(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 698 531**

(51) Int. CI.:

C12Q 1/6827 (2008.01)
C12Q 1/6858 (2008.01)
C12Q 1/6883 (2008.01)
C12Q 1/6886 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **17.07.2015 PCT/GB2015/052086**

(87) Fecha y número de publicación internacional: **21.01.2016 WO16009224**

(96) Fecha de presentación y número de la solicitud europea: **17.07.2015 E 15741290 (9)**

(97) Fecha y número de publicación de la concesión europea: **12.09.2018 EP 3169798**

(54) Título: **Un método para detectar una variante genética**

(30) Prioridad:

18.07.2014 GB 201412834

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
05.02.2019

(73) Titular/es:

CANCER RESEARCH TECHNOLOGY LIMITED (100.0%)
Angel Building 407 St. John Street
London EC1V 4AD, GB

(72) Inventor/es:

ROSENFELD, NITZAN;
FORSHEW, TIM;
MARASS, FRANCESCO y
MURTAZA, MUHAMMED

(74) Agente/Representante:

VALLEJO LÓPEZ, Juan Pedro

DESCRIPCIÓN

Un método para detectar una variante genética

Campo de la invención

La presente invención se refiere a métodos para detectar variantes genéticas en muestras de ADN.

Antecedentes de la invención

Una variante genética es uno o más nucleótidos que difieren de una secuencia de ADN de referencia para una región dada. Por ejemplo, una variante genética puede comprender una deleción, sustitución o inserción de uno o más nucleótidos.

Una muestra de ADN puede analizarse para detectar variantes genéticas conocidas o para descubrir variantes genéticas previamente desconocidas en una región de interés determinando la secuencia de ADN en la región de interés y comparando la secuencia determinada con la secuencia de referencia.

La secuenciación de ADN se puede realizar utilizando varias técnicas, como el método clásico por terminación de cadena, o una de varias metodologías de última generación (NGS) de secuenciación de alto rendimiento, revisado por Metzker, M.L., Nat Rev Genet Enero de 2010;11(1): 31-46.

La secuenciación con Illumina, pirosecuenciación 454, secuenciación de molécula simple con Heliscope, secuenciación en tiempo real de una sola molécula (SMRT) y las plataformas de secuenciación de semiconductores de iones son ejemplos de métodos de secuenciación de ADN basados en el principio de "secuenciación por síntesis". En estos métodos, la secuencia de una cadena molde de ADN se determina a través de la detección de señales que se emiten a medida que las bases de nucleótidos se incorporan en una cadena complementaria recién sintetizada.

Las plataformas de secuenciación de ADN tienen tasas de error. Por ejemplo, ocasionalmente, la polimerasa utilizada en la reacción de amplificación incorporará la base de nucleótido incorrecta en la cadena complementaria que se sintetiza, lo que lleva a una determinación incorrecta del nucleótido en esa posición en el molde de ADN. El límite de detección de los métodos NGS se define por los errores en dos etapas: la preparación de la biblioteca (que generalmente implica la amplificación por PCR) y por la secuenciación de la misma.

Esto plantea problemas, especialmente para la detección de variantes genéticas que solo estarán presentes en una muestra de ADN a baja frecuencia, por ejemplo, una frecuencia que se aproxima o es inferior a la tasa de error del método de secuenciación utilizado. En estas circunstancias, es difícil o imposible determinar si una variante genética identificada es real (es decir, realmente presente en la molécula de molde de ADN) o es un error.

Para la secuenciación con Illumina, la tasa de error de fondo varía para diferentes variantes genéticas y localizaciones genómicas y tiene una gran varianza. Por lo tanto, detectar las mutaciones que están presentes en una muestra de ADN a una frecuencia de ~ 1% o menos plantea problemas.

Los métodos existentes de secuenciación de ADN e identificación de variantes genéticas tienen limitaciones con respecto a la detección de variantes raras y nuevas en múltiples regiones, especialmente en muestras que tienen pequeñas cantidades de ADN.

Los métodos son normalmente incapaces de identificar mutaciones que ocurren a una frecuencia menor o similar a la tasa de error del método utilizado (es decir, ruido de fondo).

La PCR digital (dPCR; Vogelstein B., Kinzler K.W. Proc. Natl. Acad. Sci. U.S.A. 1999 96(16):9236-41; Sykes, P.J. et al., BioTechniques 1992 13(3): 444-9) no es útil para la identificación de nuevas variantes genéticas (es decir, no identificadas previamente), ya que la dPCR implica el uso de cebadores y ensayos diseñados para detectar variantes particulares. Por otra parte, la dPCR tiene un alcance limitado para analizar múltiples regiones de interés en paralelo, especialmente cuando la muestra de ADN es limitada.

Existen otros métodos complejos para etiquetar moléculas de ADN individuales de un conjunto único de ADN, como Safe-SeqS y sondas de inversión molecular de molécula única (Kinde I, et al., Proc Natl Acad Sci USA. 2011 108(23): 9530-5; Hiatt JB, et al., Genome Res. 2013 23(5):843-54.).

Estos métodos no son adecuados para el análisis simultáneo de múltiples genes (es decir, múltiples regiones de interés) y cuando el ADN es limitado.

Varios estudios han demostrado la detección no invasiva del ADN del cáncer (Dawson SJ, et al., N Engl J Med. 2013 368(13):1199-209; Forshew T, et al., Sci Transl Med. 2012 4(136):136ra68; Murtaza M, et al. Nature. 2013 497 (7447):108-12). Sin embargo, los principales desafíos persisten en este campo, tales como (a) seleccionar bases

suficientes del genoma para detectar mutaciones relevantes en el cáncer (b) seleccionar pequeñas cantidades de ADN fragmentado en busca de tales mutaciones, y (c) detectar moléculas de ADN tumorales mutantes de baja frecuencia entre muchas moléculas de 'tipo silvestre'.

Por ejemplo, Forshew T, et al., Sci Transl Med. 2012 4(136);136ra68 describe el análisis de grandes regiones del genoma para detectar mutaciones de cáncer en la sangre, pero el límite de detección para este método fue de ~ 1%-2% de la frecuencia alélica (AF).

Sumario de la invención

La presente invención proporciona una solución a los problemas anteriores. Las variantes genéticas reales se identifican determinando las frecuencias de fondo para las variantes genéticas en cada posición teniendo en cuenta el error del método (es decir, las tasas de error de la amplificación de ADN y de la plataforma de secuenciación).

La invención es como se define en las reivindicaciones.

En un primer aspecto, la presente divulgación proporciona un método para detectar una variante genética en una región de interés en una muestra de ADN que comprende

(i) determinar, para una plataforma de secuenciación dada, proceso de secuenciación y profundidad de secuenciación, la distribución del número de lecturas que soporta una variante genética o una pluralidad de variantes que se espera observar en los resultados de secuenciación de las reacciones de amplificación debidas al error de amplificación y secuenciación (distribución del recuento de lectura);
(ii) basándose en la distribución del recuento de lectura determinada en el paso (i), establecer una frecuencia umbral igual o superior a la cual se debe observar la variante genética en los resultados de la secuenciación de las reacciones de amplificación para asignar una determinación de positivos de la presencia de variante genética en una reacción de amplificación dada;
(iii) dividir la muestra de ADN en una pluralidad de reacciones de doble amplificación de modo que el número medio de moléculas de molde amplificables de la región de interés en una reacción de doble amplificación sea menor que el recíproco de la frecuencia umbral determinada en el paso (ii);
(iv) realizar las reacciones de amplificación del paso (iii) y secuenciar los productos de las reacciones de amplificación,
(v) basándose en el paso (ii) y en los resultados del paso (iv), determinar la presencia/ausencia de la variante genética en cada reacción de doble amplificación; y
(vi) integrar los resultados de (v) para determinar la presencia/ausencia de la variante genética en la región de interés en la muestra de ADN.

La presente divulgación proporciona un método para detectar una variante genética en una región de interés en una muestra de ADN que comprende

(i) determinar, para una plataforma de secuenciación dada, la frecuencia media y la varianza de la frecuencia a la que se espera observar una variante genética o una pluralidad de variantes genéticas en los resultados de la secuenciación de las reacciones de amplificación debidas al error de amplificación y de secuenciación;
(ii) basándose en la frecuencia media y la varianza de la frecuencia determinadas en el paso (i), establecer una frecuencia umbral igual o superior a la cual se debe observar la variante genética en los resultados de la secuenciación de las reacciones de amplificación para asignar una determinación de positivos de la presencia de variante genética en una reacción de amplificación dada;
(iii) dividir la muestra de ADN en una pluralidad de reacciones de doble amplificación de modo que el número medio de moléculas de molde amplificables de la región de interés en una reacción de doble amplificación sea menor que el recíproco de la frecuencia umbral determinada en el paso (ii);
(iv) realizar las reacciones de amplificación del paso (iii) y secuenciar los productos de las reacciones de amplificación,
(v) basándose en el paso (ii) y en los resultados del paso (iv), determinar la presencia/ausencia de la variante genética en cada reacción de doble amplificación; y
(vi) integrar los resultados de (v) para determinar la presencia/ausencia de la variante genética en la región de interés en la muestra de ADN.

Ventajosamente, el método permite la detección de variantes genéticas a muy baja frecuencia dentro de una muestra de ADN. En consecuencia, el método permite una identificación más temprana (por ejemplo, en el contexto de una patología de la enfermedad, como las mutaciones relacionadas con el cáncer) de la presencia de mutaciones genéticas (es decir, cuando la mutación está presente a una frecuencia más baja) en comparación con los métodos previamente conocidos para la detección de mutaciones. Por lo tanto, el método encuentra uso en varias aplicaciones, como la detección de la reaparición del tumor, el crecimiento tumoral, el desarrollo de resistencia a fármacos/evolución del tumor y la identificación de mutaciones terapéuticamente accionables. El método también permite la identificación de la presencia de mutaciones genéticas a bajas frecuencias con una mejor confianza estadística.

Una característica clave del método de la invención es la división de la muestra de ADN en una pluralidad de reacciones de doble amplificación (por ejemplo, por dilución y división en partes alícuotas en pocillos) de modo que, en promedio, el número de moléculas de molde amplificables de la región de interés presente en una reacción de doble amplificación es menor que el recíproco de la frecuencias umbral para una determinación de positivos. De este modo, para cada reacción de amplificación que tiene una región molde amplificable en la que está presente la variante genética, existe una alta probabilidad de que la variante se observe a una frecuencia mayor que la frecuencia umbral para determinar la presencia de la variante. De este modo, el método permite detectar la presencia de variantes genéticas en una muestra de ADN, incluso cuando está presente a muy baja frecuencia dentro de la muestra de ADN. Por ejemplo, el método permite la detección a una frecuencia inferior al 1%.

Por otra parte, el presente método es útil en el descubrimiento de variantes genéticas. Es decir, para una región de interés, pueden determinarse fácilmente los niveles de fondo para cada uno de los tres cambios de base potenciales de una secuencia de ADN de referencia y pueden identificarse posteriormente las frecuencias para la variante genética por encima de los niveles de fondo (en muestras de ADN particulares) utilizando el método. Por lo tanto, sin seleccionar ninguna variante genética particular, se pueden identificar nuevas variantes genéticas en cualquier posición dentro de la región de interés. Esta característica está asociada con claras ventajas en una amplia gama de campos, incluidos diagnósticos, pronósticos, descubrimiento de mutaciones, control de la respuesta al tratamiento y resistencia a fármacos.

En algunas realizaciones, para que la variante genética se determine como presente en la región de interés en la muestra de ADN en el paso (vi), se debe realizar una determinación de positivos de la presencia de la variante genética en más de una reacción de doble amplificación en el paso (v). En algunas realizaciones, se debe realizar una determinación de positivos de la presencia de la variante genética en al menos 3 reacciones de doble amplificación.

En algunas realizaciones, la variante genética puede ser una variante de un único nucleótido, es decir, un cambio de un nucleótido a un nucleótido diferente en la misma posición. En algunas realizaciones, la variante genética puede ser una inserción o deleción, que agrega o elimina nucleótidos. En algunas realizaciones, la variante genética puede ser una combinación de múltiples eventos que incluyen variantes de un único nucleótido e inserciones y/o deleciones. En algunas realizaciones, una variante genética puede estar compuesta de múltiples variantes genéticas presentes en diferentes regiones de interés.

Requerir una determinación de positivos para la variante genética en una pluralidad de reacciones de doble amplificación reduce la probabilidad de una determinación de positivos falsos de la variante genética que está presente en la muestra de ADN. El método que requiere múltiples determinaciones de positivos en reacciones de doble amplificación, por lo tanto, tiene mayor especificidad para la detección de variantes genéticas.

La frecuencia media y el coeficiente de variación (CV) a los que se observa una variante determinada (es decir, en los resultados de la secuenciación) como resultado del error en el método utilizado para secuenciar una muestra de ADN, se pueden usar para determinar y/o modelar los niveles de fondo (es decir, ruido) para una variante genética. Estos valores se pueden utilizar, por ejemplo, para determinar los valores de la función de distribución acumulada (CDF) y/o para calcular las puntuaciones z.

A su vez, las mediciones y/o modelos de ruido de fondo para una variante genética pueden utilizarse para establecer frecuencias umbral por encima de las cuales debe observarse que una variante genética se determina que está presente en una reacción de amplificación dada (una determinación de positivos). Para una determinación de positivos, la frecuencia de la variante debe ser mayor que la frecuencia media en los niveles de fondo.

En algunas realizaciones del método de la invención, la frecuencia umbral del paso (ii) se determina utilizando un modelo de distribución de probabilidad binomial, binomial sobredispersa, Beta, Normal, Exponencial o Gamma. En algunas realizaciones, la frecuencia umbral en la que se debe observar que una variante genética dada se determina en o por encima de la presencia de una reacción de doble amplificación es la frecuencia a la que el valor de la función de distribución acumulada (CDF) de esa variante genética alcanza un valor umbral predefinido (CDF_thresh) de 0,99, 0,995, 0,999, 0,9999, 0,99999 o mayor.

En algunas realizaciones del método de la invención, la frecuencia umbral del paso (ii) se determina utilizando un corte de puntuación z. En la presente memoria, la frecuencia media de fondo y la varianza de la frecuencia para la variante genética determinada en el paso (i) se modelan con una distribución Normal y la frecuencia umbral para identificar una mutación es la frecuencia en la puntuación z, que es un número de desviaciones estándar por encima de la frecuencia media de fondo. En algunas realizaciones, la frecuencia umbral es la frecuencia en la puntuación z de 20. En algunas realizaciones, la frecuencia umbral es la frecuencia en la puntuación z de 30.

En algunas realizaciones, establecer una frecuencia umbral igual o superior a la cual se debe observar la variante genética en los resultados de la secuenciación de las reacciones de amplificación para asignar una determinación de positivos de la presencia de la variante genética en una reacción de amplificación dada comprende

(a) basándose en la distribución de recuento de lectura determinada para una pluralidad de variantes genéticas,

que es opcionalmente una distribución normal definida por la frecuencia media y la varianza de la frecuencia determinada para una pluralidad de variantes genéticas, en el paso (i), establecer una pluralidad de frecuencias umbral en o por encima de las cuales deben observarse las variantes genéticas en los resultados de la secuenciación de las reacciones de amplificación para asignar una determinación de positivos para la presencia de la variante genética en una reacción de amplificación dada y
(b) basándose en el paso (a), establecer una frecuencia umbral general en o por encima de la cual se debe observar una variante genética en los resultados de la secuenciación de una reacción de amplificación dada para asignar una determinación de positivos de la presencia de la variante genética en esa reacción de amplificación, que es la frecuencia umbral a la cual el 90%, 95 %, 97,5 %, 99% o más de las frecuencias umbral determinadas en el paso (a) son menores que este valor.

De este modo, las frecuencias umbral no necesitan determinarse para cada posible base en cada posición de la región de interés y se puede usar un umbral global basado en una pluralidad de variantes genéticas en el método de la invención.

En algunas realizaciones, el número medio de moléculas de molde amplificables de la región de interés que estará presente en una reacción de doble amplificación dada se determina de tal manera que cuando la variante genética está presente en una molécula de molde amplificable única de una reacción de doble amplificación, la probabilidad de que se realice una determinación de positivos para esa réplica es 0,9 o mayor.

Esto minimiza la probabilidad de que una reacción de doble amplificación dada se determine incorrectamente como negativa para una variante genética cuando la variante está de hecho presente.

En algunas realizaciones, toda la pluralidad de reacciones tiene la misma cantidad de material de molde inicial de la muestra. En algunas realizaciones, las diferentes reacciones tienen diferentes cantidades de material de molde de la muestra. Estas cantidades pueden considerarse al estimar la frecuencia alélica de las variantes genéticas en la muestra.

En algunas realizaciones, el paso (i) comprende la secuenciación de una muestra de ADN varias veces, para determinar la distribución del recuento de lectura para una variante genética o una pluralidad de variantes genéticas, que es opcionalmente una distribución normal definida por la frecuencia media y la varianza de la frecuencia para una variante o pluralidad de variantes genéticas.

