(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 864 101**

(51) Int. CI.:

C12Q 1/6869 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **01.06.2017 E 17174059 (0)**

(97) Fecha y número de publicación de la concesión europea: **06.01.2021 EP 3409788**

(54) Título: **Método y sistema para la secuenciación de ácidos nucleicos**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
13.10.2021

(73) Titular/es:

SIEMENS HEALTHCARE GMBH (100.0%)
Henkestrasse 127
91052 Erlangen, DE

(72) Inventor/es:

HEUCKMANN, JOHANNES y
MARIOTTI, ERIKA

(74) Agente/Representante:

CARVAJAL Y URQUIJO, Isabel

DESCRIPCIÓN

Método y sistema para la secuenciación de ácidos nucleicos

La presente invención como se define en las reivindicaciones adjuntas se refiere a métodos y sistemas para la secuenciación de ácidos nucleicos. En particular, la presente invención se refiere a métodos y sistemas para reducir el número de falsos positivos en la secuenciación de ácidos nucleicos. El método comprende: alinear una pluralidad de lecturas genéticas; agrupar las lecturas genéticas en una pluralidad de grupos; crear una secuencia consenso para cada grupo de la pluralidad de grupos estableciendo una representación del nucleótido más abundante man_p o una etiqueta N basada en una relación r entre el número de lecturas genéticas dentro de un grupo de lecturas genéticas que tienen el nucleótido man_p más abundante en una posición específica del genoma de referencia y el número de lecturas que no tienen variante en dicha posición; e identificar la variación genética como una variación genética verdadera si una relación r* entre el número de secuencias consenso que comprenden la etiqueta N en una posición específica p y el número de secuencias consenso que comprenden la variación genética en la posición específica p está por debajo de un umbral t*.

La secuenciación de próxima generación (NGS) es una tecnología poderosa para la investigación básica y aplicaciones clínicas porque puede secuenciar cientos de miles de millones de nucleótidos de ADN en paralelo en un solo experimento. Sin embargo, ninguna tecnología de secuenciación está libre de errores, dando lugar de esta manera a llamadas de mutaciones falsas positivas. Hay diversas fuentes de errores introducidos artificialmente. Por ejemplo, pueden generarse mutaciones puntuales erróneas durante la amplificación por PCR a través de la incorporación del nucleótido incorrecto por las ADN polimerasas, durante la amplificación del complejo, la secuenciación de ADN y el análisis de imágenes de células de flujo. En consecuencia, las tecnologías de secuenciación paralela actuales muestran tasas de error de aproximadamente el 0,5 al 1 % (Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26: 1135-1145; Fox EJ, et al., (2014) Accuracy of Next Generation Sequencing Platforms. Next Generat Sequenc & Applic 1: 106. doi:10.4172/jngsa.1000106). Por lo tanto, del 0,5 al 1 % de todos los nucleótidos no son variantes reales, sino información de secuencia incorrecta que se genera durante el procesamiento de la muestra. La tasa de error intrínseca de NGS define un límite por debajo del cual la presencia de variantes verdaderas no es detectable de manera fiable. Por lo tanto, en las técnicas convencionales, las mutaciones con una fracción alélica mutante por debajo del 1 % no pueden detectarse de forma fiable. Sin embargo, la detección fiable de tales mutaciones o polimorfismo de un solo nucleótido (SNP) es particularmente crucial en caso de que tales variaciones se produzcan con una frecuencia menor.

En consecuencia, el problema técnico subyacente a la presente invención es la provisión de medios y métodos para proporcionar una forma confiable de secuenciación de ácidos nucleicos, en particular la identificación de variaciones en el genoma.

Este objeto se logra con las realizaciones que se proporcionan a continuación en el presente documento y con las características de las reivindicaciones independientes. Las reivindicaciones dependientes se refieren a aspectos adicionales de la invención. El método de acuerdo con la presente invención puede implementarse por ordenador. Sin embargo, el experto en la materia entenderá que existen también otras formas de implementar el método de acuerdo con la presente invención.

La presente invención como se define en las reivindicaciones adjuntas se refiere a métodos y sistemas para la secuenciación de ácidos nucleicos, particularmente para reducir el número de falsos positivos en la secuenciación de ácidos nucleicos. Tales métodos y sistemas de la invención pueden, por ejemplo, emplearse para identificar (una) variación o variaciones en la secuencia de ácido nucleico de un sujeto.

