas de heterogeneidad, donde una medida de heterogeneidad de las muestras biológicas puede estar asociada con el nivel de heterogeneidad determinado por el análisis de las células de los tumores directamente.

Tal procedimiento puede proporcionar una calibración de umbrales con relación a los niveles de heterogeneidad. Si la medida de heterogeneidad de la prueba cae entre dos umbrales, entonces el nivel de heterogeneidad se puede estimar que está entre los niveles correspondientes a los umbrales.

En una realización, se puede calcular una curva de calibración entre los niveles de heterogeneidad determinados a partir de las biopsias y la medida de la heterogeneidad correspondiente determinada de la muestra de plasma (u otra muestra). En dicho ejemplo, los niveles de heterogeneidad son numéricos, donde estos niveles numéricos pueden corresponder a diferentes clasificaciones. Diferentes rangos de niveles numéricos pueden corresponder a diferentes diagnósticos, por ejemplo, diferentes estadios del cáncer.

E. Método que utiliza la concentración fraccional de la representación genómica

La heterogeneidad del tumor también puede ser analizada utilizando la concentración fraccional, por ejemplo, como se determina utilizando realizaciones del método 1200. Las regiones genómicas que exhiben pérdida de una copia pueden provenir de diferentes tumores. Por lo tanto, la concentración fraccional determinada para diversas regiones del genoma puede ser diferente dependiendo de si existe la amplificación (o deleción para la pérdida de 1 copia) sólo en un tumor o múltiples tumores. Por lo tanto, las mismas medidas de heterogeneidad se pueden utilizar para las concentraciones fraccionales determinadas mediante las realizaciones del método 1200.

Por ejemplo, una región genómica puede ser identificada como correspondiente a una pérdida de 1 copia, y una concentración fraccional puede ser determinada sólo a partir de una densidad respectiva en cada región genómica (la densidad respectiva podría ser utilizada como una concentración fraccional). Se puede crear un histograma de las diversas densidades respectivas contando el número de regiones que tienen diferentes densidades. Si sólo un tumor o un clon tumoral o un depósito tumoral tiene una ganancia en una región particular, entonces la densidad de esa región sería menor que la densidad en una región que tenía una ganancia en múltiples tumores o múltiples clones de tumor o múltiples depósitos de tumor (es decir, la concentración fraccional de ADN de tumor en la región compartida sería mayor que la región específica del sitio). Las medidas de heterogeneidad descritas anteriormente por lo tanto se pueden aplicar a picos identificados usando la ganancia o pérdida de números de copias en diversas regiones, al igual que la concentración fraccional de diferentes sitios mostraba una distribución de concentraciones fraccionales.

En una aplicación, si las densidades respectivas se utilizan para el histograma, se obtendrían ganancias y pérdidas separadas. Las regiones que muestran una ganancia podrían ser analizadas por separado mediante la creación de un histograma sólo para las ganancias, y crear un histograma aparte sólo para las pérdidas. Si se utiliza la

concentración fraccional, entonces los picos de pérdidas y ganancias pueden ser analizados juntos. Por ejemplo, las concentraciones fraccionales utilizan una diferencia (por ejemplo, como un valor absoluto) respecto a la densidad de referencia, y por lo tanto las concentraciones fraccionales para las ganancias y pérdidas pueden contribuir al mismo pico.

XIII. SISTEMA INFORMÁTICO

Cualquiera de los sistemas informáticos mencionados en la presente memoria puede utilizar cualquier número adecuado de subsistemas. Los ejemplos de tales subsistemas se muestran en la FIG. 25 en un aparato informático 2500. En algunas realizaciones, un sistema informático incluye un aparato informático único, donde los subsistemas pueden ser los componentes del aparato informático. En otras realizaciones, un sistema informático puede incluir múltiples aparatos informáticos, siendo cada uno un subsistema, con componentes internos.

Los subsistemas mostrados en la FIG. 25 están interconectados a través de un bus de sistema 2575. Se muestran los subsistemas adicionales como una impresora 2574, teclado 2578, disco fijo 2579, monitor 2576, que se acoplan al adaptador de pantalla 2582, y otros. Los dispositivos periféricos y de entrada/salida (I/O), que se acoplan al controlador I/O 2571, se pueden conectar al sistema informático a través de cualquier número de medios conocidos en la técnica, tales como puerto en serie 2577. Por ejemplo, el puerto en serie 2577 o interfaz externa 2581 (por ejemplo, Ethernet, Wi-Fi, etc.) se puede utilizar para conectar el sistema informático 2500 a una red de área amplia tal como Internet, un dispositivo de entrada del ratón, o un escáner. La interconexión a través del bus de sistema 2575 permite que el procesador central 2573 se comunique con cada subsistema y controle la ejecución de instrucciones desde la memoria del sistema 2572 o el disco fijo 2579, así como el intercambio de información entre los subsistemas. La memoria del sistema 2572 y/o el disco fijo 2579 puede incorporar un medio legible por ordenador. Cualquiera de los valores mencionados en la presente memoria puede ser obtenido a través de un componente a otro componente y puede ser la salida para el usuario.

Un sistema informático puede incluir una pluralidad de los mismos componentes o subsistemas, por ejemplo, conectados entre sí por la interfaz externa 2581 o por una interfaz interna. En algunas realizaciones, los sistemas informáticos, subsistema, o aparatos pueden comunicarse a través de una red. En tales casos, un ordenador puede ser considerado un cliente y otro ordenador un servidor, donde cada uno puede ser parte de un mismo sistema informático. Un cliente y un servidor pueden incluir cada uno múltiples sistemas, subsistemas o componentes.

Debe entenderse que cualquiera de las realizaciones de la presente invención se puede implementar en la forma de control lógico utilizando el hardware (por ejemplo, una disposición de compuerta programable de campo o circuito integrado especifico de la aplicación) y/o el uso de software con un procesador generalmente programable de manera modular o integrada. Como se usa en la presente memoria, un procesador incluye un procesador de múltiples 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 enseñanzas proporcionadas en la presente memoria, una persona con experiencia ordinaria en la técnica conocerá y apreciará otras formas y/o métodos para poner en práctica las realizaciones de la presente invención usando hardware y una combinación de hardware y software.

Cualquiera de los componentes o funciones del software descritos en esta solicitud pueden implementarse como un código de software para ser ejecutado por un procesador utilizando cualquier lenguaje de programación adecuado, tal como, por ejemplo, Java, C++ o Perl usando, por ejemplo, técnicas convencionales u orientadas a objetos. El código de software puede ser almacenado como una serie de instrucciones o comandos en un medio legible por ordenador para su almacenamiento y/o transmisión, los medios adecuados incluyen memoria de acceso aleatorio (RAM), una memoria de sólo lectura (ROM), un medio magnético tal como un disco duro o disco flexible, o un medio óptico tal como un disco compacto (CD) o DVD (disco versátil digital), memoria flash, y similares. El medio legible por ordenador puede ser cualquier combinación de tales dispositivos de almacenamiento o transmisión.

Tales programas también pueden ser codificados y transmitidos utilizando señales portadoras adaptadas para la transmisión a través de redes cableadas, ópticas, y/o inalámbricas conformes a una variedad de protocolos, incluyendo la Internet. Como tal, un medio legible por ordenador de acuerdo con una realización de la presente invención puede ser creado utilizando una señal de datos codificada con tales programas. Los medios legibles por ordenador codificados con el código del programa se pueden envasar con un dispositivo compatible o disponer 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 producto de programa informático simple (por ejemplo, un disco duro, un CD, o un sistema informático completo), y puede estar presente en o dentro de diferentes productos de programa informático dentro de un sistema o red. Un sistema informático puede incluir un monitor, impresora, u otra pantalla adecuada para proporcionar cualquiera de los resultados mencionados en la presente memoria a un usuario.

Cualquiera de los métodos descritos en la presente memoria puede llevarse a cabo total o parcialmente con un sistema informático que incluye uno o más procesadores, que pueden ser configurados para realizar las etapas. Por lo tanto, las realizaciones pueden estar dirigidas a los sistemas informáticos configurados para realizar las etapas de cualquiera de los métodos descritos en la presente memoria, potencialmente con diferentes componentes que realizan una etapa respectiva o un grupo de etapas respectivas. Aunque se presentan como etapas numeradas, las

etapas de los métodos en la presente memoria se pueden realizar al mismo tiempo o en un orden diferente. Además, las porciones de estas etapas se pueden utilizar con porciones de otras etapas de otros métodos. Además, todas o partes de una etapa pueden ser opcionales. Además, cualquiera de las etapas de cualquiera de los métodos se pueden realizar con módulos, circuitos, u otros medios para realizar estas etapas.

Los detalles específicos de realizaciones particulares se pueden combinar en cualquier manera adecuada sin apartarse del espíritu y alcance de las realizaciones de la invención. Sin embargo, otras realizaciones de la invención pueden ser dirigidas a realizaciones específicas con relación a cada aspecto individual, o combinaciones específicas de estos aspectos individuales.

La descripción anterior de realizaciones ilustrativa de la invención se ha presentado con fines ilustrativos y de descripción. No se pretende que sea exhaustiva o limitar la invención a la forma precisa descrita, y muchas modificaciones y variaciones son posibles a la luz de las enseñanzas anteriores. Las realizaciones se eligieron y describieron con el fin de explicar mejor los principios de la invención y sus aplicaciones prácticas para permitir de esta manera que otros expertos en la técnica utilicen mejor la invención en varias realizaciones y con varias modificaciones que sean adecuadas al uso particular contemplado.

Una repetición de “un”, “una” o “el”, “la” se pretende que signifique “uno o más” a menos que se indique específicamente lo contrario.

REIVINDICACIONES

1. Un método para detectar cáncer o cambios premalignos en un sujeto, comprendiendo el método:

obtener una secuencia consenso de un genoma del sujeto, en el que la secuencia consenso se deriva usando etiquetas de secuencia de una muestra del sujeto que contiene más del 50 % de ADN de células sanas;
recibir una o más etiquetas de secuencia para cada una de una pluralidad de fragmentos de ADN en una muestra biológica del sujeto, incluyendo la muestra biológica ADN libre de células;
determinar posiciones genómicas para las etiquetas de secuencia;
comparar las etiquetas de secuencia con la secuencia consenso para determinar un primer número de primeros loci, en el que:

en cada uno de los primeros loci, varias de las etiquetas de secuencia que tienen una variante de secuencia respecto a la secuencia consenso están por encima de un valor de corte, siendo el valor de corte mayor que uno;
determinar un parámetro basándose en un recuento de etiquetas de secuencia que tienen una variante de secuencia en los primeros loci; y
comparar el parámetro con un valor umbral para determinar una clasificación de un nivel de cáncer en el sujeto, correspondiendo el valor umbral a un rango del parámetro para sujetos que tienen la clasificación del nivel de cáncer.