Ventajosamente, cuando las tasas de error de fondo se determinan empíricamente para las variantes genéticas en una región de interés de esta manera, la estimación de la tasa de error de fondo será más precisa, lo que permitirá una mayor sensibilidad para detectar la presencia de una variante genética en la muestra de ADN y con un mayor grado de confianza.

En algunas realizaciones, la distribución del recuento de lecturas en la que se espera observar una variante genética o una pluralidad de variantes genéticas en los resultados de la secuenciación de las reacciones de amplificación debidas al error de amplificación y de secuenciación (opcionalmente, una distribución normal definida por la frecuencia media y la varianza de la frecuencia a la que se espera observar una variante genética o una pluralidad de variantes genéticas en los resultados de la secuenciación de las reacciones de amplificación debidas al error de amplificación y de secuenciación); se determina en el paso (i) en función de las tasas de error del secuenciador y/o de la polimerasa. En algunas realizaciones, esto se determina teniendo en cuenta el contexto de la secuencia.

En algunas realizaciones, el paso (i) comprende 'buscar' un valor de referencia o una pluralidad de valores de referencia en una base de datos, gráfico, tabla, lista, catálogo, índice, directorio o registro. En algunas realizaciones, los valores de referencia se determinan mediante la secuenciación de una muestra de ADN de referencia varias veces, para determinar la distribución del recuento de lectura para una variante genética o una pluralidad de variantes genéticas, que es opcionalmente una distribución normal definida por la frecuencia media y la varianza de la frecuencia para una variante o pluralidad de variantes genéticas. En algunas realizaciones, la muestra de ADN de referencia es una muestra 'normal emparejada'.

Ventajosamente, el método de la invención se puede realizar sin tener que determinar empíricamente las tasas de error de fondo cada vez.

En algunas realizaciones, después de dividir la muestra de ADN en el paso (iii), cada reacción de doble amplificación tiene una media de más de una molécula de molde amplificable única para la región de interés por reacción de doble amplificación. En algunas realizaciones, el número medio de moléculas de molde amplificables por reacción de doble amplificación será más de 1 y menos de 1000, más de 2 y menos de 1000, o más de 5 y menos de 1000.

Ventajosamente, la muestra de ADN no necesita ser dividida más de lo necesario para lograr las eficacias del método asociado con la división de moléculas de molde. Esto tiene ventajas al menos en términos de reducir los costes de funcionamiento y materiales.

En algunas realizaciones, el método es capaz de detectar variantes genéticas que están presentes dentro de la población de moléculas de molde amplificables de una muestra de ADN a una frecuencia inferior al 2%, inferior al 1 %, inferior al 0,5 %, inferior al 0,1 % o inferior al 0,05 %.

En algunas realizaciones, la división de la muestra de ADN en una pluralidad de reacciones de doble amplificación comprende diluir la muestra de ADN y dividir en partes alícuotas en reacciones de doble amplificación. Por ejemplo, las reacciones de amplificación se pueden realizar en pocillos separados. Como alternativa, las reacciones de doble amplificación pueden dividirse por otros medios conocidos por los expertos en la materia.

En algunas realizaciones del método de la invención, las reacciones de doble amplificación se realizan en paralelo, y la secuenciación de los productos de las reacciones de amplificación se realiza en paralelo.

En algunas realizaciones de la invención, el método comprende además:

(vii) determinar la frecuencia de la variante genética en la muestra de ADN.

Ventajosamente, esto permite la determinación de cambios en las frecuencias de las variantes genéticas a lo largo del tiempo (por ejemplo, a lo largo del curso de la enfermedad y/o durante el curso del tratamiento) y/o diferencias en las frecuencias de las variantes genéticas entre las muestras.

Las reacciones de amplificación se pueden realizar mediante PCR de un paso o mediante PCR de dos pasos.

En algunas realizaciones, las reacciones de amplificación se realizan utilizando uno o más pares de cebadores que flanquean la región de interés que integran la muestra y/o secuencias identificadoras específicas de la reacción de doble amplificación en los productos de amplificación. Las secuencias identificadoras pueden definirse como cualquier serie de bases de ADN que sea lo suficientemente diferente de otra serie de bases de ADN, de manera que cuando se lee junto con una región de interés diana adjunta, el identificador se puede usar para identificar de qué muestra y/o reacción de amplificación se originó la secuencia diana. Las expresiones "secuencias identificadoras" y "códigos de barras" se usan indistintamente en este documento.

En algunas realizaciones, los cebadores integran adaptadores de secuencia en los productos de amplificación. En algunas realizaciones, los cebadores que flanquean la región de interés comprenden la muestra y/o secuencias identificadoras específicas de reacción de doble amplificación y adaptadores de secuencia, permitiendo que estas secuencias identificadoras y adaptadoras se agreguen durante una PCR de un solo paso. En algunas realizaciones, las secuencias de "etiquetado" universales se incluyen dentro de los pares de cebadores. Una secuencia de etiquetado puede definirse como cualquier serie de bases de ADN que se pueden usar como diana para rondas posteriores de PCR. La unión de secuencias de etiquetado comunes a una pluralidad de cebadores dará como resultado, después de la amplificación, una pluralidad de productos con estas secuencias de etiquetado comunes adjuntas. Una secuencia de etiquetado óptima será lo suficientemente diferente del genoma de interés para evitar la amplificación no específica de la secuencia genómica. En algunas realizaciones, la secuencia de etiquetado puede comprender características adicionales tales como un sitio de unión para cebadores de secuenciación. En algunas realizaciones, se realiza una segunda ronda de PCR, utilizando cebadores que comprenden las secuencias de etiquetado, adaptadores de secuencia y opcionalmente códigos de barras adicionales, para adjuntar adaptadores de secuenciación y opcionalmente códigos de barras adicionales al producto de PCR original. En algunas realizaciones, la ligadura se usa para unir adaptadores de secuencia y, opcionalmente, códigos de barras adicionales al producto de PCR original. En algunas realizaciones, se analizan en paralelo una pluralidad de regiones de interés.

En algunos casos, el método puede implementarse como un método de alto rendimiento, permitiendo el análisis simultáneo de múltiples regiones de interés para detectar variantes genéticas. Esto tiene ventajas en términos de velocidad, eficacia y reducción de costes de funcionamiento y materiales.

En algunas realizaciones del método de la invención, la distribución del recuento de lectura puede definirse como una distribución normal caracterizada por parámetros que son la frecuencia media y la varianza de la frecuencia a la que se observa o se espera que se observe una variante genética o una pluralidad de variantes genéticas en los resultados de secuenciación de las reacciones de amplificación debido al error de amplificación y secuenciación.

La distribución del recuento de lectura puede definirse como el conjunto general de probabilidades para observar el alelo que no sea de referencia en cualquier recuento.

La distribución del recuento de lecturas puede ser la distribución del número de lecturas que soportan una variante genética o una pluralidad de variantes que se espera observar en los resultados de secuenciación de las reacciones de amplificación debidas al error de amplificación y secuenciación. Una lectura que soporta una variante genética es una lectura positiva para una variante genética.
Por ejemplo, la distribución del recuento de lectura puede ser el conjunto general de probabilidades para observar la variante genética o la pluralidad de variantes genéticas, para una plataforma de secuenciación dada, proceso de secuenciación y profundidad de secuenciación, debido a errores de amplificación y secuenciación.

La profundidad de lectura, también denominada profundidad de secuenciación, se puede definir como el número de veces que se lee una posición genómica específica (por ejemplo, un nucleótido dado) durante el proceso de secuenciación.

El método de la invención encuentra uso en una amplia variedad de aplicaciones. De hecho, el método es útil para la detección de variantes genéticas en cualquier región de interés en cualquier muestra de ADN, para cualquier propósito.

En otro aspecto, la presente invención proporciona un método para detectar y/o cuantificar ADN tumoral en una muestra. En algunas realizaciones, el método es para detectar y/o cuantificar ADN tumoral circulante en una muestra.

En algunos casos, de acuerdo con cualquier aspecto de la presente invención, la muestra es una muestra biológica obtenida de un sujeto. En algunas realizaciones, la muestra es una muestra de tejido, por ejemplo una muestra quirúrgica. En algunas realizaciones, la muestra es una biopsia líquida, tal como sangre, plasma, orina, fluido seminal, deposiciones, esputo, fluido pleural, fluido de ascites, fluido sinovial, líquido cefalorraquídeo, linfa, líquido del pezón, o lavado bronquial. En algunas realizaciones, la muestra es una muestra citológica o frotis o un líquido que contiene material celular, como frotis cervical, cepillado nasal o toma de muestras esofágicas con una esponja (citoesponja), biopsia o cepillado endoscópico/gastroscópico/colonoscópico, moco cervical o cepillado.

Muchas de las muestras anteriores se pueden obtener de forma no invasiva y, por lo tanto, se pueden tomar regularmente sin grandes riesgos ni molestias para el sujeto.

En consecuencia, en un aspecto, la presente invención proporciona un método para el diagnóstico o pronóstico *in vitro* o para controlar una enfermedad.

El método se puede utilizar para analizar muestras de ADN para detectar variaciones genéticas asociadas con o predictivas de susceptibilidad, resistencia o respuesta a una terapia determinada.

En consecuencia, en otro aspecto, la presente divulgación proporciona un método para identificar a un sujeto que tiene una mayor probabilidad de desarrollar una enfermedad (es decir, tener una mayor susceptibilidad a una enfermedad), comprendiendo el método: obtener una muestra que contiene ADN del sujeto y llevar a cabo el método de acuerdo con el primer aspecto de la divulgación sobre la muestra que contiene ADN obtenida del sujeto. En algunos casos, de acuerdo con este aspecto, se determina que la muestra que contiene ADN comprende una variante genética asociada a la enfermedad, identificando así que un sujeto tiene una mayor probabilidad de desarrollar una enfermedad (en comparación con un sujeto que no tiene una variante genética asociada a la enfermedad detectada en una muestra que contiene ADN).

En otro aspecto, la presente divulgación proporciona un método de diagnosis, comprendiendo el método: obtener una muestra que contiene ADN de un sujeto y llevar a cabo el método de acuerdo con el primer aspecto de la divulgación sobre la muestra que contiene ADN obtenida del sujeto. En algunos casos, de acuerdo con este aspecto, se determina que la muestra que contiene ADN comprende una variante genética asociada a la enfermedad.

Un sujeto o una pluralidad de sujetos que pueden tener una susceptibilidad, resistencia o respuesta particular a una terapia dada pueden analizarse para determinar la variación genética predictiva de dicha susceptibilidad, resistencia o respuesta. En consecuencia, en un aspecto, la presente divulgación proporciona el método descrito anteriormente en el presente documento en un método para identificar la variación genética predictiva de susceptibilidad, resistencia o respuesta a una terapia.resistencia o respuesta a una terapia.

En consecuencia, la presente divulgación proporciona un método de selección de un paciente para terapia, comprendiendo el método: obtener una muestra que contiene ADN del paciente y llevar a cabo el método de acuerdo con el primer aspecto de la divulgación sobre la muestra que contiene ADN obtenida del paciente. En algunos casos, de acuerdo con este aspecto, se determina que la muestra que contiene ADN comprende una variante genética asociada a la enfermedad (por ejemplo, un alelo de riesgo) y/o una variante genética predictiva de resistencia o respuesta a una terapia. El método puede comprender además seleccionar al paciente para la terapia basándose en la determinación de que la muestra que contiene ADN del paciente comprende una variante genética. En algunos casos, el método puede comprender además una etapa de administrar una terapia y/o recomendar la administración de una terapia basándose en la determinación de que la muestra que contiene ADN del paciente comprende una variante genética.

El método también se puede usar para identificar determinantes nuevos (es decir, no identificados previamente) y/o predictores de susceptibilidad o pronóstico de una enfermedad, o de resistencia o respuesta a una terapia.
En consecuencia, en otro aspecto, la presente divulgación proporciona un método para identificar una variante genética asociada a la enfermedad, comprendiendo el método: obtener una muestra que contiene ADN de un paciente o una pluralidad de pacientes que tienen una enfermedad, y llevar a cabo el método de acuerdo con el primer aspecto de la divulgación sobre la muestra que contiene ADN obtenida del paciente o de la pluralidad de pacientes. En algunos casos, de acuerdo con este aspecto, se determina que la muestra o muestras que contienen ADN comprenden una variante genética, que de este modo se identifica como una variante genética asociada a la enfermedad. De acuerdo

con este aspecto, en algunas realizaciones, el paciente o la pluralidad de pacientes pueden tener un fenotipo clínico particular y/o pronóstico de la enfermedad.

En consecuencia, en algunas realizaciones, el método identifica determinantes o predictores de un fenotipo clínico particular y/o pronóstico de la enfermedad.

En otro aspecto, la presente divulgación proporciona un método para identificar una variante genética asociada con la resistencia o respuesta a una terapia, comprendiendo el método: obtener una muestra que contiene ADN de un paciente o una pluralidad de pacientes que tienen resistencia a o respuesta a una terapia para una enfermedad, y llevar a cabo el método de acuerdo con el primer aspecto de la divulgación en la muestra que contiene ADN obtenida del paciente o pluralidad de pacientes. En algunos casos, de acuerdo con este aspecto, se determina que la muestra o muestras que contienen ADN comprenden una variante genética, que de este modo se identifica como una variante genética asociada con la resistencia o respuesta a una terapia.

En otro aspecto, la presente divulgación proporciona métodos para controlar o evaluar la respuesta a una terapia, en donde el método de acuerdo con el primer aspecto de la divulgación se realiza en muestras de ADN obtenidas de un paciente en diferentes etapas de la terapia (por ejemplo, antes de la intervención y durante/post intervención). La comparación de la variación genética en muestras tomadas en diferentes momentos puede revelar cambios en la respuesta a la terapia. Por ejemplo, la variación genética o la frecuencia relativa de una variante genética en muestras obtenidas durante o después de la terapia y ausentes de las muestras obtenidas antes de la terapia pueden reflejar, p. ej. evolución del tumor y/o carga tumoral, u otras respuestas biológicas a la terapia.

En otro aspecto, la frecuencia de las variantes genéticas, o la proporción de frecuencias de diferentes variantes, se puede utilizar para predecir, controlar o evaluar la respuesta a la terapia. En otro aspecto, la presencia o la frecuencia de las variantes genéticas, o la proporción de frecuencias de diferentes variantes, se puede usar para predecir el nivel de riesgo o el pronóstico de un paciente si el paciente no recibe tratamiento o si se le administra uno de un conjunto de tratamientos.

Breve descripción de las figuras

Figura 1. Gráficos que muestran las distribuciones de los valores de ruido de fondo del proceso de secuenciación del amplicón. (A) la media y (B) el coeficiente de variación (CV, definido como desviación estándar dividida por la media), para diferentes variantes de un solo nucleótido observadas en un panel de secuenciación del amplicón.

Figura 2. Gráficos que muestran las distribuciones acumulativas para varios cambios de base y las funciones de distribución acumulativa para la distribución Normal y la distribución Beta, ajustadas utilizando los valores de la media y el CV que se calcularon para esos cambios de base, en comparación con las distribuciones empíricas de los datos.(A) EGFR_D0016_006_R 55248959 C> G. (B) EGFR_D0016_009_F 55249159 G> T. (C) EGFR_D0016_003_R 55241757 C> A. (D) EGFR_D0016_012_R 55259573 G> C.

Figura 3. Gráficos que muestran los valores acumulativos de las Frecuencias Mínimas Detectables (MDF) para diferentes valores de CDF_thresh. (B) Acercamiento de (A). La línea vertical indica el punto de percentil 97,5, para el cual el 97,5% de los posibles cambios de base tienen valores de MDF más bajos. Para CDF_thresh = 0,9999, el percentil 97,5 está en la frecuencia del alelo (AF) = 0,0382, indicado por una línea horizontal.

Figura 4. Las Tablas 2 y 3, que muestran los resultados de la prueba de aplicación de principio del método en una serie de dilución de líneas celulares.

Figura 5. Gráfico que muestra la probabilidad de detección de una mutación compuesta en múltiples reacciones, mostrando la probabilidad de Poisson de encontrar al menos N número de moléculas en la muestra en función del número esperado de moléculas (según la concentración y la cantidad inicial de material), las diferentes líneas muestran una N diferente, de modo que una identificación compuesta se realiza con ≥N reacciones positivas.

Figura 6. Resumen de una estrategia de preparación de biblioteca a modo de ejemplo. Los cebadores TAm-Seq se generan con múltiples códigos de barras diferentes. Cada combinación de código de barras se distribuye en un pocillo de PCR separado. A) La primera PCR amplifica la región de interés y agrega etiquetas identificadoras moleculares (códigos de barras) que identifican la reacción de las otras reacciones de doble amplificación que tienen los mismos cebadores específicos de diana, pero diferentes códigos de barras. B) Todos los productos de PCR de una muestra se agrupan y una segunda ronda de PCR incluye adaptadores específicos del secuenciador y códigos de barras específicos de la muestra.

Figura 7. Representación de combinación de los códigos de barras de cebadores específicos de diana. Cada cebador específico de diana se sintetizó 7 veces con 7 códigos de barras diferentes. Los cebadores directo e inverso se combinaron para producir 49 combinaciones diferentes de códigos de barras de pocillo mezclando cada uno de los 7 cebadores directos con cada uno de los 7 cebadores inversos.