2. El método de la reivindicación 1, en el que el valor de corte para un locus depende de un número total de etiquetas de secuencia que tienen una posición genómica en el locus.

3. El método de la reivindicación 1, en el que se usan diferentes valores de corte para al menos dos de los primeros loci, comprendiendo además el método:
determinar dinámicamente un primer valor de corte para uno de los primeros loci, residiendo uno de los primeros loci dentro de una primera región.

4. El método de la reivindicación 3, en el que el primer valor de corte se determina basándose en una profundidad de secuenciación de uno de los primeros loci, o en el que el primer valor de corte se determina basándose en una tasa de falsos positivos que depende de una tasa de error de secuenciación, una profundidad de secuenciación de la primera región, y varias posiciones de nucleótidos en la primera región.

5. El método de la reivindicación 3, en el que el primer valor de corte se determina basándose en una tasa de falsos positivos que depende de una tasa de error de secuenciación, una profundidad de secuenciación de la primera región y varias posiciones de nucleótidos en la primera región, y en el que el primer valor de corte se determina adicionalmente basándose en un número de verdaderos positivos en la primera región, comprendiendo además el método el cálculo del número de verdaderos positivos para el primer valor de corte basándose en la profundidad de secuenciación D de la primera región y una concentración fraccional *f* de ADN derivado de tumor en la muestra biológica.

6. El método de la reivindicación 5, en el que calcular el número de verdaderos positivos usa la probabilidad de distribución de Poisson de acuerdo con la fórmula:

$$Pb = 1 - \sum_{i=0}^{r-1} Poisson(i, M_p),$$

donde Pb es una probabilidad de detectar verdaderos positivos, y r es el primer valor de corte, y Mp = D x f/2.

7. El método de la reivindicación 3, en el que el primer valor de corte se determina usando uno cualquiera de los siguientes criterios:

si la profundidad de secuenciación es menor que 50, entonces el primer valor de corte es 5,
si la profundidad de secuenciación es 50 - 110, entonces el primer valor de corte es 6,
si la profundidad de secuenciación es 111-200, entonces el primer valor de corte es 7,
si la profundidad de secuenciación es 201 - 310, entonces el primer valor de corte es 8,
si la profundidad de secuenciación es 311 - 450, entonces el primer valor de corte es 9,
si la profundidad de secuenciación es 451 - 620, entonces el primer valor de corte es 10, y
si la profundidad de secuenciación es 621 - 800, entonces el primer valor de corte es 11.

8. El método de la reivindicación 1, en el que el parámetro es una suma ponderada del primer número de primeros loci, en el que una contribución de cada uno de los primeros loci se pondera basándose en un valor de importancia asignado a los primeros loci respectivos.

9. El método de la reivindicación 1, en el que el parámetro incluye una suma de las etiquetas de secuencia que indican una variante de secuencia en el primer número de primeros loci.

10. El método de la reivindicación 13, en el que la suma es una suma ponderada, y en el que uno de los primeros loci tiene un primer peso que es diferente de un segundo peso de un segundo de los primeros loci, en el que el primer peso es mayor que el segundo peso y en el que el uno de los primeros loci está asociado con cáncer, y el segundo de los primeros loci no está asociado con cáncer.

11. El método de la reivindicación 1, en el que la determinación de una posición genómica para una etiqueta de secuencia incluye:

alinear al menos una porción de las etiquetas de secuencia con un genoma de referencia, en el que la alineación de una etiqueta de secuencia permite uno o más mal apareamientos entre la etiqueta de secuencia y la secuencia consenso,
en el que la comparación de las etiquetas de secuencia con la secuencia consenso incluye:

comparar la secuencia consenso con el genoma de referencia para determinar un segundo número de segundos loci que tienen una variante respecto al genoma de referencia;
basándose en la alineación, determinar un tercer número de terceros loci, en el que:

en cada uno de los terceros loci, varias de las etiquetas de secuencia que tienen una variante de secuencia con respecto al genoma de referencia están por encima de un valor de corte; y
la diferencia entre el tercer número y el segundo número permite obtener el primer número de primeros loci,

en el que la diferencia entre el tercer número y el segundo número identifica los primeros loci, y
en el que la determinación del parámetro incluye:
para cada locus del primer número de primeros loci:

contar las etiquetas de secuencia que se alinean con el locus y tienen una variante de secuencia en el locus; y
determinar el parámetro basado en los recuentos respectivos.

12. El método de la reivindicación 1, en el que la determinación de una posición genómica para una etiqueta de secuencia incluye:

alinear al menos una porción de las etiquetas de secuencia con la secuencia consenso, en el que la alineación de una etiqueta de secuencia permite uno o más mal apareamientos entre la etiqueta de secuencia y la secuencia consenso,
en el que la comparación de las etiquetas de secuencia con la secuencia consenso incluye:

basándose en la alineación, identificar etiquetas de secuencia que tienen una variante genómica en una ubicación genómica respecto al genoma constitucional del sujeto;
presentando para cada ubicación genómica una variante de secuencia;

contar un primer número respectivo de etiquetas de secuencia que se alinean con la ubicación genómica y tienen una variante de secuencia en la ubicación genómica; y
determinar un parámetro basado en los primeros números respectivos.

13. El método de la reivindicación 12, en el que la determinación del parámetro basándose en los primeros números respectivos incluye:

sumar los primeros números respectivos para obtener una primera suma; y usar la primera suma para determinar el parámetro, y
en el que usar la primera suma para determinar el parámetro incluye:

restar el número de localizaciones genómicas que muestran una variante de secuencia de la primera suma, o normalizar la primera suma basándose en una cantidad de etiquetas de secuencia alineadas.

14. El método de la reivindicación 1, que comprende además:

obtener una muestra constitucional del sujeto que contiene más del 90 % de ADN constitucional;
realizar la secuenciación aleatoria de los fragmentos de ADN en la muestra constitucional para obtener una o más segundas etiquetas de secuencia para cada una de una pluralidad de fragmentos de ADN en la muestra constitucional;
alinear al menos una porción de las segundas etiquetas de secuencia con un genoma de referencia, en el que la

alineación de una segunda etiqueta de secuencia permite un mal apareamiento entre la segunda etiqueta de secuencia y el genoma de referencia en M o menos ubicaciones genómicas, donde M es un número entero igual a o mayor que uno; y

construir la secuencia consenso basándose en las segundas etiquetas de secuencia y la alineación, en el que la muestra constitucional es la muestra biológica, y en el que construir la secuencia consenso incluye:

determinar un locus homocigoto o un locus heterocigoto que tiene dos alelos.

15. El método de la reivindicación 1, en el que la una o más etiquetas de secuencia se generan a partir de una secuenciación aleatoria de fragmentos de ADN en la muestra biológica.

16. Un programa informático que comprende una pluralidad de instrucciones capaces de ejecución por un sistema informático, que cuando se ejecuta de esta manera controla el sistema informático para realizar el método de una cualquiera de las reivindicaciones precedentes.

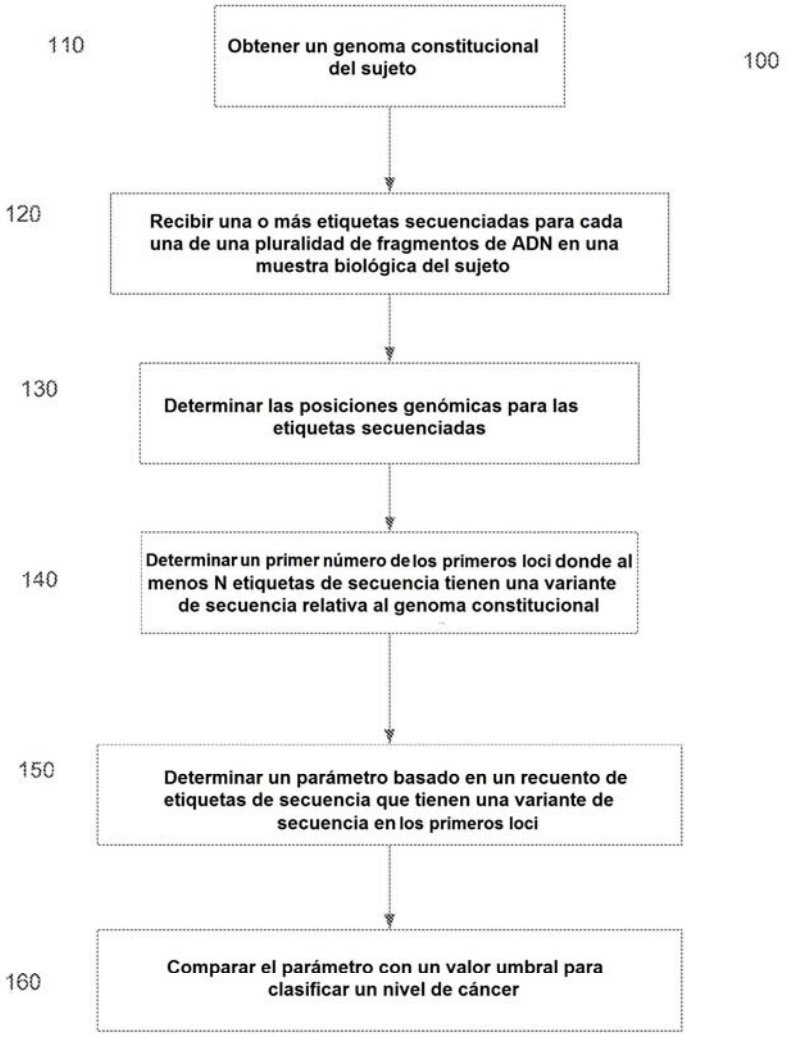

110
Obtener un genoma constitucional del sujeto
100
120
Recibir una o más etiquetas secuenciadas para cada una de una pluralidad de fragmentos de ADN en una muestra biológica del sujeto
130
Determinar las posiciones genómicas para las etiquetas secuenciadas
140
Determinar un primer número de los primeros loci donde al menos N etiquetas de secuencia tienen una variante de secuencia relativa al genoma constitucional
150
Determinar un parámetro basado en un recuento de etiquetas de secuencia que tienen una variante de secuencia en los primeros loci
160
Comparar el parámetro con un valor umbral para clasificar un nivel de cáncer