Figura 8. Gráfico que muestra la relación entre la probabilidad de identificación y el número medio de moléculas por reacción. Para diferentes valores de MDF (como se indica en la leyenda de la figura), la probabilidad de AF> MDF se representa en función del número promedio de moléculas por reacción en función de la dilución del molde. La línea horizontal muestra probabilidad = 0,90. Para MDF = 0,0382, una probabilidad de 0,90 de obtener AF> MDF en un N extrapolado = 20,59 (indicado por la línea vertical).

Figura 9. Gráficos que muestran las mutaciones identificadas (verdaderos positivos (TP)), identificaciones de falsos positivos (FP) y mutaciones perdidas (falsos negativos (FN)) al realizar el método en una serie de dilución de ADN de tres líneas celulares. (A) Corte de CDF = 0,9999 en una distribución Beta con N ≥2. (b) Corte de CDF = 0,9999 en una distribución Beta con N ≥3.

Figura 10. Gráfico que muestra la concordancia entre la AF esperada y medida. Los negativos falsos se muestran como "No detectado" (ND) en el eje vertical. Los valores esperados se muestran como cruces, conectados por una línea de puntos como una guía para el ojo.

Descripción detallada de la invención

El método de la invención utiliza la división de la muestra de ADN en réplicas dentro de las cuales una variante dada, si está presente, tendrá una frecuencia mayor que en la muestra de ADN total, combinada con la detección de la mutación en múltiples pocillos/reacciones, lo que permite a una variante ser "identificada" sobre las tasas de error inherentes en el método utilizado para la determinación de la secuencia; es decir, la plataforma de secuenciación.

Las tasas de error del secuenciador y la polimerasa se pueden determinar de varias maneras bien conocidas por los expertos en la materia (Tindall, K.R., Kunkel, T.A. Biochemistry 1988 27(16): 6008-13; Forshew, T. et al. Sci. Transl. Med. 2012 4(136):136ra68). A menudo, estos se publican y/o están disponibles por parte de los fabricantes. Las tasas de error se pueden usar para determinar la frecuencia esperada en la que se observará una variante genética debido al ruido de fondo.

El error del método también se puede introducir durante, por ejemplo, la preparación de la biblioteca, en el caso de los métodos NGS.

A menudo, el error difiere dependiendo de la posición de la base en la molécula que se está secuenciando (véase, por ejemplo, Loman NJ et al., Nature Biotechnology 30: 434-439, 2012; Forshew, T. et al. Sci. Transl. Med. 2012 4(136):136ra68). Una forma de abordar esto es mediante el modelado del ruido usando parámetros como el contexto de secuencia (Ross, MG et al., Genome Biology 2013 14-.R51), sin embargo, esto no puede proporcionar un valor preciso para cada cambio posible en cada locus y no tiene en cuenta los cambios en las propiedades del sistema de secuenciación o el proceso de amplificación a lo largo del tiempo.

Las frecuencias de fondo para variantes genéticas en una base particular también pueden determinarse empíricamente, secuenciando la región de interés para una muestra de ADN de referencia varias veces. Por ejemplo, la muestra de ADN de referencia se puede obtener de sangre total o plasma de sujetos sanos o de una línea celular. Los resultados de la secuenciación de la región de interés de la muestra de referencia se pueden usar varias veces para establecer frecuencias de fondo para cada una de las cuatro posibles bases (A, G, T y C) en cada posición de la secuencia de ADN de la región de interés. Esto permite determinar la frecuencia con la que se identifican las variantes genéticas debido al error del método (es decir, las tasas de error de la plataforma de secuenciación de ADN utilizada) para cada posición de la secuencia de ADN de la región de interés. En algunas realizaciones, la muestra de ADN de referencia puede ser una muestra 'normal emparejada'. Por ejemplo, la muestra de ADN de referencia se puede obtener del mismo tejido y/o tipo de muestra, de un sujeto sano.

En algunas realizaciones del método de la invención, la media y la varianza de la frecuencia de la variante genética se determinan para un panel específico de amplicones que cubren la región de interés.

"Región de interés" como se usa en el presente documento significa la porción o porciones del genoma que se está investigando. Puede ser una secuencia única de ADN o una pluralidad de secuencias de ADN. Cuando la región de interés es una pluralidad de regiones, estas pueden extenderse por todo el genoma. Es decir, "región de interés" abarca una pluralidad de secuencias de ADN que se están investigando y se intercalan con porciones del genoma que no se están investigando. En algunos casos, la región de interés dependerá, por ejemplo, de cualquier interrogante clínica que se esté investigando.

Por ejemplo, en algunas realizaciones, la región de interés puede ser una pluralidad de regiones que se sabe que son o candidatas para albergar variantes genéticas asociadas a enfermedades, o variantes genéticas que influyen en la respuesta a la terapia. Por ejemplo, en algunas realizaciones, la región de interés puede ser un panel de genes asociados con el cáncer.

La frecuencia de un nucleótido dado en una posición dada (frecuencia alélica, AF) se determina como la proporción de lecturas para esa posición que identifican ese nucleótido en esa posición, del número total de determinaciones del

nucleótido en esa posición.

La AF para cada nucleótido, en cada posición de la región de interés determinada a partir de cada reacción de secuenciación de la muestra de referencia, se puede usar a continuación, para calcular la media de AF y el coeficiente de variación (CV) de la AF. El CV de la AF se calcula como la desviación estándar de la AF dividida por la media de la AF.

La media de la AF y el CV de la AF se puede usar a continuación para modelar el ruido de fondo (es decir, el error de secuenciación) para una variante genética dada. Por ejemplo, se puede usar una distribución Normal, Beta, Exponencial o Gamma u otra función para modelar el ruido de fondo para una variante genética, dependiendo de qué modelo se ajuste mejor a la distribución empírica de los datos. De forma similar, el número de lecturas de mutantes y las profundidades de secuenciación se pueden modelar con funciones discretas.

Las distribuciones de probabilidad y los métodos de modelado tales como los datos son bien conocidos por los expertos en la materia. Preferentemente, se modelarán varias distribuciones a los datos de la AF y el modelo con el mejor ajuste se seleccionará como el modelo para los pasos posteriores del método de la invención. Por ejemplo, se pueden analizar diferentes modelos para los valores de Kolmogorov-Smirnov para la bondad de ajuste a los datos empíricos.

En los Ejemplos a continuación, se utilizan las distribuciones de probabilidad Beta y Normal.

Umbrales para determinar la presencia de una variante genética

Las distribuciones de probabilidad ajustadas a los valores de error de fondo de la AF y el CV determinados empíricamente, o basadas en el error de fondo previsto en función de las tasas de error del secuenciador y de la polimerasa, la región de interés y la variante genética, se pueden usar para establecer el umbral de la AF en o por encima del cual debe observarse una variante genética en los resultados de la secuenciación de una reacción de doble amplificación que se determinará como presente en la muestra de ADN, denominada en el presente documento como frecuencia mínima detectable (MDF).

Por ejemplo, en el Ejemplo 7 debajo de la AF se seleccionaron puntuaciones z de 20 y 30. Es decir, para una variante genética que se determine como presente en la muestra de ADN en una reacción de doble amplificación dada, la frecuencia de ese alelo tenía que ser 20 o más, o 30 o más desviaciones estándar por encima de la media de fondo de la AF para esa variante en particular.

Dependiendo de la variante genética particular en cuestión, la región de interés, la plataforma de secuenciación utilizada, etc., son útiles varios cortes de puntuación z con el método de la invención. Por ejemplo, el umbral de puntuación z se puede seleccionar entre 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 22, 24, 26, 28, 30, 35, 40, 45, 50, 60, 70, 80, 90 y 100.

Mediante el uso de distribuciones de probabilidad, se puede determinar la probabilidad acumulada de observar una variante genética en particular en una AF dada, cuando la variante está realmente presente en una reacción de doble amplificación. Esto, a su vez, se puede usar para establecer una frecuencia umbral en la que se debe observar una variante genética en o por encima de los productos de una reacción de doble amplificación para determinar que está presente en la muestra de ADN.

La probabilidad de que las supuestas mutaciones observadas en los datos de secuenciación a una frecuencia igual o superior a la MDF sean "reales", en lugar del resultado del error de secuenciación, es muy alta. Por lo tanto, se puede determinar que tales variantes están presentes en una reacción de amplificación dada con alta confianza.

En algunas realizaciones del método de la invención, la frecuencia umbral del paso (ii) se determina utilizando un modelo de distribución de probabilidad Beta de las frecuencias de fondo para las variantes genéticas. En la presente memoria, para cada posible variante genética en cada posición de la región de interés, la frecuencia media de fondo y la varianza de la frecuencia para la variante genética determinada en el paso (i) se utilizan para definir una distribución Beta. La frecuencia umbral en la que se debe observar que una variante genética dada se determina en o por encima de la presencia de una reacción de doble amplificación es la frecuencia a la que el valor de la función de distribución acumulada (CDF) de esa variante genética alcanza un valor umbral predefinido (CDF_thresh). En algunas realizaciones, CDF_thresh es 0,99, 0,995, 0,999, 0,9999, 0,99999 o mayor.

Por ejemplo, en el Ejemplo 6, se usó una probabilidad acumulada (basada en una distribución Beta) de observar una variante genética en particular a una AF dada de 0,9999 como umbral de la AF para una determinación de positivos (es decir, MDF). Es decir, la probabilidad de observar una variante genética dada en los resultados de la secuenciación para una reacción de doble amplificación, donde la variante genética no está presente, es de 0,01% o menos. En algunas realizaciones, la frecuencia umbral es la frecuencia a la cual la función de distribución acumulada para la probabilidad de observar la variante genética en esa frecuencia (CDF_thresh) es 0,8, 0,85, 0,9, 0,95, 0,96, 0,97, 0,98, 0,99, 0,995, 0,999, 0,9995, 0,9999, 0,99995, 0,99999, 0,999995, 0,999999 o mayor, correspondiente a una

probabilidad de observar la variante genética en los resultados de secuenciación para una reacción de amplificación donde la variante genética no está presente en un 20%, 15 %, 10 %, 5 %, 4 %, 3 %, 2 %, 1 %, 0,5 %, 0,1 %, 0,05 %, 0,01 %, 0,005 %, 0,001 %, 0,0005 %, 0,0001 % o menos.

En algunas realizaciones, los valores de la MDF se determinan para una pluralidad de posibles variantes genéticas y se usan posteriormente para establecer una MDF general para una región de interés dada y/o un panel de variantes genéticas de interés. En consecuencia, en algunas realizaciones, establecer una frecuencia umbral igual o superior a la cual se debe observar la variante genética en los resultados de la secuenciación de las reacciones de amplificación para asignar una determinación de positivos de la presencia de la variante genética en una reacción de amplificación dada comprende

(a) basándose en la frecuencia media y la varianza de la frecuencia determinada para la pluralidad de variantes genéticas en el paso (i), establecer una pluralidad de las frecuencias umbral en o por encima de las cuales deben observarse las variantes genéticas en los resultados de la secuenciación de las reacciones de amplificación para asignar una determinación de positivos para la presencia de la variante genética en una reacción de amplificación dada (es decir, estableciendo las MDF para la pluralidad de variantes genéticas), y
(b) basándose en el paso (a), establecer una frecuencia umbral general en o por encima de la cual se debe observar una variante genética en los resultados de la secuenciación de una reacción de amplificación dada para asignar una determinación de positivos de la presencia de la variante genética en esa reacción de amplificación (es decir, estableciendo una MDF general), que es la frecuencia umbral a la cual el 90%, 95 %, 97.5%, o 99% de la pluralidad de frecuencias umbral determinadas en el paso (a) son menores que este valor.

En algunas realizaciones, la MDF global es la frecuencia a la que el 80%, 85 %, 90 %, 91 %, 92 %, 93 %, 94 %, 95 %, 96 %, 97 %, 97,5 %, 98 %, 98,5 %, 99% o 99.5% o más de los valores de MDF determinados para la pluralidad de variantes genéticas son más pequeños (es decir, menos) que la MDF general. En algunas realizaciones, la MDF global se determina basándose en los valores de la MDF establecidos para una pluralidad de variantes genéticas distribuidas en la región de interés. En algunas realizaciones, la pluralidad de variantes genéticas será 5, 10, 20, 25, 30 o más variantes.

El umbral para determinar la presencia de una variante genética se selecciona, junto con el número de copias del molde de ADN de la región de interés en una reacción de doble amplificación dada, de manera que, si está presente, la frecuencia de la variante genética en un la reacción de doble amplificación dada sea significativamente más alta que el nivel de fondo para esa variación genética. De esta manera, la presencia/ausencia de la variante genética en una reacción de amplificación dada puede determinarse con confianza, minimizando la posibilidad de una determinación de positivos falsos.

De forma similar, el umbral para determinar la presencia de una variante genética se elige para minimizar el número de falsas determinaciones negativas.

Números de moléculas de molde amplificables por reacción

El límite inferior en el que se puede identificar una mutación se determina por la frecuencia de fondo de una variante genética (es decir, la frecuencia observada debido al error del método). Es decir, por ejemplo, en un método en el que se realizan determinaciones incorrectas con una frecuencia del 5%, las variantes reales solo pueden distinguirse de los errores si su frecuencia es sustancialmente mayor que el 5%.

El presente método supera este problema al colocar un límite superior en el número promedio de moléculas de molde amplificables de la región de interés en una reacción de doble amplificación dada.

En el método de la presente invención, el número medio de moléculas de molde amplificables de la región de interés (Mol_mean) en una reacción de doble amplificación dada se determina utilizando el umbral para determinar la presencia de una variante genética (MDF), como se establece a continuación:

$$Mol\_mean < 1/MDF$$

Por lo tanto, se esperará que una variante genética, si está presente en una reacción de doble amplificación dada, se represente (como una fracción) dentro del número total medio de copias de molde amplificables de la región de interés dentro de esa reacción de doble amplificación a una frecuencia mayor que la frecuencia umbral para determinar la presencia de la variante genética en una reacción de amplificación dada.
De este modo, una sola molécula que tenga la variante tendrá una alta probabilidad de ser detectada a una frecuencia que sea igual o mayor que la MDF determinada en el paso (ii).

En la presente memoria, la MDF puede ser la MDF calculada para una variante genética determinada, o la MDF general, que es un valor paraguas representativo de una pluralidad de MDF y que se describe en el presente documento anteriormente.

Por ejemplo, si una variante genética dada solo se puede determinar con confianza como presente en una muestra de ADN cuando se observa con una frecuencia de, por ejemplo, el 5% o más (según lo define la MDF), el método requiere que en promedio, estén presentes menos de 20 moléculas en una reacción de doble amplificación dada. De este modo, para cualquier réplica dada donde esté presente la variante genética (y estén presentes hasta 19 moléculas no variantes), la variante se observará a una frecuencia igual o superior al 5%.

En los ejemplos proporcionados a continuación, la distribución de moléculas en reacciones de doble amplificación se modela con una distribución de Poisson. La probabilidad de una determinación de positivos para una molécula amplificable que tenga la variante genética en una reacción de amplificación dada puede determinarse en función de una distribución de Poisson. Como alternativa, la distribución de moléculas en reacciones de doble amplificación podría modelarse con una distribución binomial negativa.

En consecuencia, en algunas realizaciones, El Mol_mean se determina utilizando una distribución de Poisson para modelar el número de moléculas amplificables que estarán presentes en una reacción de doble amplificación dada y el número medio esperado de moléculas amplificables. Para diferentes valores posibles de Mol_mean, se calcula la función de distribución acumulada de la distribución de Poisson. El Mol_mean se selecciona de tal manera que el valor de la función de distribución acumulada para la distribución de Poisson será mayor que un umbral tal como 0,8, 0,85, 0,9, 0,95 o 0,99.

En algunas realizaciones de los métodos de la presente invención la muestra de ADN se divide de manera tal que la cantidad media de moléculas de molde amplificables por reacción es la que se basa en una distribución de moléculas de Poisson esperada en las reacciones, una variante genética, si está presente, se detectará en más del 90% de las veces. Este umbral puede ser mayor o menor y meseta a medida que el número medio de moléculas por reacción disminuye a 1 (ver, por ejemplo, la Figura 8). En algunas realizaciones, la cantidad media de moléculas de molde amplificables por reacción de doble amplificación se selecciona de modo que una variante genética, si está presente, se detectará una de más del 80%, más del 81%, más del 82 %, más del 83 %, más del 84 %, más del 85 %, más del 86 %, más del 87 %, más del 88 %, más del 89 %, más del 90 %, más del 91 %, más del 92 %, más del 93 %, más del 94 %, más del 95 %, más del 96 %, más del 97 %, más del 98% y más del 99% de las veces.

En aras de la eficacia y el uso rentable del tiempo y los recursos, se prefiere que una muestra de ADN no se divida más de lo necesario para lograr las eficacias del método asociado con la división de moléculas de molde. Esto tiene ventajas al menos en términos de reducir los costes de funcionamiento y materiales. Por supuesto, el número mínimo de moléculas por reacción de amplificación dependerá de la variante genética que se determine, la región de interés, el error del método (es decir, las tasas de error de la plataforma de secuenciación del ADN) y los umbrales calculados, etc. En ciertas realizaciones del método de la invención, se prefiere que se proporcione más de una molécula de molde amplificable de la región de interés por reacción de doble amplificación.

Por ejemplo, en algunas realizaciones el número medio de moléculas de molde amplificables por reacción de doble amplificación será más de 1 y menos del Mol_mean.

En algunas realizaciones, el número medio de moléculas de molde amplificables por reacción de doble amplificación es más de 1 y menos de 1000, más de 1 y menos de 750, más de 1 y menos de 500, más de 1 y menos de 250, más de 1 y menos de 200, más de 1 y menos de 150, más de 1 y menos de 100, más de 1 y menos de 80, más de 1 y menos de 60, más de 1 y menos de 50, más de 1 y menos de 40, más de 1 y menos de 35, más de 1 y menos de 30, más de 1 y menos de 29, más de 1 y menos de 28, más de 1 y menos de 27, más de 1 y menos de 26, más de 1 y menos de 25, más de 1 y menos de 24, más de 1 y menos de 23, más de 1 y menos de 22, más de 1 y menos de 21, más de 1 y menos de 20, más de 1 y menos de 19, más de 1 y menos de 18, más de 1 y menos de 17, más de 1 y menos de 16, más de 1 y menos de 15, más de 1 y menos de 14, más de 1 y menos de 13, más de 1 y menos de 12, más de 1 y menos de 11, más de 1 y menos de 10, más de 1 y menos de 9, más de 1 y menos de 8, más de 1 y menos de 7, más de 1 y menos de 6, más de 1 y menos de 5, más de 1 y menos de 4, más de 1 y menos de 3, o más de 1 y menos de 2.

En algunas realizaciones, el número medio de moléculas de molde amplificables por reacción de doble amplificación es más de 2 y menos de 1000, más de 2 y menos de 100, más de 2 y menos de 75, más de 2 y menos de 50, más de 2 y menos de 40, más de 2 y menos de 30, más de 2 y menos de 29, más de 2 y menos de 28, más de 2 y menos de 27, más de 2 y menos de 26, más de 2 y menos de 25, más de 2 y menos de 24, más de 2 y menos de 23, más de 2 y menos de 22, más de 2 y menos de 21, más de 2 y menos de 20, más de 2 y menos de 19, más de 2 y menos de 18, más de 2 y menos de 17, más de 2 y menos de 16, más de 2 y menos de 15, más de 2 y menos de 14, más de 2 y menos de 13, más de 2 y menos de 12, más de 2 y menos de 11, más de 2 y menos de 10, más de 2 y menos de 9, más de 2 y menos de 8, más de 2 y menos de 7, más de 2 y menos de 6, más de 2 y menos de 5, más de 2 y menos de 4, o más de 2 y menos de 3.

En algunas realizaciones, el número medio de moléculas de molde amplificables por reacción de doble amplificación es más de 5 y menos de 1000, más de 5 y menos de 100, más de 5 y menos de 75, más de 5 y menos de 50, más de 5 y menos de 40, más de 5 y menos de 30, más de 5 y menos de 29, más de 5 y menos de 28, más de 5 y menos de 27, más de 5 y menos de 26, más de 5 y menos de 25, más de 5 y menos de 24, más de 5 y menos de 23, más de 5

y menos de 22, más de 5 y menos de 21, más de 5 y menos de 20, más de 5 y menos de 19, más de 5 y menos de 18, más de 5 y menos de 17, más de 5 y menos de 16, más de 5 y menos de 15, más de 5 y menos de 14, más de 5 y menos de 13, más de 5 y menos de 12, más de 5 y menos de 11, más de 5 y menos de 10, más de 5 y menos de 9, más de 5 y menos de 8, más de 5 y menos de 7, o más de 5 y menos de 6.

En algunas realizaciones, el número medio de moléculas de molde amplificables por reacción de doble amplificación está en el rango de 2- 40, 3-30, 4-30, 5-30, 5-25, 10-25, 15-25, o 18-22.

La "división" se puede lograr por cualquier medio adecuado para separar las reacciones de doble amplificación. Por ejemplo, la división se puede lograr dividiendo en partes alícuotas la muestra de ADN (diluida) en pocillos separados. El experto en la materia conocerá otros varios métodos para compartimentar reacciones de doble amplificación separadas (es decir, discretas, individuales o independientes).

Requerir réplicas plurales positivas para una determinación de positivos

Para minimizar la probabilidad de una determinación de positivos falsos de la presencia de una variante genética en una muestra de ADN, el método de la invención requiere que se determine que una variante genética está presente en más de una réplica.

El número medio de moléculas por reacción y el número de identificaciones repetidas necesarias para minimizar las identificaciones de falsos positivos, determinan el número de reacciones a realizar para obtener la sensibilidad requerida.

La probabilidad teórica de que se produzca un falso positivo en N reacciones fuera de las reacciones totales T (P_fpNT) depende de la probabilidad de un falso positivo en una reacción (P_fp1T, que es igual a 1-CDF_thresh) y el número de reacciones, y es igual a

$$P\_fpNT = ((P\_fp1T)^N)*C\_NT = ((1-CDF\_thresh)^N)*C\_NT$$

donde C_NT es el coeficiente combinatorio que depende del número total de reacciones T, como sigue ("!" indica la función factorial)

$$C\_NT = T!/((T-N)!*N!)$$

El número total de falsos positivos esperados de este proceso es igual a la probabilidad de falsos positivos (P_fpNT), multiplicado por el número de variantes examinadas en este proceso. Por ejemplo, cuando se investigan variantes de un solo nucleótido, este será el número de posiciones multiplicado por 3 (para las tres posibles alteraciones que no sean de referencia).

Requerir una determinación de positivos para la variante genética en una pluralidad de reacciones de doble amplificación reduce la probabilidad de una determinación de positivos falsos de la variante genética que está presente en la muestra de ADN. Esto queda claro a partir de los ejemplos experimentales a continuación.

El método que requiere múltiples determinaciones de positivos en reacciones de doble amplificación, por lo tanto, tiene mayor especificidad para la detección de variantes genéticas.

Sin embargo, requerir que una pluralidad de réplicas sea positiva para una variante genética dada también aumenta la probabilidad de que se realice una determinación de negativos falsos. La Figura 5 muestra la probabilidad de Poisson de que al menos N moléculas de molde de ADN estén presentes en la muestra de ADN en función del número esperado de moléculas (según la concentración y la cantidad inicial de material).

El número de reacciones y el número de moléculas variantes en estas reacciones, debe ser lo suficientemente alto para que, dada la distribución aleatoria de las moléculas mutantes, la probabilidad de que al menos el número requerido de réplicas contenga un mutante sea suficientemente alta.

Por lo tanto, el número de réplicas positivas requeridas para que se determine que una variante genética está presente en una muestra de ADN en el paso (vi) se elegirá para minimizar falsos positivos y falsos negativos. Es decir, el número de réplicas positivas requeridas se determinará para lograr una sensibilidad y especificidad del método deseadas. En algunas realizaciones, el número de réplicas positivas será el número que maximice la sensibilidad y la especificidad del método.

El número siempre será mayor que uno. En algunas realizaciones, el número será 2, 3, 4, 5, 6, 7, 8, 9, o 10 réplicas. En ciertas realizaciones, el número será 2 o 3.

El número puede variar según el número y el tamaño de las regiones que se analizan, la cantidad de ADN disponible, el número total de reacciones de doble amplificación etc.

El número también puede variar según el tipo particular de variación del nucleótido de referencia; por ejemplo, si la variante es una transición o una transversión del alelo de referencia.

El número de réplicas positivas que se requieren se puede determinar en parte por el número total de reacciones de doble amplificación para la muestra de ADN. Por ejemplo, en algunas realizaciones, para una variante genética que se determine como presente en una muestra de ADN en el paso (iv), se debe realizar una determinación de positivos para la presencia de la variante en más del 1%, 2 %, 3 %, 4 %, 5 %, 6 %, 7 %, 8 %, 9% o 10% de las réplicas.

Se puede determinar el número óptimo de réplicas positivas (es decir, minimizar falsos positivos y falsos negativos) para una variante y/o región de interés determinadas, por ejemplo, realizando el método de la invención en muestras de ADN de referencia.

El método de la invención requiere la integración de los resultados de las determinaciones de presencia/ausencia para las reacciones de doble amplificación para la determinación final de la presencia/ausencia de la variante genética en la región de interés en la muestra de ADN.

Como se usa en el presente documento, 'integrar' significa combinar o consolidar los resultados de la presencia/determinaciones para las reacciones de doble amplificación.

Relación entre el umbral de identificación de variante, el número de moléculas de ADN molde amplificables y el número de reacciones de doble amplificación

Tal como se describe anteriormente y en los ejemplos experimentales más adelante, será inmediatamente evidente para la persona experta que el umbral de la AF para determinar la presencia de una variante genética en una reacción de doble amplificación, el número de moléculas de ADN molde amplificables que están presentes en una reacción de doble amplificación dada y el número de repeticiones requeridas para ser "positivo" para que se determine que una variante genética está presente en una muestra de ADN están interrelacionados.

Esencialmente, el principio fundamental es que el umbral para determinar la presencia de una variante genética y el número de copias del molde de ADN de la región de interés debe seleccionarse de manera tal que, si está presente, la frecuencia de la variante genética en una reacción de doble amplificación dada sea significativamente más alta que el nivel de fondo para esa variación genética, y que el número de réplicas para ser positivo debe seleccionarse para optimizar la sensibilidad y especificidad del método.

De esta manera, la presencia/ausencia de la variante genética en una reacción de amplificación dada puede determinarse con confianza, minimizando la posibilidad de una determinación de positivos falsos.

La importancia relativa de la sensibilidad y la especificidad para establecer estos parámetros también dependerá de la investigación en particular. Por ejemplo, en los casos en que el método de la invención se utiliza para identificar nuevas mutaciones, la consideración principal puede ser la sensibilidad. Por el contrario, cuando se analiza una muestra de ADN en busca de una variante para comunicar decisiones terapéuticas, la consideración principal puede ser la especificidad.

Secuenciación profunda de Amplicones Etiquetados

El método de la invención es adecuado para su uso con metodologías de secuenciación de ADN de alto rendimiento. Ventajosamente, se pueden analizar múltiples regiones de interés para mutaciones en paralelo.

En los ejemplos experimentales a continuación, se utiliza el método de secuenciación profunda de amplicones etiquetados (TAm-Seq). El método se describe en detalle en Forshew et al. 2012 Sci Transl Med 4 (136) 136ra68. El TAm-Seq permite la amplificación y la secuenciación profunda de regiones genómicas que abarcan miles de bases desde tan solo copias individuales de ADN fragmentado.

En resumen, los cebadores están diseñados para generar amplicones que agrupan regiones de interés en segmentos de un intervalo de tamaño seleccionado en función de la calidad (por ejemplo, la longitud del fragmento promedio) del ADN molde, que incorpora secuencias adaptadoras de secuenciación universales y para etiquetar cada réplica con secuencias de identificación o “código de barras” (véase Figuras 6 y 7). A continuación, los productos se secuencian y las secuencias de identificación se utilizan para demultiplexar las lecturas y alinearlas con el genoma.

Usos del método de la invención

El método es útil para una amplia gama de aplicaciones, que serán inmediatamente evidentes para el experto. Esencialmente, el método es útil para la detección de cualquier variante genética en cualquier muestra de interés.

Por ejemplo, el método de la invención es útil en métodos de diagnóstico y/o pronóstico, o en análisis de microbios o virus.

En particular, el método es útil para la detección de variantes genéticas raras y/o mutaciones raras. El método permite la detección de variantes presentes a una frecuencia inferior al 1% en la muestra de ADN. Además, el método es adecuado para detectar nuevas variantes a tan baja frecuencia, así como para analizar muestras de variantes conocidas.

En consecuencia, el método permite una identificación más temprana de la presencia de variantes genéticas que pueden aparecer en el cuerpo, como en el cáncer o durante el embarazo (es decir, cuando la variante genética está presente en una frecuencia más baja).

En algunos casos, de acuerdo con el método de cualquier aspecto de la presente invención, el método comprende analizar una muestra que se haya obtenido previamente de un sujeto. La muestra puede ser en general cualquier muestra biológica adecuada a partir de la cual se pueda aislar ADN. En algunos casos, la muestra se selecciona del grupo que consiste en: orina, saliva, sangre, suero, heces, otros fluidos biológicos, pelo, células y tejidos.

En algunas realizaciones, de acuerdo con cualquier aspecto de la presente invención, el ADN puede tratarse con bisulfito antes de llevar a cabo el método. El tratamiento con bisulfito del ADN convierte los restos de citosina no metilados en uracilo, pero deja los restos de citosina metilados sin afectar. De este modo, el método es útil para detectar la variación en la metilación del ADN.

La capacidad del método para detectar la presencia de variantes a baja frecuencia permite el análisis de muestras obtenidas por medios no invasivos y/o mínimamente invasivos. Por ejemplo, biopsias líquidas, tal como sangre, plasma, orina, fluido seminal, deposiciones, esputo, fluido pleural, fluido de ascites, fluido sinovial, líquido cefalorraquídeo o muestras de lavado bronquial. Esto tiene la ventaja de que se pueden tomar muestras regularmente sin grandes riesgos ni molestias para el sujeto.

La muestra puede ser nueva o se ha almacenado previamente (por ejemplo, congelada) y/o se puede haber procesado previamente.

Por otra parte, el método es cuantitativo, lo que permite una medición precisa, p.ej. de los niveles de mutantes de ADN circulante tumoral (ADNct) para las correlaciones clínicas o de otros ADN, como el ADN viral o microbiano.

La cuantificación se puede llevar a cabo, por ejemplo, contando el número de moléculas determinadas como que tienen la variante. Esto es adecuado cuando el número de moléculas que tienen la variante es bajo. La cuantificación también se puede realizar mediante la corrección de Poisson del recuento, o modelando las frecuencias alélicas observadas. La cuantificación también se puede realizar teniendo en cuenta las frecuencias alélicas de la variante genética en cada una de la pluralidad de reacciones y las cantidades relativas de molde en cada reacción, por ejemplo, utilizando un promedio o un promedio ponderado por las cantidades de material o considerando la fracción de reacciones en la que se detectó la variante para diferentes cantidades iniciales de material.

Por ejemplo, el método es útil para el control de la carga tumoral, en donde el método se usa para controlar los niveles de mutaciones del ADNct, o para controlar la reaparición de mutaciones, para informar como un indicador temprano del recrecimiento tumoral. El método claramente tiene usos diagnósticos y pronósticos.

El método también es útil para controlar la resistencia al fármaco y/o la evolución del tumor en respuesta a la terapia. El método es particularmente adecuado para la evaluación de un gran número de regiones de interés en paralelo, y se puede utilizar para detectar mutaciones en vías de resistencia probables que podrían ser indicadores tempranos de resistencia a fármacos. Tales usos pueden estar más allá de las capacidades de los métodos anteriores.

Además, el método es útil en el diagnóstico primario de tumores. El método se puede utilizar para buscar mutaciones del ADNct en pacientes sanos para el diagnóstico temprano. Por ejemplo, se pueden controlar los oncogenes clave y los genes supresores de tumores.

El método también es útil con muestras obtenidas de sujetos sanos, para el análisis o prueba de rutina (es decir, el control) de variantes genéticas. En algunas realizaciones, el sujeto sano puede haberse recuperado de una enfermedad, por ejemplo, el sujeto puede estar en remisión de un cáncer. Como alternativa, el sujeto sano puede tener un mayor riesgo de contraer una enfermedad, por ejemplo, el sujeto sano puede tener antecedentes familiares de una enfermedad. El método de la invención se puede usar para analizar de forma rutinaria las variantes genéticas, opcionalmente para las variantes genéticas que se sabe que están asociadas con una enfermedad dada. De esta manera, las muestras obtenidas de sujetos sanos se pueden usar para detectar variantes genéticas asociadas a la enfermedad antes de que se manifiesten los síntomas clínicos, lo que facilita la intervención terapéutica temprana.

El método de la invención es útil para comunicar decisiones terapéuticas. Es decir, la identificación de variantes genéticas predictivas de susceptibilidad, resistencia o respuesta a la terapia puede usarse para seleccionar un curso de tratamiento apropiado para un sujeto.

A modo de ejemplo, se sabe que la mutación T790M en EGFR está asociada con la resistencia al tratamiento de

cánceres utilizando los inhibidores de EGFR gefitinib y erlotinib. Los sujetos en los que se identifica esta mutación utilizando el método de la invención, por lo tanto, se identificarían como candidatos inadecuados para el tratamiento utilizando estos inhibidores.

De forma similar, la mutación V600E en B-Raf se asocia con una mayor sensibilidad a los inhibidores de B-Raf. Los sujetos en los que se identifica esta mutación utilizando el método de la invención, por lo tanto, se identificarían como buenos candidatos para el tratamiento utilizando estos inhibidores.

El método se puede realizar en muestras de ADN obtenidas de un sujeto en diferentes momentos, por ejemplo, en diferentes etapas de una enfermedad o curso de tratamiento.