FIG. 1

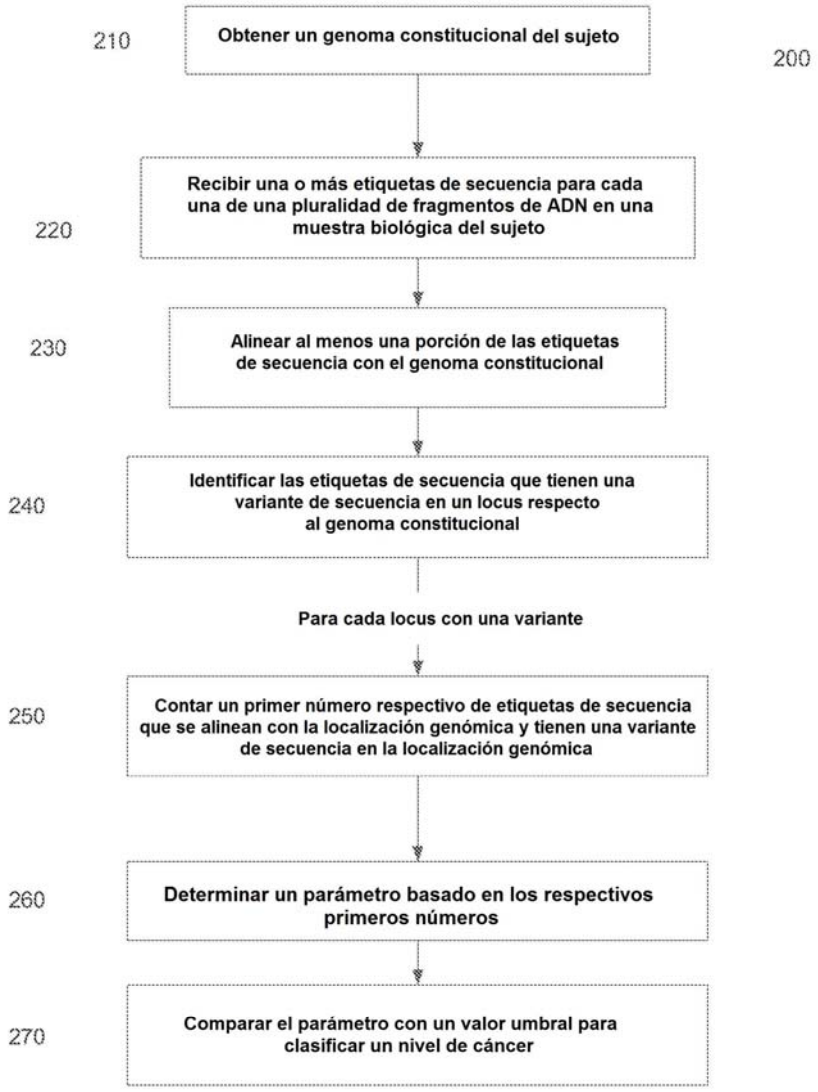

210
Obtener un genoma constitucional del sujeto
200
Recibir una o más etiquetas de secuencia para cada una de una pluralidad de fragmentos de ADN en una muestra biológica del sujeto
220
230
Alinear al menos una porción de las etiquetas de secuencia con el genoma constitucional
240
Identificar las etiquetas de secuencia que tienen una variante de secuencia en un locus respecto al genoma constitucional
Para cada locus con una variante
250
Contar un primer número respectivo de etiquetas de secuencia que se alinean con la localización genómica y tienen una variante de secuencia en la localización genómica
260
Determinar un parámetro basado en los respectivos primeros números
270
Comparar el parámetro con un valor umbral para clasificar un nivel de cáncer

FIG. 2

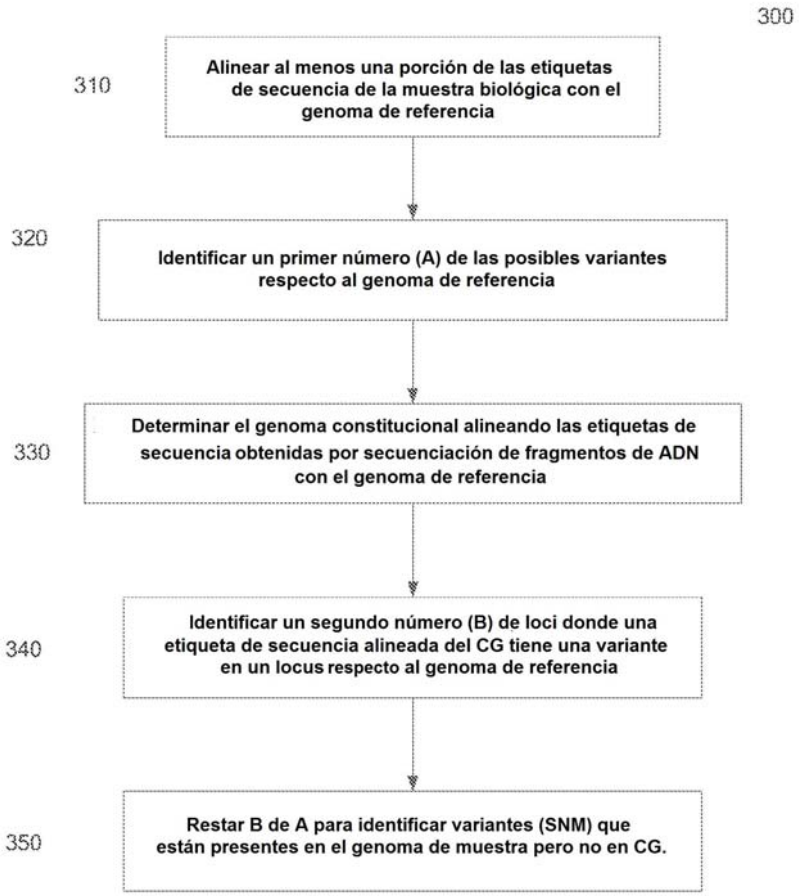

300
310
Alinear al menos una porción de las etiquetas de secuencia de la muestra biológica con el genoma de referencia
320
Identificar un primer número (A) de las posibles variantes respecto al genoma de referencia
330
Determinar el genoma constitucional alineando las etiquetas de secuencia obtenidas por secuenciación de fragmentos de ADN con el genoma de referencia
340
Identificar un segundo número (B) de loci donde una etiqueta de secuencia alineada del CG tiene una variante en un locus respecto al genoma de referencia
350
Restar B de A para identificar variantes (SNM) que están presentes en el genoma de muestra pero no en CG.

FIG. 3

N.º de veces observado	1X		2X		3X		4X		5X		6X	
Profundidad de secuenciación	N.º esperado de locis falsos positivos por errores de secuenciación	N.º esperado de mutaciones identificadas	N.º esperado de locis falsos positivos por errores de secuenciación	N.º esperado de mutaciones identificadas	N.º esperado de locis falsos positivos por errores de secuenciación	N.º esperado de mutaciones identificadas	N.º esperado de locis falsos positivos por errores de secuenciación	N.º esperado de mutaciones identificadas	N.º esperado de locis falsos positivos por errores de secuenciación	N.º esperado de mutaciones identificadas	N.º esperado de locis falsos positivos por errores de secuenciación	N.º esperado de mutaciones identificadas
1	2,968,500	1.46	1,499	4	0	0	0	0	0	0	0	0
10	29,850,499	1,188	45,000	271	40	43	0	5	0	1	0	0
20	59,403,980	1,896	190,000	793	780	241	1	57	0	11	0	2
40	117,631,683	2,584	780,000	1,782	3,293	970	10	429	0	158	0	50
60	174,706,399	2,851	1,770,000	2,403	11,407	1,730	54	1,058	0	564	0	252
80	230,650,961	2,945	3,160,000	2,725	27,387	2,286	176	1,709	1	1,113	0	645
100	285,487,746	2,980	4,950,000	2,879	53,900	2,626	436	2,205	3	1,679	0	1,152
120	339,238,690	2,993	7,140,000	2,948	93,613	2,814	913	2,546	7	2,145	0	1,663
140	391,925,294	2,997	9,730,000	2,978	149,193	2,911	1,703	2,755	15	2,481	0	2,098
160	443,568,633	2,999	12,720,000	2,991	223,807	2,959	2,922	2,873	30	2,701	0	2,426
180	494,189,366	3,000	16,110,000	2,996	318,620	2,981	4,709	2,936	55	2,835	1	2,653
200	543,807,741	3,000	19,900,000	2,999	437,800	2,992	7,187	2,969	94	2,912	1	2,799

400

FIG. 4

N.º de veces observado	1x		2x		3x		4x		5x		6x	
Profundidad de secuenciación	N.º esperado de locis falsos positivos por errores de secuenciación	N.º esperado de mutaciones identificadas	N.º esperado de locis falsos positivos por errores de secuenciación	N.º esperado de mutaciones identificadas	N.º esperado de locis falsos positivos por errores de secuenciación	N.º esperado de mutaciones identificadas	N.º esperado de locis falsos positivos por errores de secuenciación	N.º esperado de mutaciones identificadas	N.º esperado de locis falsos positivos por errores de secuenciación	N.º esperado de mutaciones identificadas	N.º esperado de locis falsos positivos por errores de secuenciación	N.º esperado de mutaciones identificadas
1	2,998,500	74	1,499	1	0	0	0	0	0	0	0	0
10	29,850,499	664	45,000	70	60	6	0	0	0	0	0	0
20	59,403,980	1,180	190,000	271	389	43	1	5	0	1	0	0
40	117,631,683	1,896	780,000	793	3,293	241	10	57	0	11	0	2
60	174,706,599	2,331	1,770,000	1,327	11,407	573	54	197	0	56	0	13
80	230,650,961	2,594	3,160,000	1,782	27,387	970	176	429	1	158	0	50
100	285,487,746	2,754	4,950,000	2,138	53,900	1,369	426	727	3	326	0	126
120	339,238,600	2,853	7,140,000	2,403	93,613	1,738	913	1,058	7	554	0	252
140	391,925,294	2,969	9,730,000	2,592	149,193	2,037	1,763	1,390	15	824	0	427
160	443,568,633	2,945	12,720,000	2,725	223,367	2,286	2,922	1,700	30	1,113	0	645
180	494,189,366	2,967	16,110,000	2,817	318,620	2,479	4,760	1,973	55	1,404	1	891
200	543,807,741	2,980	19,900,000	2,879	437,800	2,626	7,187	2,205	94	1,629	1	1,152