El método se puede utilizar para analizar muestras de ADN para detectar variaciones genéticas asociadas con o predictivas de susceptibilidad, progresión (es decir, pronóstico), resistencia o respuesta a una terapia determinada. Por ejemplo, el método se puede usar para identificar determinantes nuevos (es decir, no identificados previamente) y/o predictores de susceptibilidad o pronóstico de una enfermedad, o de resistencia o respuesta a una terapia.

Siendo cuantitativo, el método de la invención es útil para investigar la frecuencia relativa de variantes genéticas en diferentes etapas de la enfermedad, en respuesta al tratamiento y/o en sujetos sanos a lo largo del tiempo.

Esta información, a su vez, puede usarse para comunicar decisiones terapéuticas. Por ejemplo, un aumento o disminución en la frecuencia de una variante genética conocida y/o prevista para asociarla con, por ejemplo, la susceptibilidad o la resistencia a una terapia en particular guiarán la decisión sobre qué terapia es la más adecuada para el tratamiento de un sujeto dado.

Ejemplos

Ejemplo 1 - Determinación de ruido de fondo

Se diseñaron 41 amplicones, que cubren todas las regiones TP53 y de punto caliente en EGFR, *KRAS,* BRAF y PIK3CA, y se leyeron en las direcciones directa e inversa, dando 82 familias leídas en total (Tabla 1). Los amplicones cubren 5.038 bases incluyendo lecturas de superposición directa e inversa (excluyendo secuencias de cebadores).

Tabla 1:

Cromosoma	Coordenada izquierda	Coordenada derecha	Nombre del amplicón
chr7	140453098	140453187	BRAF_D0016_001_F
chr7	140453128	140453217	BRAF_D0016_001_R
chr7	55241589	55241678	EGFR_D0016_001_F
chr7	55241620	55241709	EGFR_D0016_001_R
chr7	55241658	55241745	EGFR_D0016_002_F
chr7	55241659	55241746	EGFR_D0016_002_R
chr7	55241705	55241792	EGFR_D0016_003_F
chr7	55241706	55241793	EGFR_D0016_003_R
chr7	55242385	55242474	EGFR D0016_004 F
chr7	55242397	55242486	EGFR D0016 004_R
chr7	55242427	55242516	EGFR_D0016_005_F
chr7	55242448	55242537	EGFR_D0016_005_R
chr7	55248931	55249020	EGFR_D0016_006_F
chr7	55248938	55249027	EGFR_D0016_006_R
chr7	55248989	55249078	EGFR_D0016_007_F
chr7	55249011	55249100	EGFR_D0016_007_R

chr7	55249062	55249151	EGFR_D0016_008_F
chr7	55249101	55249190	EGFR_D0016_008_R
chr7	55249144	55249233	EGFR D0016_009_F
chr7	55249160	55249249	EGFR_D0016_009_R
chr7	55259388	55259477	EGFR_D0016_010_F
chr7	55259409	55259498	EGFR_D0016_010_R
chr7	55259456	55259545	EGFR_D0016_011_F
chr7	55259485	55259574	EGFR_D0016_011_R
chr7	55259526	55259615	EGFR_D0016_012_F
chr7	55259546	55259635	EGFR_D0016_012_R
chr12	25378518	25378607	KRAS_D0016_001_F
chr12	25378524	25378613	KRAS_D0016_001_R
chr12	25380216	25380305	KRAS_D0016_002_F
chr12	25380248	25380337	KRAS_D0016_002_R
chr12	25398246	25398335	KRAS D0016_003_F
chr12	25398248	25398337	KRAS_D0016_003_R
chr3	178936028	178936117	PIK3CA_D0016_001_F
chr3	178936046	178936135	PIK3CA_D0016_001_R
chr3	178952022	178952111	PIK3CA_D0016_002_F
chr3	178952056	178952145	PIK3CA_D0016_002_R
chr17	7572903	7572992	TP53_D0016_001_F
chr17	7572942	7573031	TP53_D0016_001_R
chr17	7573904	7573993	TP53_D0016_002_F
chr17	7573930	7574019	TP53_D0016_002_R
chr17	7573975	7574064	TP53_D0016_003_F
chr17	7573988	7574077	TP53_D0016_003_R
chr17	7576789	7576878	TP53_D0016_004_F
chr17	7576828	7576917	TP53_D0016_004_R
chr17	7576873	7576960	TP53_D0016_005_F
chr17	7576874	7576961	TP53_D0016_005_R
chr17	7576996	7577085	TP53_D0016_006_F
chr17	7577026	7577115	TP53_D0016_006_R
chr17	7577074	7577163	TP53_D0016_007_F
chr17	7577093	7577182	TP53_D0016_007_R

chr17	7577434	7577523	TP53_D0016_008_F
chr17	7577439	7577528	TP53_D0016_008_R
chr17	7577484	7577573	TP53_D0016_009_F
chr17	7577523	7577612	TP53_D0016_009_R
chr17	7577561	7577650	TP53_D0016_010_F
chr17	7577578	7577667	TP53_D0016_010_R
chr17	7578120	7578209	TP53_D0016_011_F
chr17	7578124	7578213	TP53_D0016_011_R
chr17	7578173	7578262	TP53_D0016_012_F
chr17	7578189	7578278	TP53_D0016_012_R
chr17	7578228	7578316	TP53_D0016_013_F
chr17	7578229	7578317	TP53_D0016_013_R
chr17	7578343	7578432	TP53_D0016_014_F
chr17	7578361	7578450	TP53_D0016_014_R
chr17	7578407	7578496	TP53_D0016_015_F
chr17	7578434	7578523	TP53_D0016_015_R
chr17	7578483	7578572	TP53_D0016_016_F
chr17	7578492	7578581	TP53_D0016_016_R
chr 17	7579284	7579373	TP53_D0016_017_F
chr17	7579319	7579408	TP53_D0016_017_R
chr17	7579364	7579452	TP53_D0016_018_F
chr17	7579365	7579453	TP53_D0016_018_R
chr17	7579405	7579494	TP53_D0016_019_F
chr17	7579444	7579533	TP53_D0016_019_R
chr17	7579488	7579577	TP53_D0016_020_F
chr17	7579527	7579616	TP53_D0016_020_R
chr17	7579677	7579765	TP53_D0016_021_F
chr17	7579678	7579766	TP53_D0016_021_R
chr17	7579816	7579905	TP53_D0016_022_F
chr17	7579839	7579928	TP53_D0016_022_R
chr17	7579887	7579976	TP53_D0016_023_F
chr17	7579892	7579981	TP53_D0016_023_R

Cada locus puede cambiar a uno de tres posibles alelos que no sean de referencia. Excluyendo los polimorfismos conocidos, los polimorfismos identificados en las líneas celulares utilizadas y los loci que no se pueden modelar, se consideraron un total de 13.278 posibles cambios de una sola base.

Los datos de secuenciación para este panel de amplicones se recopilaron a partir de 336 dobles amplificaciones de muestras de ADN de la línea celular LNCaP (un adenocarcinoma de próstata humano, línea celular sensible al andrógeno; ATCC CRL-1740).

Para cada cambio de base en cada familia de lectura (amplicón y dirección de lectura), distintos de los loci excluidos anteriormente, se calcularon la frecuencia alélica media (AF) y el coeficiente de variación (CV) de la AF. Para un alelo dado (es decir, una variante genética), la AF se calcula como la proporción de lecturas que contienen este alelo de todas las lecturas de la familia leídas.

La Figura 1A muestra el número de ocurrencias de diferentes valores de la media de AF, agrupados al aumentar los valores de la AF. La línea vertical indica la media = 0,00022, que se encuentra en el percentil del 95% de la distribución.

La Figura 1A muestra el número de ocurrencias de diferentes valores del CV, agrupados al aumentar los valores del CV. La línea vertical indica el CV= 18,16, que se encuentra en el percentil del 95% de la distribución.

Ejemplo 2: determinación de la AF requerida para una identificación de positivos para un cambio de base en una reacción dada

Se definió una reacción como positiva para un cambio de secuencia dado si se encontró que la AF observada era mayor que la frecuencia mínima detectable (MDF) para ese alelo. La MDF para un alelo se calculó como la AF en el que la función de distribución acumulada (CDF) para ese cambio de base particular cruza un umbral predefinido (CDF_thresh).

La Figura 2 muestra las distribuciones acumuladas de varios cambios básicos del panel del Ejemplo 1. Se muestran las funciones de distribución acumulada para las distribuciones Normal y Beta, ajustadas utilizando los valores para la media de la AF y el CV que se calcularon para esos cambios de base, en comparación con las distribuciones empíricas de los datos. El valor de Kolmogorov-Smirnov para la bondad de ajuste para las distribuciones Normal (KS Normal) y Beta (KS Beta) a las distribuciones empíricas de los datos es el siguiente:

Cambio de Base	**KS Normal**	**KS Beta**
EGFR D0016_006_R 55248959 C>G	8,55	0,78
EGFR_D0016_009_F 55249159 G>T	9,26	2,89
EGFR D0016_003_R 55241757 C>A	9,45	0,96
EGFR_D0016_012_R 55259573 G>C	8,83	0,45

El proceso descrito se utilizó para escanear 13.278 posibles cambios de una sola base en múltiples muestras. Para mantener una tasa baja de falsos positivos, se llamaron mutaciones si estaban presentes en una AF tal que AF> MDF con una CDF_thresh = 0,9999.

El ruido de fondo se modeló como una distribución Beta, y para cada uno de los 13.278 cambios de base posibles, dada la media de la AF y CV medidos para el cambio de base en las muestras de control (Figura 1), se calculó la MDF (CDF_thresh), que es la AF en la que

$$F(AF|Media,CV) = CDF\_thresh$$

donde "F" es la CDF de una distribución Beta.

La Figura 3 muestra la MDF (CDF_thresh) para diferentes valores de CDF_thresh (como se indica en la leyenda). Los datos en cada caso se ordenan por los valores de la MDF obtenidos.

La Figura 3B muestra un acercamiento. La línea vertical indica el punto de percentil 97,5, para el cual el 97,5% de los cambios de base posibles tienen un valor de AF más bajo. Para CDF_thresh = 0,9999, el percentil 97,5 está en AF = 0.0382, indicado por una línea horizontal.

Por lo tanto, se determinó que para los 13.278 posibles cambios de una sola base, una MDF a CDF_thresh = 0,9999 sería 0,0382 o menor para el 97,5% de los posibles cambios de una sola base (véase la Figura 3).

La probabilidad de una determinación de positivos falsos para un cambio de base dado en cada reacción (P_fp1), de acuerdo con este modelo, es un CDF thresh menos, que es el valor de la CDF en AF = MDF:

$$P\_fp1T = 1-F(AF = MDF[Media,CV) = 1-CDF\_thresh$$

Ejemplo 3: Determinación del número de reacciones de doble amplificación para que sean positivas para un cambio de base dado para determinar la presencia del cambio de base en una muestra

Para minimizar las determinaciones de positivos falsos para la presencia de un cambio de base en una muestra, se requirieron múltiples reacciones de doble amplificación positiva (N) del número total de reacciones de doble amplificación para esa muestra (T).

La probabilidad teórica de que aparezca un positivo falso en N de T reacciones (P_fpNT) es

$$P\_fpNT = (P\_fp1T)^N) * C\_NT = ((1 - CDF\_thresh)^N) * C\_NT$$

donde "C_NT" es el número de posibilidades de elegir N pozos de T sin prestar atención al orden, y es

$$C\_NT = T!/((T-N)!*N!)$$

donde “!” indica la función factorial.

Por ejemplo, para elegir 3 reacciones de 48 hay 17.296 posibilidades. Para elegir 2 reacciones de 48 hay 1.128 posibilidades. La tasa esperada de positivos falsos se calculó para cada muestra y se muestra en las Tablas 2 y 3 (Figura 4).

El requisito de observar múltiples reacciones positivas también aumenta la probabilidad de negativos falsos y disminuye la probabilidad de que se observe un positivo verdadero en múltiples reacciones.

La Figura 5 muestra la probabilidad de Poisson de encontrar al menos N número de moléculas en la muestra en función del número esperado de moléculas (según la concentración y la cantidad inicial de material). Las diferentes líneas muestran una N diferente, de modo que una identificación compuesta se realiza con ≥N reacciones positivas.

El sistema se diseñó de modo que

(i) el número total de reacciones de doble amplificación (T) para cada muestra fuera mucho mayor que N (aquí, T ≥48), y
(ii) la probabilidad de una identificación en cualquier reacción sería cerca de 1 si estaba presente una molécula mutante.

Ejemplo 4 - Determinación del número de moléculas amplificables que estarán presentes en cada reacción de amplificación

Para una tasa de dilución prevista y el número promedio correspondiente de moléculas por reacción, se espera que las reacciones reales tengan un número aleatorio de moléculas que puedan modelarse con una distribución de Poisson.

Una sola molécula mutante dentro de cada grupo podría conducir a una AF observada igual al recíproco del número de moléculas (asumiendo que las fracciones leídas son representativas de la frecuencia del alelo mutante, véase Forshew et al. 2012 Sci Transl Med 4 (136) 136ra68, Figura 3B).

Para diferentes valores de MDF (como se indica en la leyenda de la figura), la probabilidad de AF> MDF se representó en función del número promedio de moléculas por reacción, en función de la dilución del molde (Figura 8).

La línea horizontal muestra probabilidad = 0,90. Para MDF = 0,0382 (para el cual el 97,5% de los cambios en la base pasarían un CDF_thresh = 0,9999, véase la Figura 3), se obtiene una probabilidad de 0,90 de que AF> MDF con un N extrapolado = 20,59 (indicado por la línea vertical).

Se seleccionó una tasa de dilución para la muestra de ADN, de modo que cada reacción de amplificación tendría unas 20 moléculas de molde amplificables esperadas, de manera que para cada reacción que tenga presente una molécula mutante amplificable, la probabilidad de tener una AF mutante> 0,0382 sería > 0,9.

La probabilidad de Poisson de observar ≥N moléculas en las muestras fue aproximadamente igual a la probabilidad de una identificación compuesta basándose en la identificación de la secuencia mutante por encima de las tasas de fondo en ≥N reacciones. La probabilidad aproximada de un negativo falso se calculó para cada muestra y se muestra en las Tablas 2 y 3 (Figura 4).

Ejemplo 5 - Cebadores de PCR y diseño para pruebas de experimentos de principio

Se diseñaron 41 pares de cebadores para el panel de 41 amplicones. Cada cebador se produjo con una secuencia de etiquetado en el extremo 5' y una de las 7 secuencias de códigos de barras diferentes en el medio (Figura 6). Los

cebadores específicos de diana se dividieron en dos grupos de PCR múltiple optimizados, y los cebadores directos e inversos se agruparon de tal manera que produjeron 49 combinaciones diferentes de códigos de barras (Figura 7). Se dispensaron 48 de los grupos de cebadores en diferentes pocillos de PCR utilizando la plataforma Fluidigm, para hacerlo en alto rendimiento.

El ADN se cuantificó digitalmente y agregó para que un promedio de 20 moléculas entrara en cada PCR. La matriz de Acceso Fluidigm tiene un volumen muerto del 68,32% que se tuvo en cuenta al agregar el ADN (volumen muerto = 33 nl*48/5 µl).

Esto permitió la amplificación de 48 grupos separados de ~ 20 moléculas por muestra. Para muestras en las que se esperaban alelos de baja frecuencia, se procesaron múltiples conjuntos de 48 pocillos.

Tras la recolección de esta primera PCR, se realizó una segunda ronda de PCR para unir adaptadores del secuenciador y códigos de barras de muestra. A continuación, se limpiaron y secuenciaron las bibliotecas en los secuenciadores Illumina MiSeq y HiSeq 2000.

Ejemplo 6 - Prueba de principio en unas series de dilución de líneas celulares

Se creó una línea celular de ADN de dilución en serie donde se esperaba que estuvieran presentes mutaciones heterocigotas en una AF entre 3-0,04%. Las líneas celulares NCI-H1975 y VCAP tienen 5 mutaciones conocidas/SNP detectables usando el panel de cebador/amplicón descrito en el Ejemplo 5 y no están presentes en la línea celular LNCaP (Tabla 4). El ADN de VCAP, NCI-H1975 y LNCaP se cuantificó y normalizó digitalmente, y a continuación se diluyó en serie en el ADN de LNCaP.

Tabla 5:

Línea celular	**Gen**	**cambio de ADNc**	**Cambio proteico**	**Cambio genómico (hg19)**
NCI-H1975	EGFR	c.2369C>T	P.T790M	chr7:55249071 C>T
NCI-H1975	EGFR	c.2573T>G	p.L858R	chr7:55259515 T>G
NCI-H1975	TP53	c.818G>A	p.R273H	chr17:7577120 C>T
NCI-H1975	TP53	SNP	rs17880604	Chr17:7577644 C>G
VCAP	TP53	p.R248W	c.742C>T	chr17:7577539 G>A

El método se realizó en unas series de dilución de dos líneas celulares, en un total de 5 muestras diferentes, utilizando diferentes números de reacciones para diferentes muestras (véase Tablas 2 y 3 (Figura 4)).