500

FIG. 5

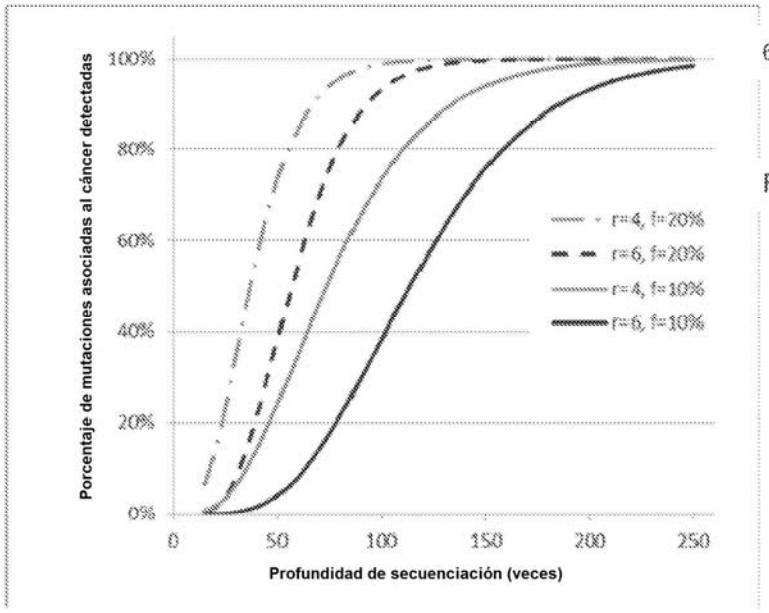

600
Porcentaje de mutaciones asociadas al cáncer detectadas
100%
80%
60%
40%
20%
0%
0
50
100
150
200
250
Profundidad de secuenciación (veces)
r=4, f=20%
r=6, f=20%
r=4, f=10%
r=6, f=10%

FIG. 6A

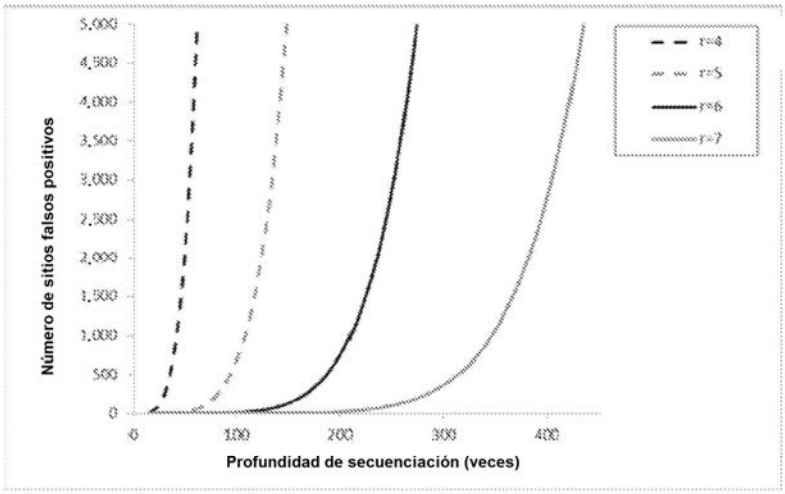
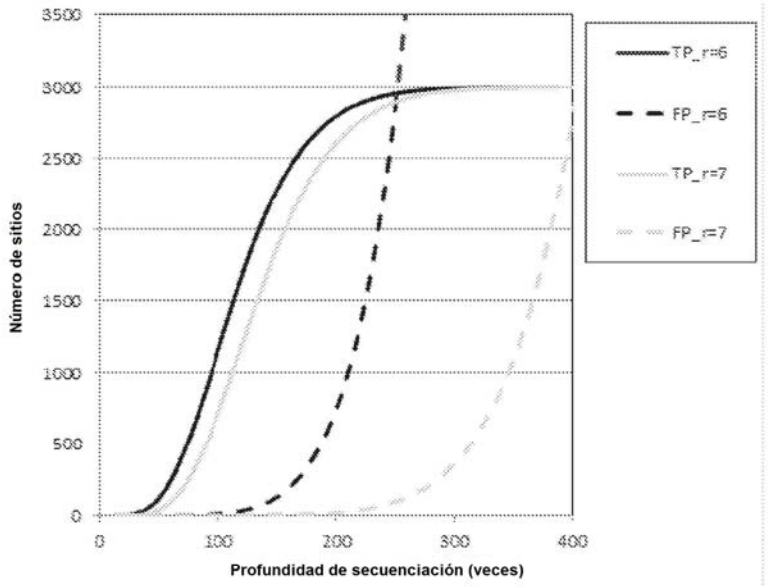
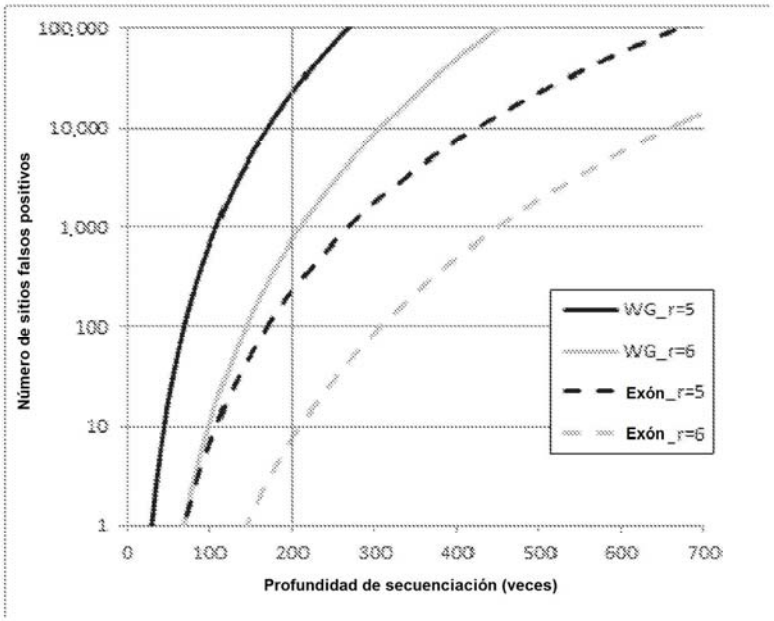

650
Número de sitios falsos positivos
5.000
4.500
4.000
3.500
3.000
2.500
2.000
1.500
1.000
500
0
0
100
200
300
400
Profundidad de secuenciación (veces)
r=4
r=5
r=6
r=7

FIG. 6B



700

FIG. 7A



750

FIG. 7B

Caso	N.º de SNV detectadas en el tejido de tumor	Punto temporal	N.º de SNV asociadas al tumor secuenciadas del plasma (% de SNV observadas en el tejido de tumor)	N.º de lecturas de secuencia de ADN de plasma que muestran SNV (*p*)	N.º de lecturas de secuencia de ADN de plasma que muestran secuencia de tipo silvestre (*q*)	Concentración fraccional deducida de ADN derivado de tumor por análisis de SNV $(\frac{2p}{p+q})$	Concentración fraccional deducida de ADN derivado de tumor por análisis de GAAL
HCC1	2.840	Pre-tratam.	2.569 (94%)	11.389	31.602	53%	52%
		Post-tratam.	44 (1,5%)	91	46.898	0,4%	1,4%
HCC2	3.105	Pre-tratam.	1.097 (35%)	1.490	57.865	5,0%	4,3%
		Post-tratam.	72 (2,3%)	206	66.692	0,6%	0,9%
HCC3	3.171	Pre-tratam.	461 (15%)	525	48.886	2,1%	5,6%
		Post-tratam.	31 (1%)	67	58.862	0,2%	0,9%
HCC4	1.334	Pre-tratam.	201 (15%)	248	18.527	2,6%	7,6%
		Post-tratam.	74 (5,5%)	149	22.144	1,3%	2,7%

800

FIG. 8

	SNV DE HCC1		SNV DE HCC2		SNV DE HCC3		SNV DE HCC4		Todas las SNV		
Controles	Lecturas de secuencia de ADN de plasma que muestran SNV asociadas al tumor	Lecturas de secuencia de ADN de plasma que muestran la secuencia de tipo silvestre	Lecturas de secuencia de ADN de plasma que muestran SNV asociadas al tumor	Lecturas de secuencia de ADN de plasma que muestran la secuencia de tipo silvestre	Lecturas de secuencia de ADN de plasma que muestran SNV asociadas al tumor	Lecturas de secuencia de ADN de plasma que muestran la secuencia de tipo silvestre	Lecturas de secuencia de ADN de plasma que muestran SNV asociadas al tumor	Lecturas de secuencia de ADN de plasma que muestran la secuencia de tipo silvestre	Lecturas de secuencia de ADN de plasma que muestran SNV asociadas al tumor (p)	Lecturas de secuencia de ADN de plasma que muestran la secuencia de tipo silvestre (q)	Concentración fraccional de todas las SNV asociadas al tumor en plasma $(\frac{2p}{p+q})$
C01	40	31.261	66	38.172	30	33.823	69	14.389	205	117.645	0,35%
C02	60	34.368	112	42.660	37	37.462	58	16.097	367	130.587	0,41%
C03	55	37.931	100	46.297	50	41.181	74	17.039	279	142.448	0,39%
C04	41	40.277	87	48.441	39	43.541	87	17.358	254	149.617	0,34%
C05	42	31.655	71	38.787	46	34.566	70	14.821	229	119.829	0,38%
C06	41	34.604	97	43.999	26	37.477	79	16.762	243	132.842	0,37%
C07	52	38.664	118	47.200	61	41.840	72	17.601	303	145.305	0,42%
C08	57	38.514	131	46.464	42	41.662	81	17.275	311	143.915	0,43%
C09	41	34.945	115	43.760	33	38.039	81	16.503	270	133.247	0,41%
C10	44	35.197	89	43.622	51	38.370	65	16.009	249	133.198	0,37%
C11	41	36.693	95	45.108	39	39.732	72	16.801	247	138.334	0,36%
C12	45	39.506	104	50.085	46	42.536	91	18.598	286	150.725	0,38%
C13	36	33.799	71	40.309	39	36.241	70	14.450	216	124.799	0,35%
C14	42	38.199	104	47.832	42	41.245	91	17.807	279	145.083	0,38%
C15	49	36.106	90	43.129	41	38.432	73	16.154	253	133.821	0,38%
C16	38	33.633	66	40.023	41	36.431	52	14.502	197	124.589	0,32%
										Media	0,38%