Las dos líneas celulares diferían de LNCaP en 5 polimorfismos de base única en las regiones del genoma cubiertas, para un total de 25 positivos (50 cuando se identifica de forma directa e inversa independientemente).

Para modelar las tasas de fondo, se utilizaron la media de la AF y el CV en cada posición en la región de interés para estimar los parámetros y ajustar una distribución Beta específica para cada locus y cada posible cambio de base.

Basándose en el modelo anterior y los cálculos en las Tablas 2 y 3 (Figura 4), utilizando CDF_thresh = 0,9999 y una identificación compuesta que requiere múltiples reacciones positivas (N), se obtuvieron los siguientes resultados.

N ≥2

Sensibilidad global de 0,96, 36 identificaciones compuestas de positivos falsos. La Tabla 2 (Figura 4A) muestra la sensibilidad a diferentes tasas de dilución.

N ≥3

Sensibilidad global de 0,90, 1 identificación compuesta de positivos falsos. La Tabla 3 (Figura 4B) muestra la sensibilidad a diferentes tasas de dilución.

Se muestran las mutaciones identificadas (positivos verdaderos (TP)), identificaciones de positivos falsos (FP) y mutaciones perdidas (negativos falsos (FN)) al realizar el método en una serie de dilución de ADN de dos líneas celulares en la Tabla 4 y en la Figura 9. Los puntos de datos en la Figura 9 se ordenan al aumentar la AF medida, mostrando para cada mutación la AF medida, a excepción de las mutaciones FN, que muestran la AF esperada.

La Figura 10 muestra la concordancia entre la AF esperada y la medida en función de las identificaciones compuestas

de mutación utilizando la CDF_thresh = 0,9999 y N≥3. Los negativos falsos se muestran como "No detectado" (ND) en el eje vertical. Los valores esperados se muestran como cruces, conectados por una línea de puntos como una guía para el ojo.

Ejemplo 7 - Prueba de principio con un método alternativo de identificación de mutación

El método se realizó en unas series de dilución como se describe en el Ejemplo 6.

El ADN de VCAP y NCI-H1975 se diluyó en serie en el ADN de LNCaP como se muestra en la Tabla 6.

Tabla 6:

Nombre	% mutante	Moles mutantes por reacción	Repeticiones	Total de moles mut
DIL 3 %	3	28,8	1x48=48	28,8
DIL 1 %	1	9,6	2x48=96	19,2
DIL 0,33 %	0,33	3,20	2x48=96	6,4
DIL 0,11 %	0,11	1,07	6x48=288	6,4
DIL 0,037 %	0,04	0,36	12x48=576	4,266667

En el primer experimento, se secuenciaron 1 o más de las diluciones, 7 (x48) de LNCaP, 1 (x48) de VCAP, 1 (x48) de NCI-H1975 (Tabla 6, "Repeticiones").

Identificación de mutaciones

Se usaron 336 reacciones de PCR de LNCaP para determinar la AF de fondo en cada base. Se usó una distribución normal para modelar el fondo para todas las bases posibles desde la base que no era de referencia en todas las posiciones de la región de interés.

A continuación, cada reacción se examinó para detectar cambios en cada base que difieran del fondo por una puntuación z específica y profundidad. Se determinó que los cambios en la base identificados un número específico de veces por encima de esta puntuación tienen la mutación.

Resultados

La identificación de la mutación se realizó por primera vez con un corte de puntuación z de 20 y 3 pocillos positivos. La tabla 7 muestra los resultados de detección de mutaciones.

Todas las mutaciones/SNP se detectaron hasta el 0,33%. De las 1.920 moléculas analizadas en la dilución del 0,33% (2 muestras x 48 pocillos x 20 moléculas de ADN), se detectaron entre 4 y 12 pocillos/moléculas positivos, que corresponden a casi exactamente el 0,33% (Tabla 7, Dil 0,33).

En la dilución al 0,11% se detectaron 4 de los 5 cambios. Se perdió 1 ya que solo 2 pocillos fueron positivos (Tabla 7, Dil 0.11).

Finalmente, en la dilución del 0,037% se omitió por completo 1 de 5 mutaciones; se omitió otra (chr17: 7577644 C> G) en uno de los dos amplicones superpuestos. Estos resultados concuerdan con la distribución aleatoria de estas moléculas mutantes (Tabla 7, Dil 0,037).

Las 2 reacciones de positivos falsos fueron más probables debido al error de la polimerasa durante las primeras rondas de amplificación de la biblioteca. Tales errores deben ser normalmente en frecuencias más bajas que los cambios normales.

Al aumentar nuestro corte de puntuación z a 30, se mantuvieron todos los cambios reales y se eliminaron los 2 positivos falsos (Tabla 7).

Tabla 7: número de reacciones positivas determinadas por la puntuación z superior a 30. Para los verdaderos positivos, se hace media de este número sobre las dos amplificaciones superpuestas (directa e inversa). Se muestran identificaciones de positivos falsos si al menos fue positivo un mínimo de 3 reacciones.

Cambios esperados	**NCI-H1975**	**VCAP**	**DIL 3**	**DIL 1**	**DIL 0,33**	**DIL 0,11**	**DIL 0,037**
chr17:7577120 C>T	48	0	26	27,5	4	6,5	3
chr17:7577539 G>A	0	48	19,5	18	5,5	2	0
chr17:7577644 C>G	46,5	0	29	22	11,5	3	2,5
chr7:55249071 C>T	47,5	0	21,5	15	6,5	8	4
chr7:55259515 T>G	41	0	17	9,5	8	7	3
Positivos falsos							
chr17:7576861 T>C	0	0	0	0	0	2*	0
*identificado a puntuación z = 20							

REIVINDICACIONES

1. Un método para detectar una variante genética en una región de interés en una muestra de ADN, que comprende:

(i) establecer, para una plataforma de secuenciación dada, proceso de secuenciación y profundidad de secuenciación, una frecuencia umbral igual o superior a la cual se debe observar la variante genética en los resultados de la secuenciación de las reacciones de amplificación para asignar una determinación de positivos de la presencia de la variante genética en una reacción de amplificación dada, en donde la frecuencia umbral se basa en la distribución del número de lecturas que soportan la variante genética que se espera observar en los resultados de la secuenciación de las reacciones de amplificación debido al error de amplificación y secuenciación (distribución del recuento de lectura);
(ii) dividir la muestra de ADN en una pluralidad de reacciones de doble amplificación, de modo que, para la variante genética, el número promedio de moléculas de molde amplificables de la región de interés en una reacción de doble amplificación es menor que el recíproco de la frecuencia umbral determinada en el paso (i);
(iii) realizar las reacciones de amplificación del paso (ii) y secuenciar los productos de las reacciones de amplificación;
(iv) basándose en el paso (i) y en los resultados del paso (iii), determinar la presencia/ausencia de la variante genética en cada reacción de doble amplificación; y
(v) integrar los resultados de (iv) para determinar la presencia/ausencia de la variante genética en la región de interés en la muestra de ADN.

2. El método de acuerdo con la reivindicación 1, en donde para determinar que la variante genética está presente en la región de interés en la muestra de ADN en el paso (v), se debe realizar una determinación de positivos de la presencia de la variante genética en más de una reacción de doble amplificación en el paso (iv), opcionalmente, en donde se debe realizar una determinación de positivos de la presencia de la variante genética en al menos 3 reacciones de doble amplificación en el paso (iv).

3. El método de acuerdo con una cualquiera de las reivindicaciones anteriores, en donde la frecuencia umbral del paso (i) es la frecuencia a la cual el valor de la función de distribución acumulada para la probabilidad de observar la variante genética a esa frecuencia es 0,99 o superior, 0,995 o superior, 0,999 o superior, 0,9999 o superior o 0,99999 o superior.

4. El método de acuerdo con una cualquiera de las reivindicaciones anteriores, en donde la frecuencia umbral del paso (i) se determina utilizando un modelo de distribución de probabilidad binomial, binomial sobredispersa, Beta, Normal, Exponencial o Gamma.

5. El método de acuerdo con una cualquiera de las reivindicaciones anteriores, en donde el paso (i) comprende:

(a) basándose en la distribución del recuento de lectura para una pluralidad de variantes genéticas, establecer una pluralidad de frecuencias umbral en o por encima de las cuales deben observarse las variantes genéticas en los resultados de la secuenciación de las reacciones de amplificación para asignar una determinación de positivos para la presencia de la variante genética en una reacción de amplificación dada; y
(b) basándose en el paso (a), establecer una frecuencia umbral general en o por encima de la cual se debe observar una variante genética en los resultados de la secuenciación de una reacción de amplificación dada para asignar una determinación de positivos de la presencia de la variante genética en esa reacción de amplificación, que es la frecuencia umbral a la cual el 90 %, 95 %, 97,5 %, 99 % o más de la pluralidad de frecuencias umbral determinadas en el paso (a) son menores de este valor.

6. El método de acuerdo con una cualquiera de las reivindicaciones anteriores, en donde el número medio de moléculas de molde amplificables que están presentes en una reacción de doble amplificación es el número en el que cuando la variante genética está presente en una molécula de molde amplificable única de una reacción de doble amplificación, la probabilidad de que se realice una determinación de positivos para esa réplica es 0,9 o mayor.

7. El método de acuerdo con una cualquiera de las reivindicaciones anteriores, en donde el método comprende un paso para determinar, para una plataforma de secuenciación dada, proceso de secuenciación y profundidad de secuenciación, la distribución del número de lecturas que soportan la variante genética o una pluralidad de variantes genéticas que se espera observar en los resultados de secuenciación de las reacciones de amplificación debido al error de amplificación y secuenciación (distribución del recuento de lectura),
opcionalmente,en donde el método para determinar la distribución del recuento de lectura comprende además secuenciar una muestra de ADN varias veces para determinar la distribución del recuento de lectura para la variante genética o para una pluralidad de variantes genéticas.

8. El método según una cualquiera de las reivindicaciones 1 a 6, en donde la distribución del recuento de lectura se basa en las tasas de error del secuenciador y/o la polimerasa, teniendo en cuenta, opcionalmente, el contexto de secuencia.

9. El método según una cualquiera de las reivindicaciones 1 a 5, en donde el método comprende "buscar" un valor de

referencia o una pluralidad de valores de referencia para la distribución del recuento de lectura para la variante genética o para una pluralidad de variantes genéticas en una base de datos, un gráfico, una tabla, una lista, un catálogo, un índice, un directorio o un registro, opcionalmente, en donde el valor de referencia o la pluralidad de valores de referencia se determinaron secuenciando una muestra de ADN varias veces, para determinar la distribución del recuento de lectura para la variante genética o para una pluralidad de variantes genéticas.

10. El método de acuerdo con una cualquiera de las reivindicaciones anteriores, en donde después de dividir la muestra de ADN en el paso (ii), cada reacción de doble amplificación tiene más de una molécula de molde amplificable única para la región de interés por reacción de doble amplificación, opcionalmente, en donde el número medio de moléculas de molde amplificables por reacción de doble amplificación es más de 1 y menos de 1000, más de 2 y menos de 1000 o más de 5 y menos de 1000.

11. El método de acuerdo con cualquiera de las reivindicaciones anteriores, que es capaz de detectar una variante genética que está presente dentro de la población de moléculas de molde amplificables de una muestra de ADN a una frecuencia inferior al 2 %, inferior al 1 %, inferior al 0,5 %, inferior al 0,2 %, inferior al 0,1 % o inferior al 0,05 %.

12. El método de acuerdo con una cualquiera de las reivindicaciones anteriores, en donde la división de la muestra de ADN en una pluralidad de reacciones de doble amplificación comprende diluir la muestra de ADN y dividir en partes alícuotas en reacciones de doble amplificación.

13. El método de acuerdo con una cualquiera de las reivindicaciones anteriores, en donde las reacciones de doble amplificación se realizan en paralelo y la secuenciación de los productos de las reacciones de amplificación se realiza en paralelo,
opcionalmente, en donde las reacciones de amplificación se realizan utilizando uno o más cebadores que flanquean la región de interés y que integran la muestra y/o las secuencias identificadoras específicas de reacción de doble amplificación en los productos de amplificación,
y opcionalmente además, en donde los cebadores integran adaptadores de secuencia en los productos de amplificación.

14. El método de acuerdo con una cualquiera de las reivindicaciones anteriores, en donde se analizan en paralelo una pluralidad de regiones de interés.

15. El método de acuerdo con una cualquiera de las reivindicaciones anteriores, en donde el método comprende además:

(vi) determinar la frecuencia de la variante genética en la muestra de ADN.

16. Uso del método según una cualquiera de las reivindicaciones 1 a 15:

para detectar y/o cuantificar el ADN tumoral circulante en una muestra obtenida de un sujeto;
en un método de diagnóstico o de pronóstico *in vitro* o en el seguimiento de una enfermedad;
en un método de selección de un paciente para una terapia;
en un método de identificación de variación genética predictiva de la susceptibilidad, la resistencia o la respuesta a una terapia; o
para hacer el seguimiento o evaluar la respuesta a una terapia.

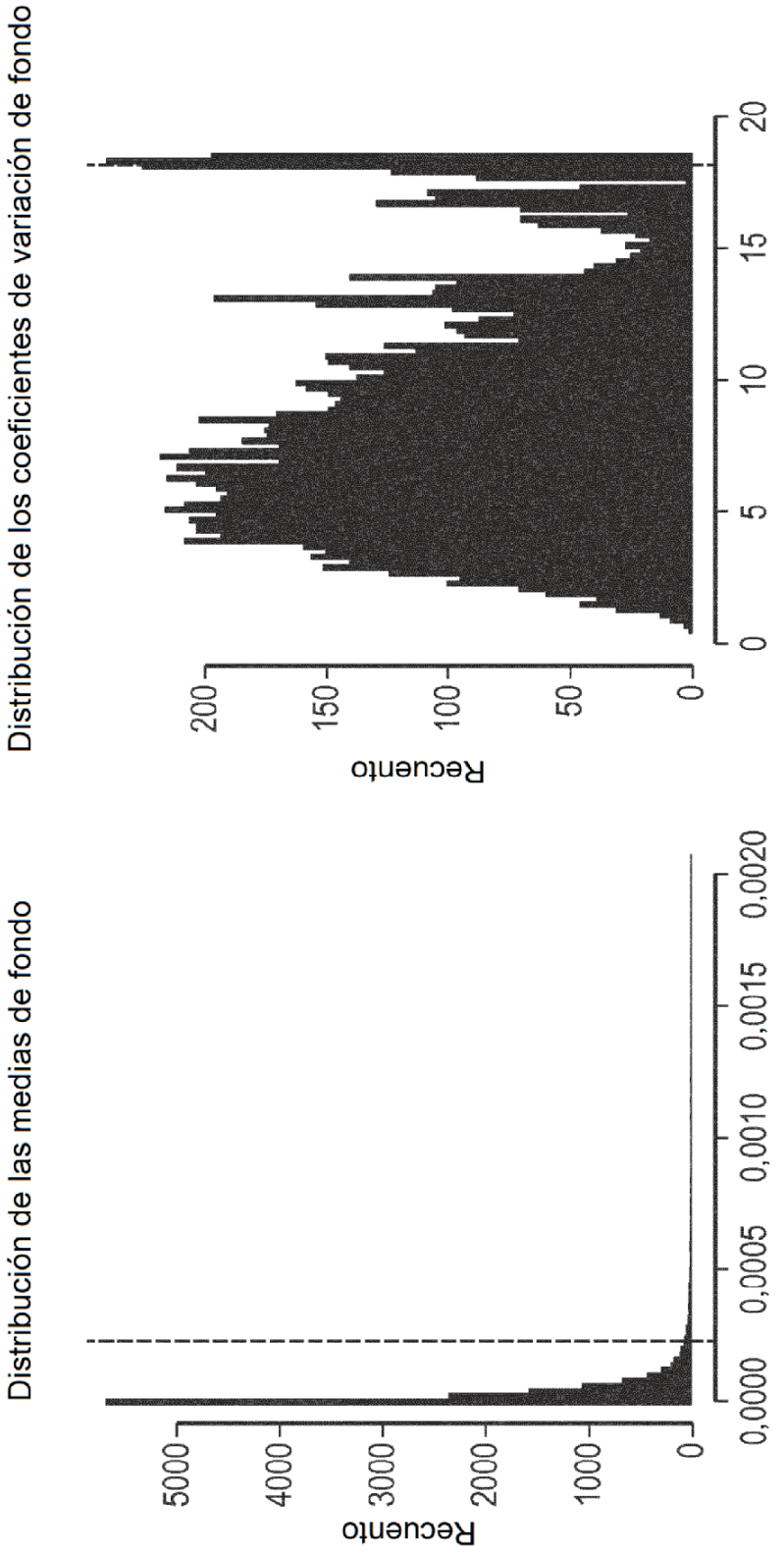
Distribución de las medias de fondo
Recuento
5000
4000
3000
2000
1000
0
0,0000
0,0005
0,0010
0,0015
0,0020
Media de AF

FIG. 1A

Distribución de los coeficientes de variación de fondo
Recuento
200
150
100
50
0
0
5
10
15
20
CV