FIG. 9

900

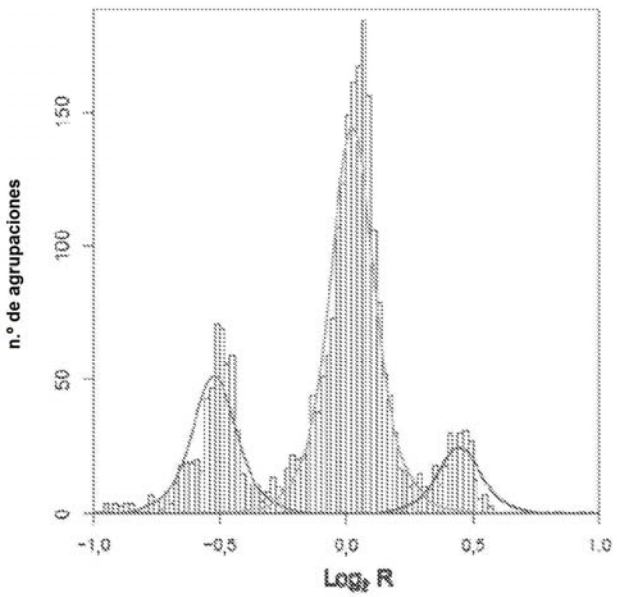


FIG. 10A

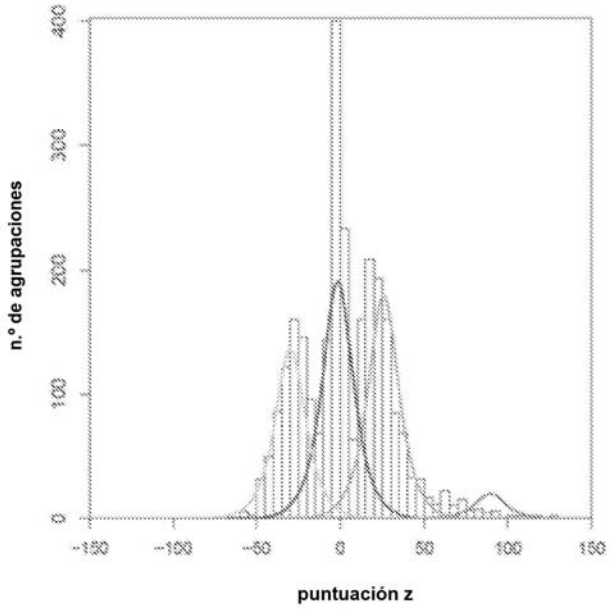


FIG. 10B

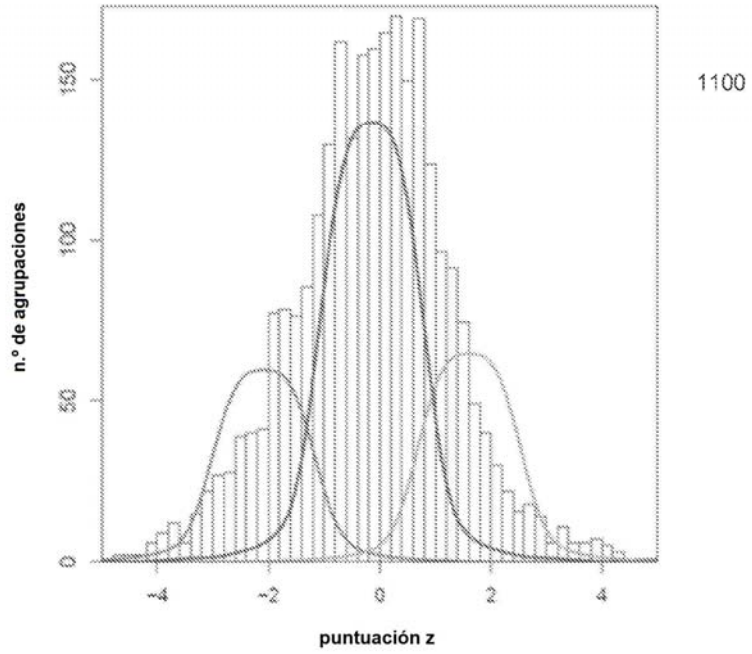


FIG. 11

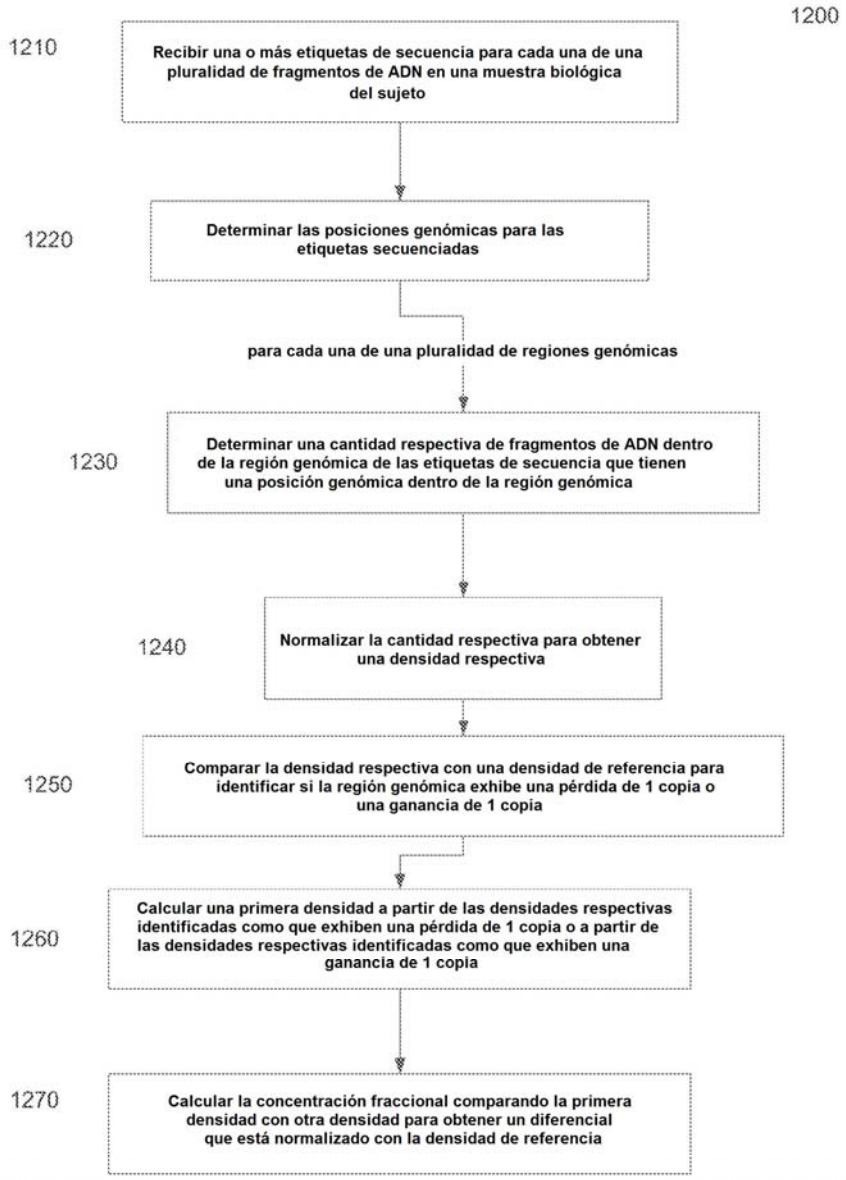

1200
1210
Recibir una o más etiquetas de secuencia para cada una de una pluralidad de fragmentos de ADN en una muestra biológica del sujeto
1220
Determinar las posiciones genómicas para las etiquetas secuenciadas
para cada una de una pluralidad de regiones genómicas
1230
Determinar una cantidad respectiva de fragmentos de ADN dentro de la región genómica de las etiquetas de secuencia que tienen una posición genómica dentro de la región genómica
1240
Normalizar la cantidad respectiva para obtener una densidad respectiva
1250
Comparar la densidad respectiva con una densidad de referencia para identificar si la región genómica exhibe una pérdida de 1 copia o una ganancia de 1 copia
1260
Calcular una primera densidad a partir de las densidades respectivas identificadas como que exhiben una pérdida de 1 copia o a partir de las densidades respectivas identificadas como que exhiben una ganancia de 1 copia
1270
Calcular la concentración fraccional comparando la primera densidad con otra densidad para obtener un diferencial que está normalizado con la densidad de referencia

FIG. 12

N.º de veces detectado en plasma (N)	N.º de mutaciones potenciales detectadas N veces en plasma	N.º de mutaciones potenciales detectadas ≥ N veces en plasma	Mutaciones de tumor detectadas N veces en plasma	Mutaciones de tumor detectadas ≥ N veces en plasma	Porcentaje de mutaciones detectadas comparado con N =1	Valor predictivo positivo
1	3,605,085	3,623,970	61	2,064	100%	0,06%
2	13,518	18,885	35	2,003	97%	11%
3	2,029	5,367	71	1,968	95%	37%
4	766	3,338	83	1,897	92%	57%
5	437	2,572	140	1,814	88%	71%
6	313	2,135	161	1,674	81%	78%
7	265	1,822	165	1,513	73%	83%
8	261	1,557	192	1,348	65%	87%
9	231	1,296	185	1,156	56%	89%
10	191	1,065	160	971	47%	91%
11	179	874	165	811	39%	93%
12	148	695	137	646	31%	93%
13	122	547	109	509	25%	93%
14	99	425	92	400	19%	94%
15	74	326	71	308	15%	94%
16	54	252	49	237	11%	94%
17	39	198	39	188	9%	95%
18	35	159	33	149	7%	94%
19	24	124	23	116	6%	94%
20	20	100	20	93	5%	93%
21	18	80	17	73	4%	91%
22	11	62	10	56	3%	90%
23	9	51	9	46	2%	90%
24	8	42	5	37	2%	88%
25	6	34	6	32	2%	94%
26	7	28	5	26	1%	93%
27	8	21	8	21	1%	100%
28	1	13	1	13	1%	100%
29	4	12	4	12	1%	100%
30	8	8	8	8	0%	100%