FIG. 1B

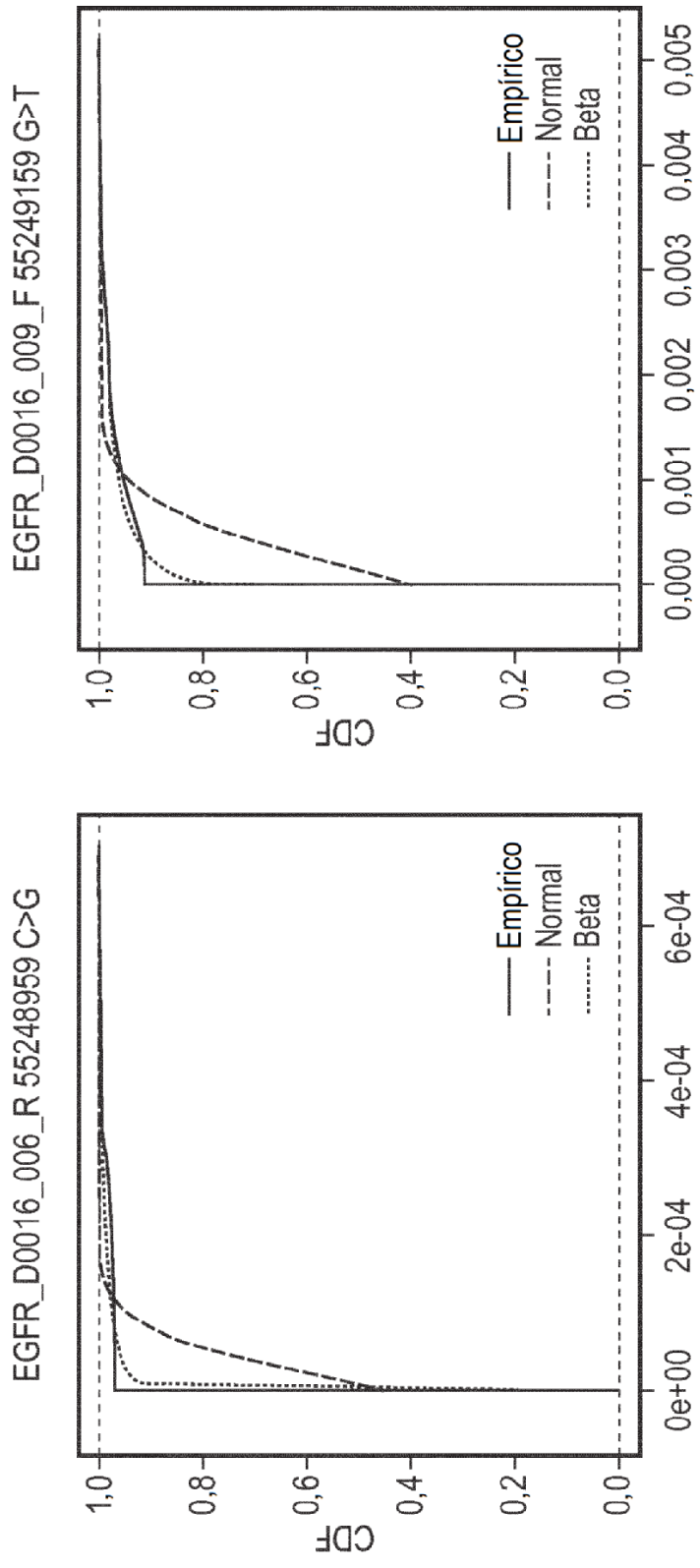
EGFR_D0016_009_F 55249159 G>T
CDF
1,0
0,8
0,6
0,4
0,2
0,0
0,000
0,001
0,002
0,003
0,004
0,005
AF
Empírico
Normal
Beta

FIG. 2B

EGFR_D0016_006_R 55248959 C>G
CDF
1,0
0,8
0,6
0,4
0,2
0,0
0e+00
2e-04
4e-04
6e-04
AF
Empírico
Normal
Beta

FIG. 2A

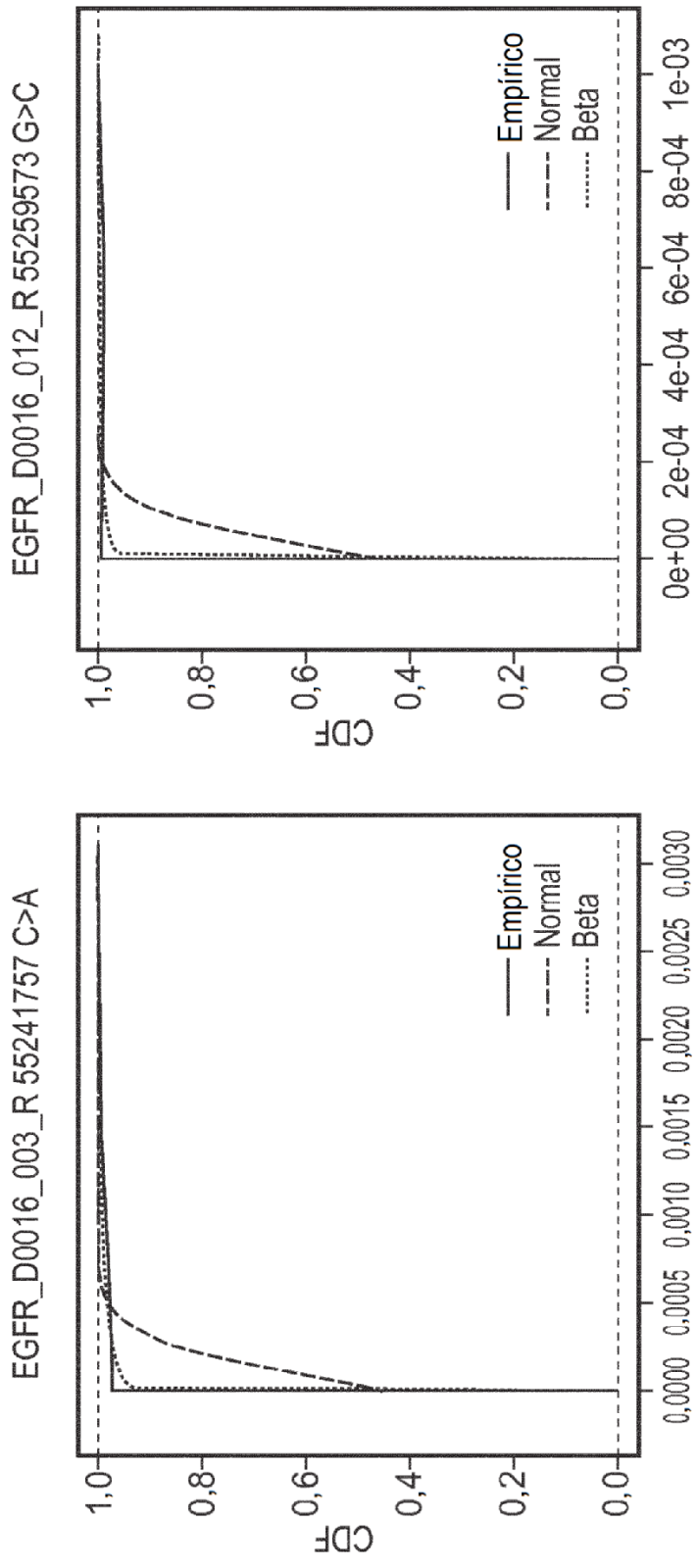
EGFR_D0016_003_R 55241757 C>A
CDF
1,0
0,8
0,6
0,4
0,2
0,0
0,0000
0,0005
0,0010
0,0015
0,0020
0,0025
0,0030
AF
Empírico
Normal
Beta

FIG. 2C

EGFR_D0016_012_R 55259573 G>C
CDF
1,0
0,8
0,6
0,4
0,2
0,0
0e+00
2e-04
4e-04
6e-04
8e-04
1e-03
AF
Empírico
Normal
Beta

FIG. 2D

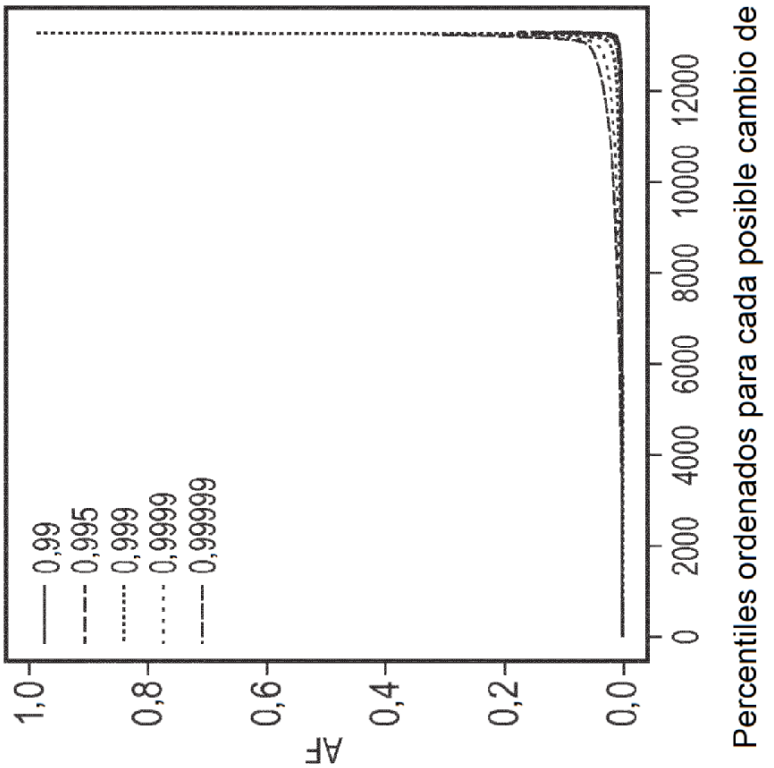
1,0
0,8
0,6
0,4
0,2
0,0
ΔF
0
2000
4000
6000
8000
10000
12000
0,99
0,995
0,999
0,9999
0,99999
Percentiles ordenados para cada posible cambio de base

FIG. 3A

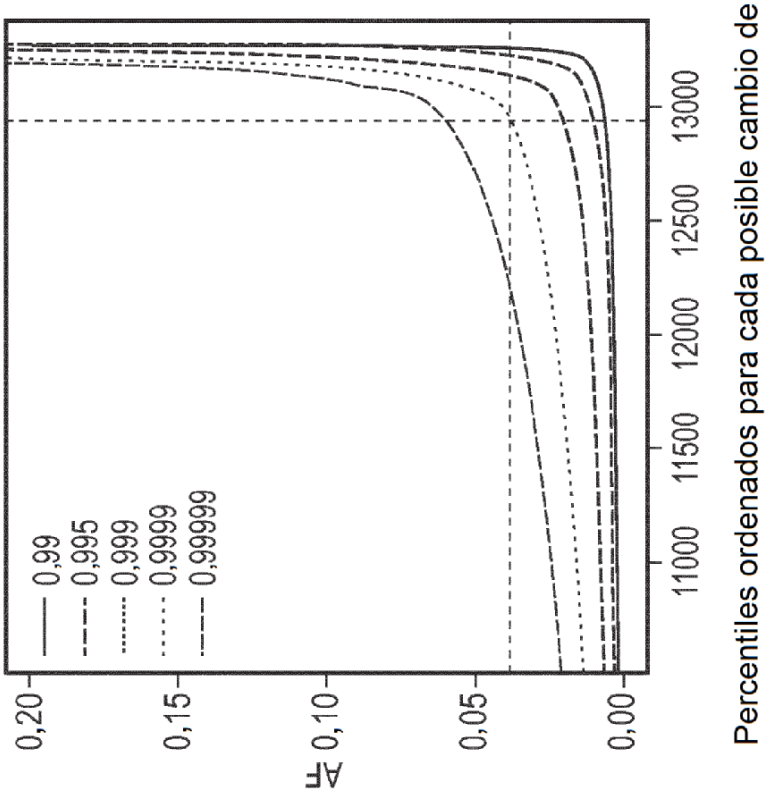
0,20
0,15
0,10
0,05
0,00
ΔF
11000
11500
12000
12500
13000
0,99
0,995
0,999
0,9999
0,99999
Percentiles ordenados para cada posible cambio de base

FIG. 3B

CDF_thresh=0,9999, identificación compuesta usando N≥2. Tabla 2

Número de muestra	Fracción de ADN mutante	Fracción de ADN de tipo silvestre	Número esperado de moléculas mutantes	Reacciones por muestra (T)	Elección entre T-2	Tasa esperada de FP	Probabilidad aproximada de un falso negativo	TP esperados	FN esperados	FP observados	TP observados	FN observados
1	0,03	0,97	28,8	48	1128	0,14977584	1,26E-11	10	1,26E-10	2	10	0
2	0,01	0,99	19,2	96	4560	0,6054768	1,13E-07	9,999999	1,13E-06	0	10	0
3	0,0033	0,9967	6,4	96	4560	0,6054768	0,01302792	9,869721	0,130279	3	10	0
4	0,0011	0,9989	6,4	288	41328	5,48753184	0,01302792	9,869721	0,130279	4	10	0
5	0,0004	0,9996	4,27	576	165600	21,988368	0,07621843	9,237816	0,762184	28	8	2
					Total	28,8366293		48,97726	1,022744	36	48	2
					Promedio		0,020454877					

*Probabilidad aproximada de un falso negativo basado en la probabilidad de Poisson de <N copias, más la probabilidad de que caigan en menos de N reacciones diferentes

FIG. 4A

CDF_thresh=0,9999, identificación compuesta usando N≥3. Tabla 3:

Número de muestra	Fracción de ADN mutante	Fracción de ADN de tipo silvestre	Número esperado de moléculas mutantes	Reacciones por muestra (T)	Elección entre T-3	Tasa esperada de FP	Probabilidad aproximada de un falso negativo	TP esperados	FP observados	FP esperados	TP observados	FN observados
1	0,03	0,97	28,8	48	17296	0,00022966	1,68E-10	10	1,68E-09	0	10	0
2	0,01	0,99	19,2	96	142880	0,00189716	1,06E-06	9,999989	1,06E-05	0	10	0
3	0,0033	0,9967	6,4	96	142880	0,00189716	0,04781593	9,521841	0,478159	0	10	0
4	0,0011	0,9989	6,4	288	3939936	0,05231447	0,04781593	9,521841	0,478159	0	8	2
5	0,0004	0,9996	4,27	576	31684800	0,42071077	0,204236	7,95764	2,04236	1	7	3
					Total	0,47704922		47,00131	2,998689	1	45	5
					Promedio		0,059973785					

*Probabilidad aproximada de un falso negativo basado en la probabilidad de Poisson de <N copias, más la probabilidad de que caigan en menos de N reacciones diferentes

FIG. 4B

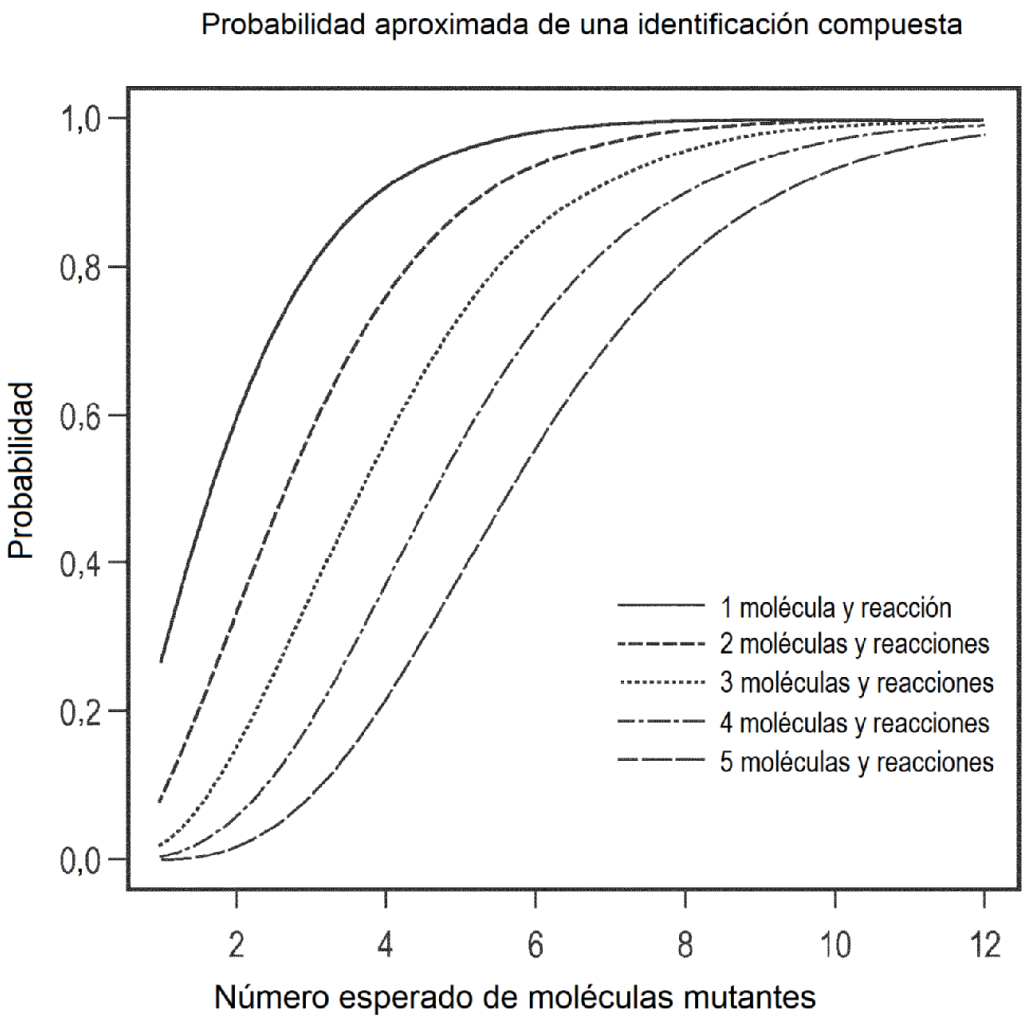
Probabilidad aproximada de una identificación compuesta
Probabilidad
1,0
0,8
0,6
0,4
0,2
0,0
2
4
6
8
10
12
Número esperado de moléculas mutantes
1 molécula y reacción
2 moléculas y reacciones
3 moléculas y reacciones
4 moléculas y reacciones
5 moléculas y reacciones

FIG. 5

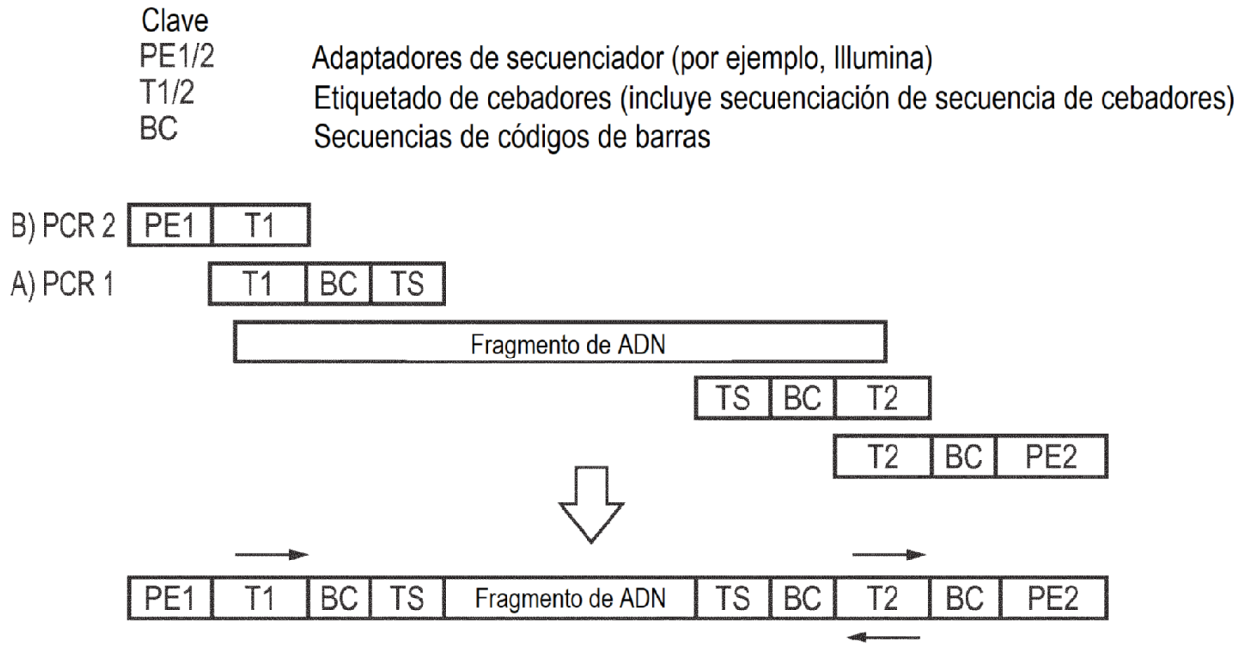
Clave
PE1/2 Adaptadores de secuenciador (por ejemplo, Illumina)
T1/2 Etiquetado de cebadores (incluye secuenciación de secuencia de cebadores)
BC Secuencias de códigos de barras
B) PCR 2
PE1
T1
A) PCR 1
T1
BC
TS
Fragmento de ADN
TS
BC
T2
T2
BC
PE2
PE1
T1
BC
TS
Fragmento de ADN
TS
BC
T2
BC
PE2

FIG. 6

	BC1	BC2	BC3	BC4	BC5	BC6	BC7
BC1	BC1/BC1	BC1/BC2	BC1/BC3	BC1/BC4	BC1/BC5	BC1/BC6	BC1/BC7
BC2	BC2/BC1	BC2/BC2	BC2/BC3	BC2/BC4	BC2/BC5	BC2/BC6	BC2/BC7
BC3	BC3/BC1	BC3/BC2	BC3/BC3	BC3/BC4	BC3/BC5	BC3/BC6	BC3/BC7
BC4	BC4/BC1	BC4/BC2	BC4/BC3	BC4/BC4	BC4/BC5	BC4/BC6	BC4/BC7
BC5	BC5/BC1	BC5/BC2	BC5/BC3	BC5/BC4	BC5/BC5	BC5/BC6	BC5/BC7
BC6	BC6/BC1	BC6/BC2	BC6/BC3	BC6/BC4	BC6/BC5	BC6/BC6	BC6/BC7
BC7	BC7/BC1	BC7/BC2	BC7/BC3	BC7/BC4	BC7/BC5	BC7/BC6	BC7/BC7

FIG. 7

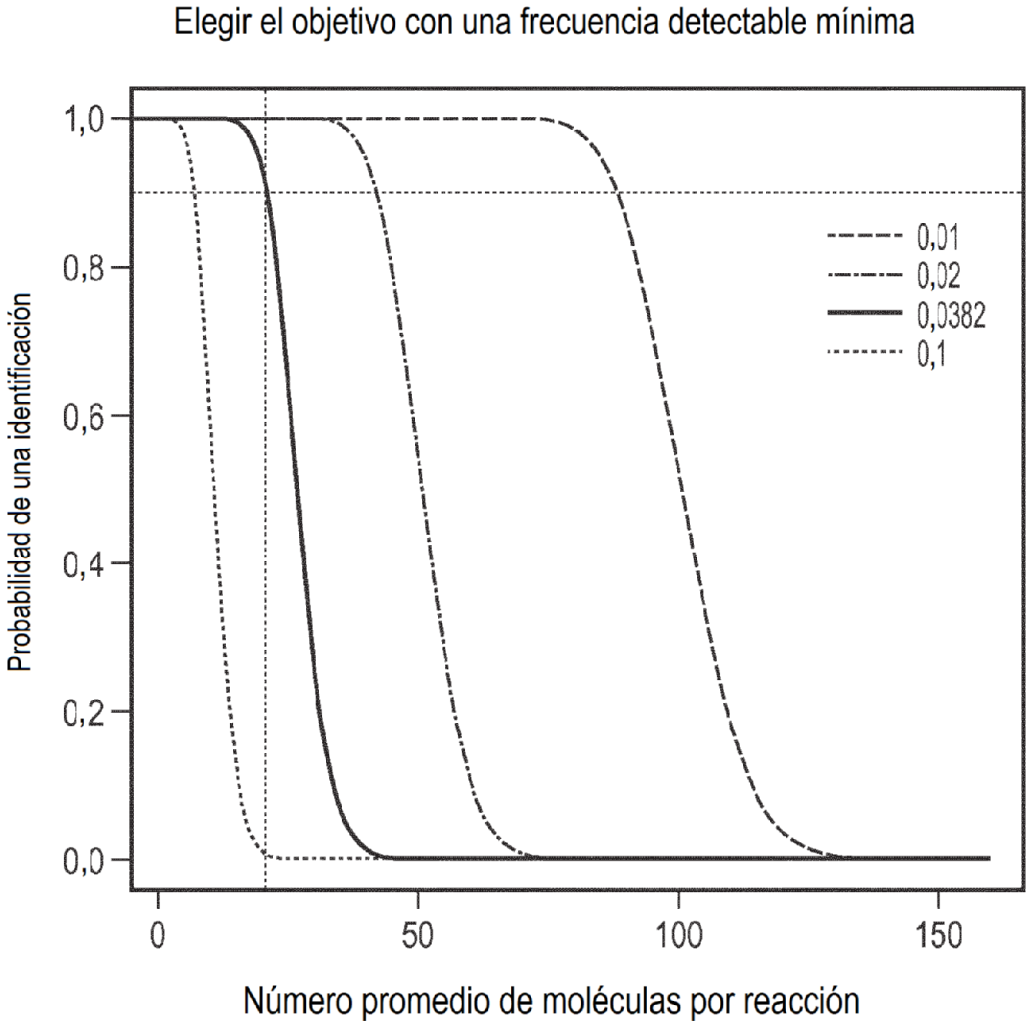
Elegir el objetivo con una frecuencia detectable mínima
Probabilidad de una identificación
1,0
0,8
0,6
0,4
0,2
0,0
0
50
100
150
Número promedio de moléculas por reacción
0,01
0,02
0,0382
0,1

FIG. 8

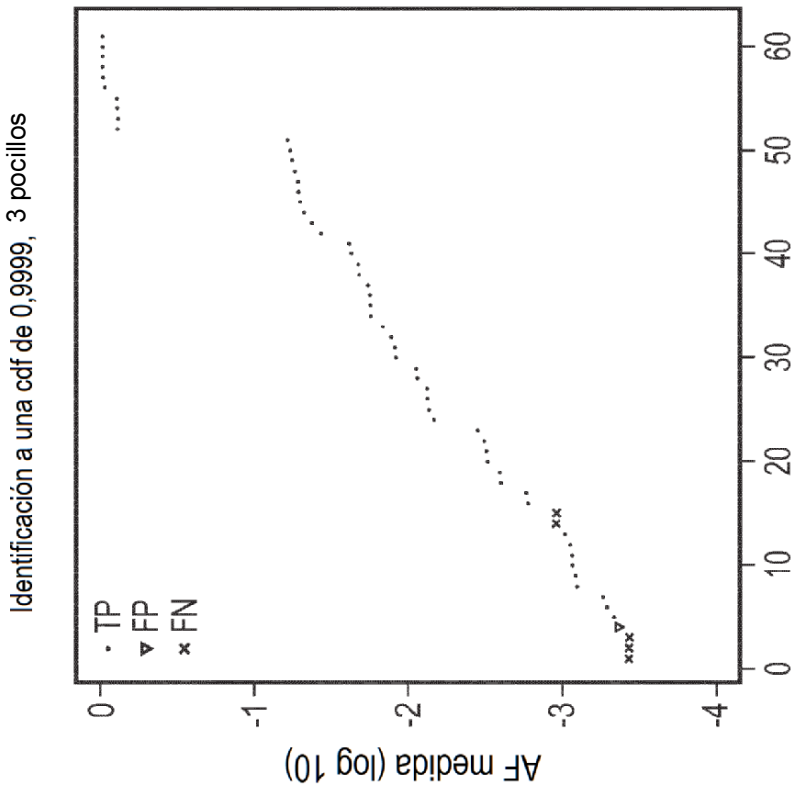
Identificación a una cdf de 0,9999, 3 pocillos
TP
FP
FN
AF medida (log 10)
0
-1
-2
-3
-4
0
10
20
30
40
50
60
Identificaciones

FIG. 9B

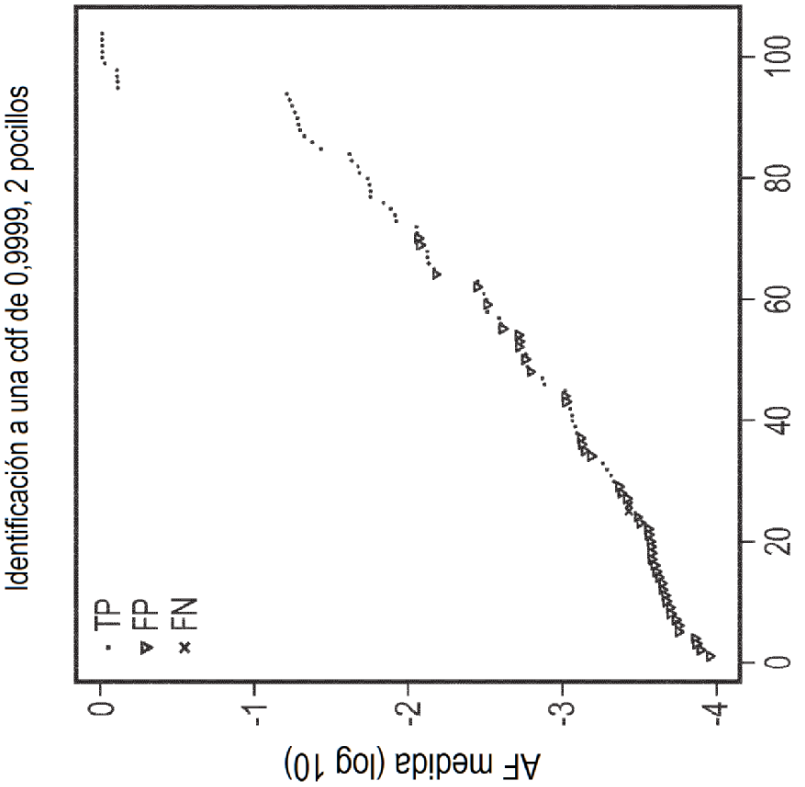
Identificación a una cdf de 0,9999, 2 pocillos
TP
FP
FN
AF medida (log 10)
0
-1
-2
-3
-4
0
20
40
60
80
100
Identificaciones

FIG. 9A

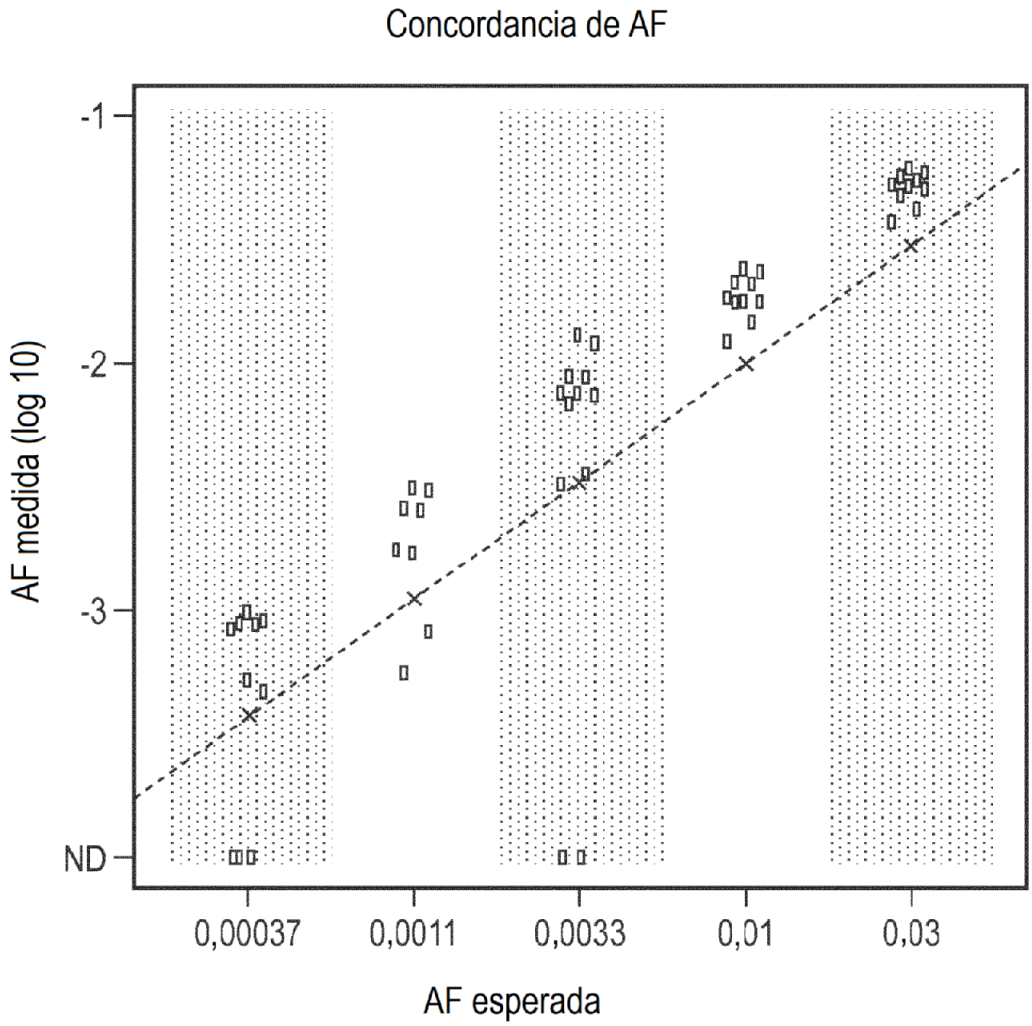
Concordancia de AF
AF medida (log 10)
-1
-2
-3
ND
0,00037
0,0011
0,0033
0,01
0,03
AF esperada

FIG. 10