1300

FIG. 13A

N.º de veces detectado en plasma (N)	N.º de mutaciones potenciales detectadas N veces en plasma	N.º de mutaciones potenciales detectadas ≥ N veces en plasma	Mutaciones de tumor detectadas N veces en plasma	Mutaciones de tumor detectadas ≥ N veces en plasma	Porcentaje de mutaciones detectadas comparado con N =1	Valor predictivo positivo
1	3,155,634	3,167,144	51	54	100%	<0,01%
2	9,835	11,510	2	3	6%	0,03%
3	1,123	1,675	1	1	2%	0,06%
4	314	552	-	-	-	-
5	132	238	-	-	-	-
6	49	106	-	-	-	-
7	22	57	-	-	-	-
8	11	35	-	-	-	-
9	6	24	-	-	-	-
10	5	18	-	-	-	-
11	4	13	-	-	-	-
12	2	9	-	-	-	-
13	2	7	-	-	-	-
14	3	5	-	-	-	-
15	-	2	-	-	-	-
16	-	2	-	-	-	-
17	1	2	-	-	-	-
18	-	1	-	-	-	-
19	1	1	-	-	-	-
20	-	-	-	-	-	-
21	-	-	-	-	-	-
22	-	-	-	-	-	-
23	-	-	-	-	-	-
24	-	-	-	-	-	-
25	-	-	-	-	-	-
26	-	-	-	-	-	-
27	-	-	-	-	-	-
28	-	-	-	-	-	-
29	-	-	-	-	-	-
30	-	-	-	-	-	-

1300

FIG. 13B

Profundidad de la secuenciación	Plasma pre-tratamiento			Plasma post-tratamiento		
	Porcentaje de nucleótidos con la profundidad de secuencia especificada	N.º total de posiciones de nucleótidos con variaciones de nucleótido único detectadas en plasma	N.º de sitios con cambios correspondientes en el tumor	Porcentaje de nucleótidos con la profundidad de secuencia especificada	N.º total de posiciones de nucleótidos con variaciones de nucleótido único detectadas en plasma	N.º de sitios con cambios correspondientes en el tumor
<50	36,3%	9	8	30,5%	0	0
50 – 125	29,9%	15	15	31,2%	0	0
126 – 235	18,6%	21	21	20,8%	0	0
236 – 380	9,4%	7	6	10,9%	0	0
381 – 560	4,2%	2	2	4,8%	0	0
561 – 760	1,6%	3	3	1,7%	0	0
Total		57	55		0	0

1400

FIG. 14A

Profundidad de la secuenciación	Plasma pre-tratamiento			Plasma post-tratamiento		
	Porcentaje de nucleótidos con la profundidad de secuencia especificada	N.º total de posiciones de nucleótidos con variaciones de nucleótido único detectadas en plasma	N.º de sitios con cambios correspondientes en el tumor	Porcentaje de nucleótidos con la profundidad de secuencia especificada	N.º total de posiciones de nucleótidos con variaciones de nucleótido único detectadas en plasma	N.º de sitios con cambios correspondientes en el tumor
<50	30,0%	0	0	27,1%	0	0
50 – 125	33,6%	2	2	32,5%	0	0
126 – 235	21,6%	6	6	22,2%	0	0
236 – 380	10,2%	5	5	11,5%	0	0
381 – 560	3,7%	4	4	4,9%	0	0
561 – 760	0,9%	1	1	1,7%	0	0
Total		18	18		0	0

1450

FIG. 14B

Profundidad de la secuenciación	Plasma pre-tratamiento			Plasma post-tratamiento		
	Porcentaje de nucleótidos con la profundidad de secuencia especificada	N.º total de posiciones de nucleótidos con variaciones de nucleótido único detectadas en plasma	N.º de sitios con cambios correspondientes en el tumor	Porcentaje de nucleótidos con la profundidad de secuencia especificada	N.º total de posiciones de nucleótidos con variaciones de nucleótido único detectadas en plasma	N.º de sitios con cambios correspondientes en el tumor
<50	38,8%	0	0	18,5%	0	0
50 – 125	45,7%	0	0	28,6%	0	0
126 – 235	15,2%	0	0	25,1%	0	0
236 – 380	0,2%	0	0	16,2%	0	0
381 – 560	<0,1%	0	0	8,3%	0	0
561 – 760	<0,1%	0	0	3,4%	0	0
Total		0	0		0	0

FIG. 15A

1500

Profundidad de la secuenciación	Plasma pre-tratamiento			Plasma post-tratamiento		
	Porcentaje de nucleótidos con la profundidad de secuencia especificada	N.º total de posiciones de nucleótidos con variaciones de nucleótido único detectadas en plasma	N.º de sitios con cambios correspondientes en el tumor	Porcentaje de nucleótidos con la profundidad de secuencia especificada	N.º total de posiciones de nucleótidos con variaciones de nucleótido único detectadas en plasma	N.º de sitios con cambios correspondientes en el tumor
<50	26,7%	0	0	24,4%	0	0
50 – 125	35,9%	1	1	31,2%	1	1
126 – 235	24,7%	2	2	23,1%	0	0
236 – 380	10,4%	0	0	13,1%	0	0
381 – 560	2,2%	0	0	6,0%	0	0
561 – 760	0,1%	0	0	2,2%	0	0
Total		3	3		1	1

FIG. 15B

1550

Profundidad de la secuenciación	Plasma pre-tratamiento			Plasma post-tratamiento		
	Porcentaje de nucleótidos con la profundidad de secuencia especificada	N.º total de posiciones de nucleótidos con variaciones de nucleótido único detectadas en plasma	N.º de sitios con cambios correspondientes en el tumor	Porcentaje de nucleótidos con la profundidad de secuencia especificada	N.º total de posiciones de nucleótidos con variaciones de nucleótido único detectadas en plasma	N.º de sitios con cambios correspondientes en el tumor
<50	31,1%	7	7	30,8%	0	0
50 – 125	30,3%	14	14	33,2%	1	0
126 – 235	20,5%	19	19	21,6%	0	0
236 – 380	11,3%	13	13	10,3%	0	0
381 – 560	5,2%	7	4	3,5%	0	0
561 – 760	1,8%	4	2	0,7%	0	0
Total		64	59		1	0

FIG. 16

1600

Profundidad de secuenciación (veces)	Sensibilidad cuando el mutante se define como observado en los datos de secuenciación del tejido de tumor durante al menos n veces, donde n=					
	4	5	6	7	8	9
20	73%	56%	38%	24%	13%	7%
21	77%	60%	43%	28%	16%	9%
22	80%	64%	47%	31%	19%	11%
23	83%	68%	51%	35%	22%	13%
24	85%	71%	55%	39%	26%	15%
25	87%	75%	59%	43%	29%	18%
26	89%	78%	63%	47%	33%	21%
27	90%	80%	67%	51%	36%	24%
28	92%	83%	70%	55%	40%	27%
29	93%	85%	73%	59%	44%	30%
30	94%	87%	76%	62%	48%	34%
31	95%	89%	78%	66%	51%	37%
32	96%	90%	81%	69%	55%	41%
33	96%	91%	83%	72%	58%	44%
34	97%	93%	85%	74%	61%	48%
35	97%	94%	87%	77%	65%	51%
36	98%	95%	88%	79%	68%	54%
37	98%	95%	90%	82%	70%	58%
38	99%	96%	91%	84%	73%	61%
39	99%	97%	92%	85%	76%	64%
40	99%	97%	93%	87%	78%	67%
41	99%	98%	94%	88%	80%	69%
42	99%	98%	95%	90%	82%	72%
43	99%	98%	96%	91%	84%	75%
44	100%	98%	96%	92%	86%	77%
45	100%	99%	97%	93%	87%	79%
46	100%	99%	97%	94%	89%	81%
47	100%	99%	98%	95%	90%	83%
48	100%	99%	98%	95%	91%	84%
49	100%	99%	98%	96%	92%	86%
50	100%	99%	99%	97%	93%	88%
51	100%	100%	99%	97%	94%	89%
52	100%	100%	99%	97%	95%	90%
53	100%	100%	99%	98%	95%	91%
54	100%	100%	99%	98%	96%	92%
55	100%	100%	99%	98%	96%	93%
56	100%	100%	99%	99%	97%	94%
57	100%	100%	100%	99%	97%	95%
58	100%	100%	100%	99%	98%	95%
59	100%	100%	100%	99%	98%	96%
60	100%	100%	100%	99%	98%	96%
61	100%	100%	100%	99%	98%	97%
62	100%	100%	100%	99%	99%	97%
63	100%	100%	100%	100%	99%	97%
64	100%	100%	100%	100%	99%	98%
65	100%	100%	100%	100%	99%	98%
66	100%	100%	100%	100%	99%	98%
67	100%	100%	100%	100%	99%	99%
68	100%	100%	100%	100%	99%	99%
69	100%	100%	100%	100%	100%	99%
70	100%	100%	100%	100%	100%	99%

1700

FIG. 17

Profundidad de secuenciación (veces)	N.º de falsos positivos en el genoma completo cuando un mutante se identifica como observado en los datos de secuenciación del tumor durante al menos n veces, donde n=					
	4	5	6	7	8	9
20	1	0	0	0	0	0
21	1	0	0	0	0	0
22	1	0	0	0	0	0
23	1	0	0	0	0	0
24	1	0	0	0	0	0
25	1	0	0	0	0	0
26	2	0	0	0	0	0
27	2	0	0	0	0	0
28	2	0	0	0	0	0
29	3	0	0	0	0	0
30	3	0	0	0	0	0
31	3	0	0	0	0	0
32	4	0	0	0	0	0
33	5	0	0	0	0	0
34	5	0	0	0	0	0
35	6	0	0	0	0	0
36	7	0	0	0	0	0
37	7	0	0	0	0	0
38	8	0	0	0	0	0
39	9	0	0	0	0	0
40	10	0	0	0	0	0
41	11	0	0	0	0	0
42	12	0	0	0	0	0
43	14	0	0	0	0	0
44	15	0	0	0	0	0
45	17	0	0	0	0	0
46	18	0	0	0	0	0
47	20	0	0	0	0	0
48	22	0	0	0	0	0
49	24	0	0	0	0	0
50	26	0	0	0	0	0
51	28	0	0	0	0	0
52	30	0	0	0	0	0
53	33	0	0	0	0	0
54	35	0	0	0	0	0
55	38	0	0	0	0	0
56	41	0	0	0	0	0
57	44	0	0	0	0	0
58	47	0	0	0	0	0
59	51	0	0	0	0	0
60	54	0	0	0	0	0
61	58	0	0	0	0	0
62	62	0	0	0	0	0
63	66	0	0	0	0	0
64	71	0	0	0	0	0
65	75	0	0	0	0	0
66	80	0	0	0	0	0
67	85	0	0	0	0	0
68	90	0	0	0	0	0
69	96	0	0	0	0	0
70	102	0	0	0	0	0

FIG. 18

1800

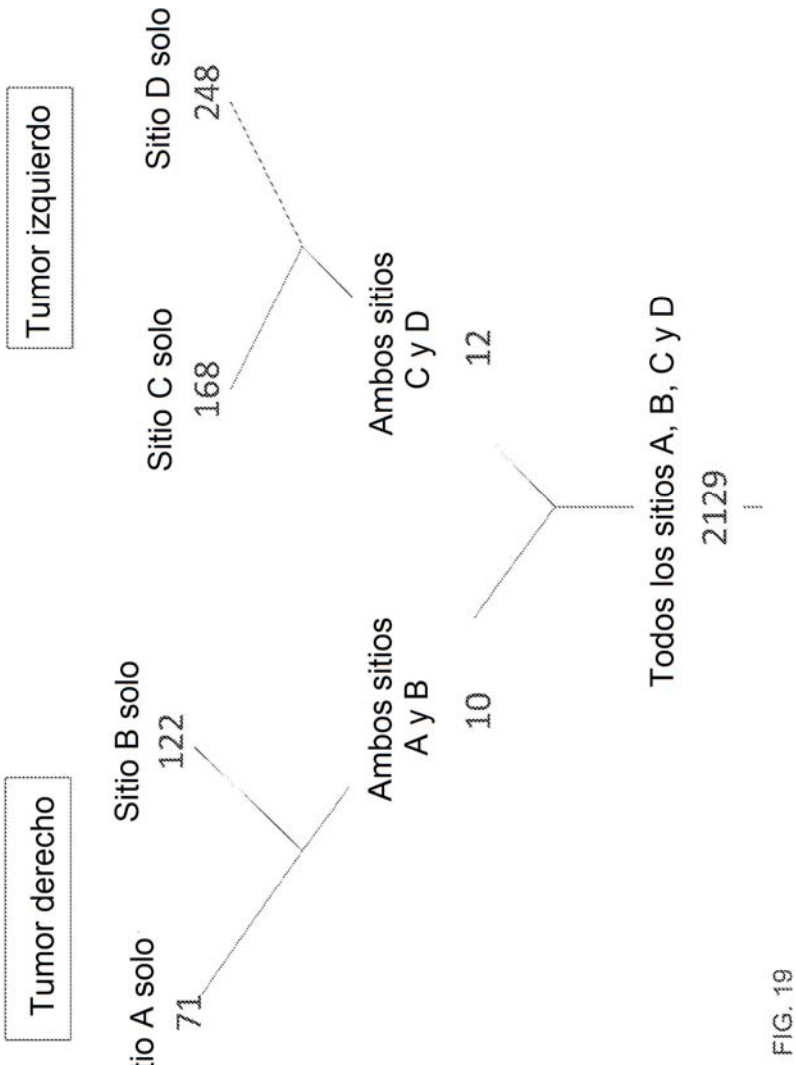

Tumor derecho
Tumor izquierdo
Sitio A solo
71
Sitio B solo
122
Ambos sitios
A y B
10
Sitio C solo
168
Sitio D solo
248
Ambos sitios
C y D
12
Todos los sitios A, B, C y D
2129

FIG. 19

Categoría	N.º de loci con una mutación	Plasma pre-tratamiento			Plasma post-tratamiento		
		N.º de fragmentos de ADN que llevan alelos de tipo silvestre	N.º de fragmentos de ADN que llevan alelos mutantes	Concentración fraccional de ADN derivado de tumor	N.º de fragmentos de ADN que llevan alelos de tipo silvestre	N.º de fragmentos de ADN que llevan alelos mutantes	Concentración fraccional de ADN derivado de tumor
A	71	3.321	37	2,20%	2.149	5	0,46%
B	122	5.633	8	0,28%	3.516	1	0,06%
C	168	8.507	51	1,19%	5.438	2	0,07%
D	248	10.880	26	0,48%	7.060	2	0,06%
AB	10	423	21	9,46%	297	0	0,00%
CD	12	569	3	1,05%	333	0	0,00%
ABCD	2129	70.940	21.344	46,26%	61.417	54	0,18%

FIG. 20

2000

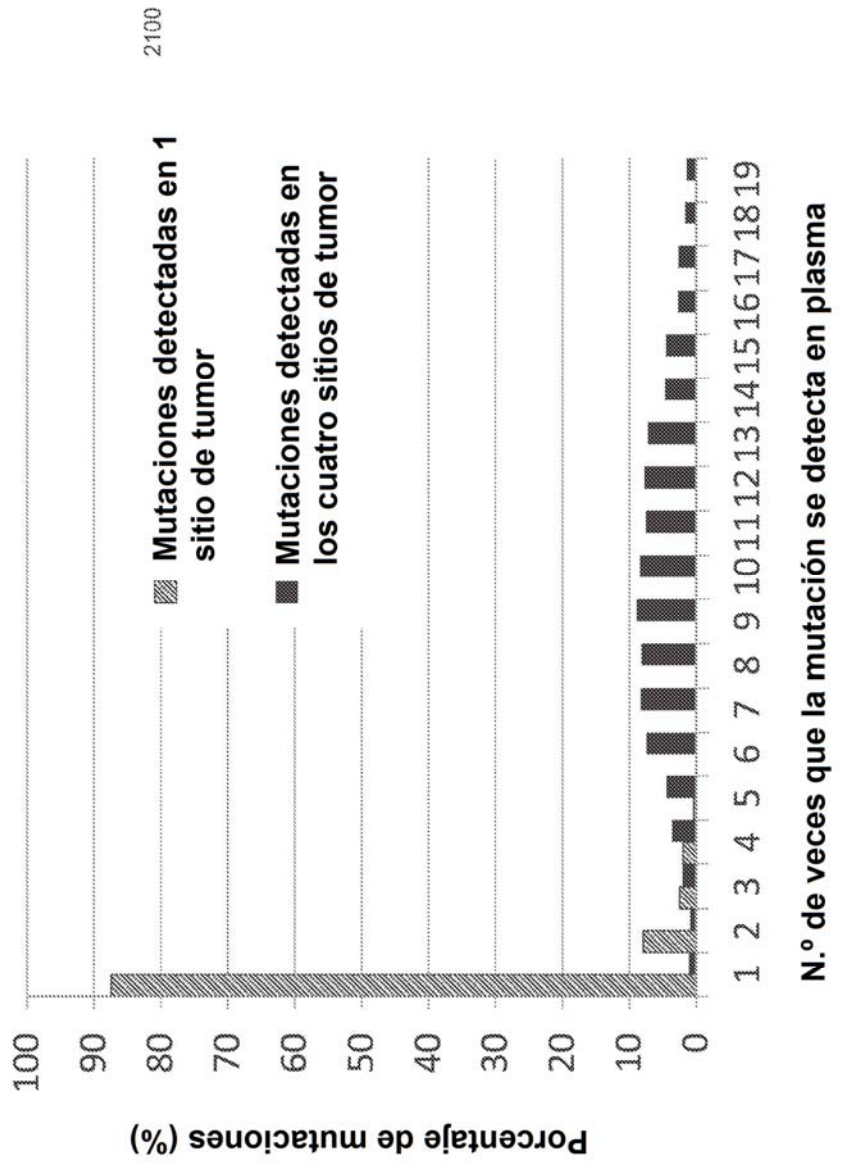

2100
Mutaciones detectadas en 1 sitio de tumor
Mutaciones detectadas en los cuatro sitios de tumor
Porcentaje de mutaciones (%)
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
N.º de veces que la mutación se detecta en plasma

FIG. 21

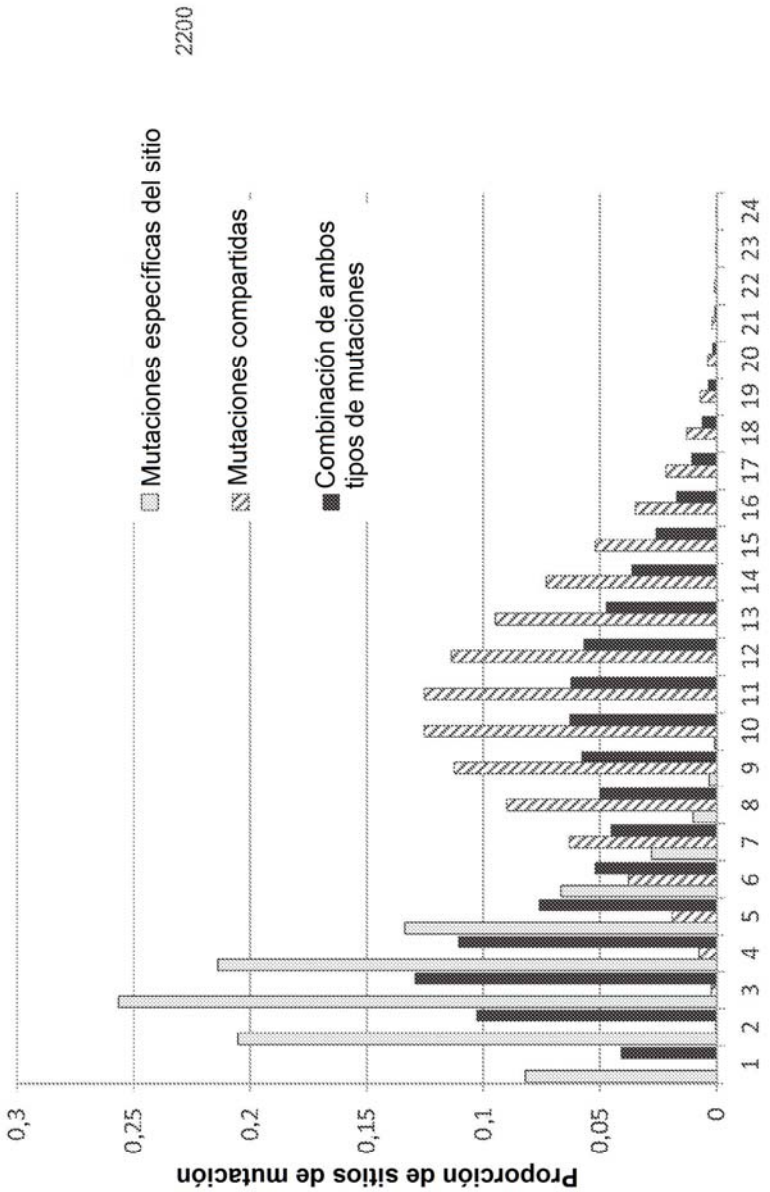
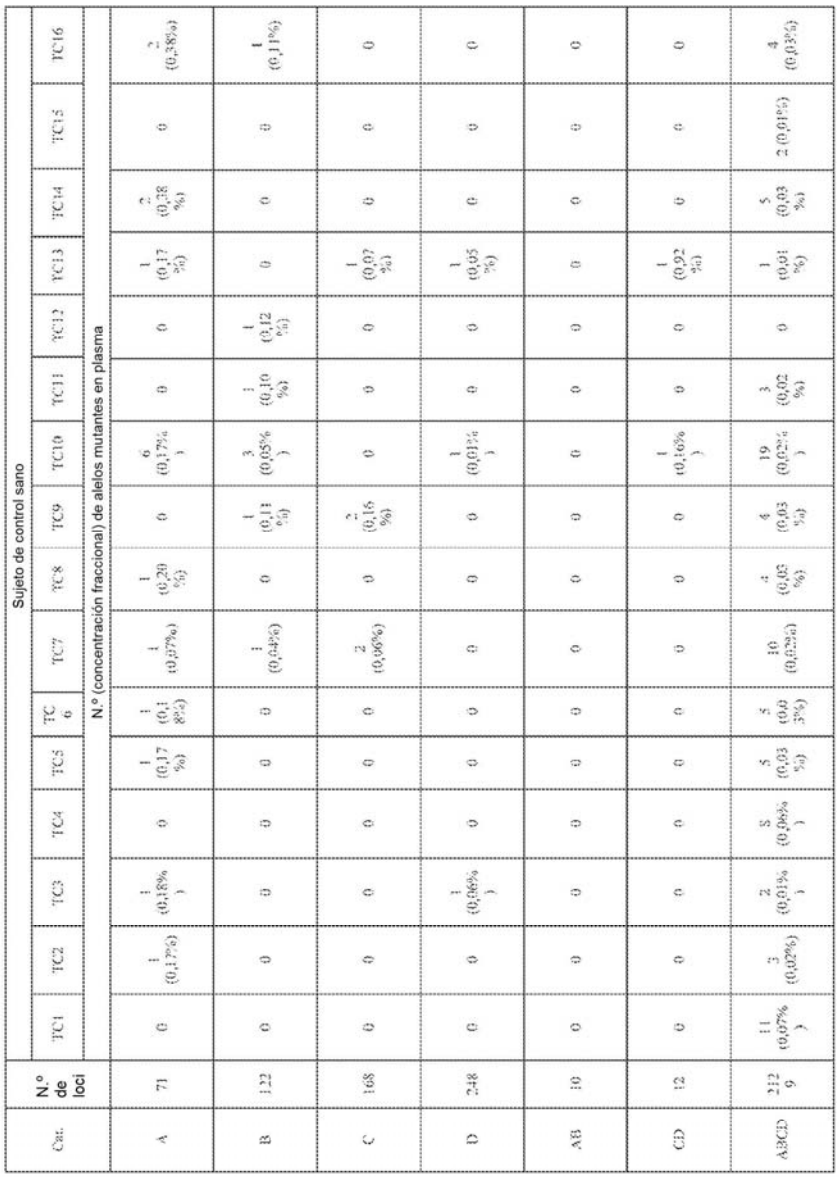

2200
Mutaciones específicas del sitio
Mutaciones compartidas
Combinación de ambos tipos de mutaciones
Proporción de sitios de mutación
0,3
0,25
0,2
0,15
0,1
0,05
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
N.º de veces detectada en plasma

FIG. 22



Cat.	N.º de loci	Sujeto de control sano															
		N.º (concentración fraccional) de alelos mutantes en plasma															
		TC1	TC2	TC3	TC4	TC5	TC 6	TC7	TC8	TC9	TC10	TC11	TC12	TC13	TC14	TC15	TC16
A	71	0	1 (0,17%)	1 (0,18%)	0	1 (0,17%)	1 (0,18%)	1 (0,07%)	1 (0,20%)	0	6 (0,17%)	0	0	1 (0,17%)	2 (0,28%)	0	2 (0,38%)
B	122	0	0	0	0	0	0	1 (0,04%)	0	1 (0,11%)	3 (0,05%)	1 (0,10%)	1 (0,12%)	0	0	0	1 (0,11%)
C	168	0	0	0	0	0	0	2 (0,06%)	0	2 (0,16%)	0	0	0	1 (0,07%)	0	0	0
D	248	0	0	1 (0,06%)	0	0	0	0	0	0	1 (0,01%)	0	0	1 (0,05%)	0	0	0
AB	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CD	12	0	0	0	0	0	0	0	0	0	1 (0,16%)	0	0	1 (0,92%)	0	0	0
ABCD	2129	11 (0,07%)	3 (0,02%)	2 (0,01%)	8 (0,06%)	5 (0,03%)	5 (0,03%)	10 (0,07%)	4 (0,03%)	4 (0,03%)	19 (0,02%)	3 (0,02%)	0	1 (0,01%)	5 (0,03%)	2 (0,01%)	4 (0,03%)

FIG. 23

2300

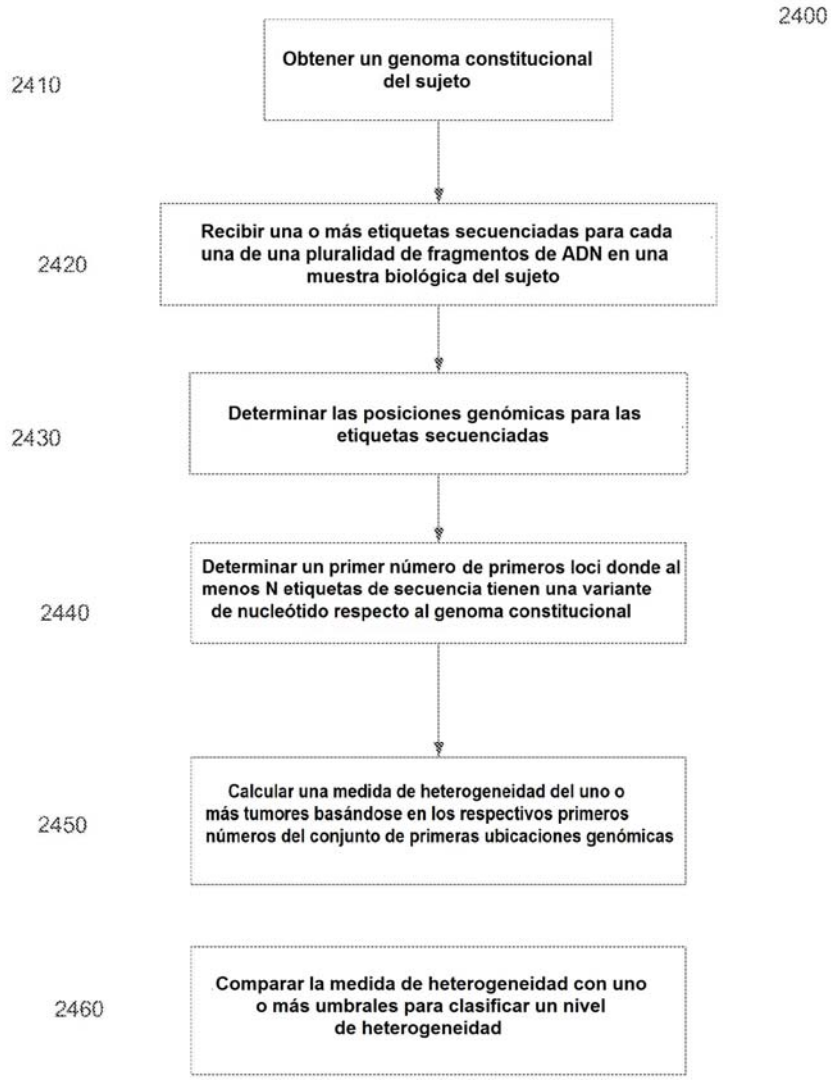

2400
2410
Obtener un genoma constitucional del sujeto
2420
Recibir una o más etiquetas secuenciadas para cada una de una pluralidad de fragmentos de ADN en una muestra biológica del sujeto
2430
Determinar las posiciones genómicas para las etiquetas secuenciadas
2440
Determinar un primer número de primeros loci donde al menos N etiquetas de secuencia tienen una variante de nucleótido respecto al genoma constitucional
2450
Calcular una medida de heterogeneidad del uno o más tumores basándose en los respectivos primeros números del conjunto de primeras ubicaciones genómicas
2460
Comparar la medida de heterogeneidad con uno o más umbrales para clasificar un nivel de heterogeneidad

FIG. 24

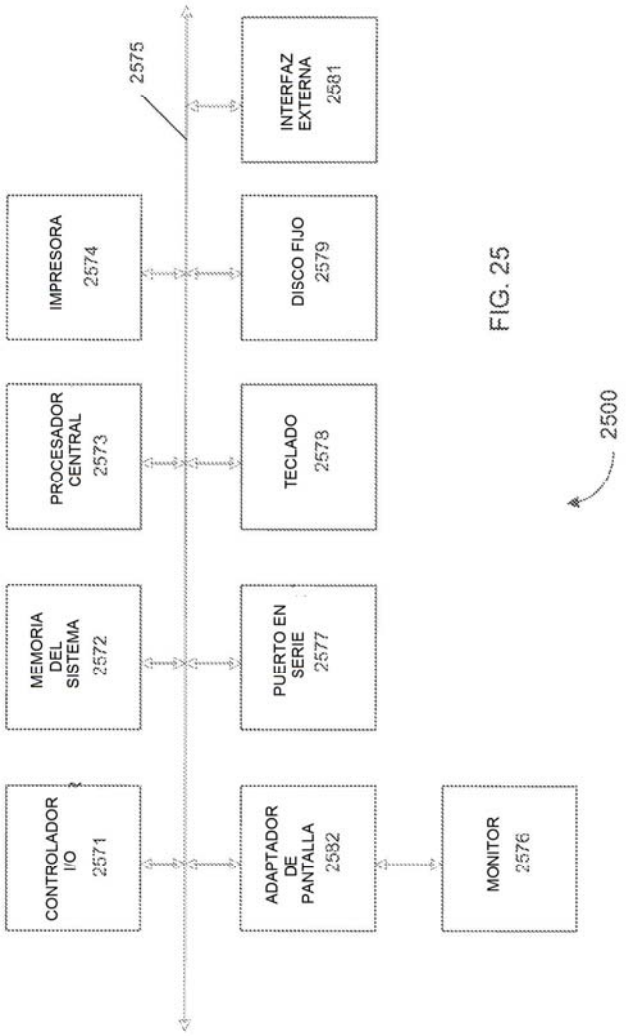

2575
INTERFAZ EXTERNA 2581
IMPRESORA 2574
DISCO FIJO 2579
PROCESADOR CENTRAL 2573
TECLADO 2578
MEMORIA DEL SISTEMA 2572
PUERTO EN SERIE 2577
CONTROLADOR I/O 2571
ADAPTADOR DE PANTALLA 2582
MONITOR 2576
2500

FIG. 25