El método comprende las siguientes etapas:

(a) obtener una pluralidad de lecturas genéticas mediante la secuenciación de una muestra de ácido nucleico,
(b) alinear una pluralidad de lecturas genéticas con al menos una secuencia genética de referencia;
(c) agrupar las lecturas genéticas que comparten una posición genética en una secuencia genética de referencia de al menos una secuencia genética de referencia en una pluralidad de grupos;
(d) crear una secuencia consenso para cada grupo de la pluralidad de grupos, en donde se crea una secuencia consenso correspondiente determinando un nucleótido más abundante *man_p* en cada posición específica *p* de una pluralidad de posiciones dentro del grupo correspondiente de lecturas genéticas y

(i) establecer una representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación r entre el número de lecturas genéticas dentro del grupo correspondiente que tiene el nucleótido más abundante *man_p* en la posición específica *p* y el número de lecturas genéticas es superior o igual a un umbral predeterminado *t;* y
(ii) establecer una etiqueta *N* si la relación r está por debajo del umbral predeterminado *t* en donde dicho umbral es al menos aproximadamente el 76 %;

(e) comparar las secuencias consenso de la pluralidad de grupos con la secuencia genética de referencia en cada posición específica *p* de una pluralidad de posiciones de las secuencias consenso y en donde una diferencia en una posición específica entre las secuencias consenso y la secuencia genética de referencia indica una variación

en la posición específica;
(f) determinar el número de secuencias consenso que comprenden la variación en cada posición específica *p* de una pluralidad de posiciones y determinar el número de secuencias consenso que comprenden la etiqueta *N* en cada posición específica *p* de una pluralidad de posiciones; y
(g) identificar la variación en cada posición específica *p* de una pluralidad de posiciones como una verdadera variación si una relación *r** entre el número de secuencias consenso que comprenden la etiqueta *N* en la posición específica *p* y el número de secuencias consenso que comprenden la variación en la posición específica *p* está por debajo de un umbral *t** en donde dicho umbral t* es igual o inferior a 1,8.

Alternativamente, el método puede comprender

(a) obtener una pluralidad de lecturas genéticas mediante la secuenciación de una muestra de ácido nucleico, en donde la etapa de obtener es opcionalmente;
(b) alinear una pluralidad de lecturas genéticas con al menos una secuencia genética de referencia;
(c) agrupar las lecturas genéticas que comparten una posición genética en una secuencia genética de referencia en una pluralidad de grupos;
(d) crear una secuencia consenso en cada posición p específica de una pluralidad de posiciones para cada grupo de la pluralidad de grupos,
en donde una representación de un nucleótido en una posición específica de la pluralidad de posiciones se establece en una secuencia consenso correspondiente, si una relación entre el número del nucleótido en la posición específica y el número de las lecturas genéticas dentro del grupo correspondiente está por encima o es igual a un umbral predeterminado t, y en donde una etiqueta N se establece en la secuencia consenso correspondiente en la posición específica de la pluralidad de posiciones si la relación está por debajo del umbral predeterminado t;
(e) comparar las secuencias consenso de la pluralidad de grupos con la secuencia genética de referencia en cada posición específica *p* de una pluralidad de posiciones de las secuencias consenso y en donde una diferencia en una posición específica entre las secuencias consenso y la secuencia genética de referencia indica una variación en la posición específica;
(f) determinar el número de secuencias consenso que comprenden la variación en cada posición específica *p* de una pluralidad de posiciones y determinar el número de secuencias consenso que comprenden la etiqueta N en cada posición específica *p* de una pluralidad de posiciones; y
(g) identificar la variación en cada posición específica *p* de una pluralidad de posiciones como una verdadera variación si una relación *r** entre el número de secuencias consenso que comprenden la etiqueta *N* en la posición específica *p* y el número de secuencias consenso que comprenden la variación en la posición específica *p* está por debajo de un umbral *t**.

Las explicaciones y definiciones proporcionadas en lo siguiente se aplican a todos los métodos y sistemas proporcionados en el presente documento cambiando lo que se tenga que cambiar. Los métodos y sistemas de la invención se refieren a la secuenciación de ácidos nucleicos. En consecuencia, el método de la invención podría comprender opcionalmente obtener una pluralidad de lecturas genéticas mediante la secuenciación de una muestra de ácido nucleico y los sistemas de la invención podrían comprender opcionalmente una unidad de obtención estando configurada para obtener la pluralidad de lecturas genéticas de una muestra de ácido nucleico.

En el Ejemplo ilustrativo 1, se demuestra que los métodos y sistemas de la invención muestran una alta sensibilidad y especificidad (PPV) en la detección de variaciones (mutaciones), por ejemplo, que se producen en el genoma con una frecuencia alélica menor (MAF ≤ 1 %). Los ejemplos adjuntos documentan el análisis de variaciones genéticas de cuatro diluciones de tres líneas celulares normales HapMap; véase, por ejemplo, la Tabla 1. El ADN de tales líneas celulares se analizó mediante los métodos y sistemas de la invención. El método y sistema proporcionados en el presente documento incluyen etapas de filtrado particulares que usan el nivel de confiabilidad para cada llamada de nucleótidos. Estas etapas se describen en lo siguiente. En las Figuras 2 a 5 se ilustra un método de ejemplo. En los ejemplos adjuntos, las secuencias adaptadoras se retiran de los datos de secuenciación sin procesar (las lecturas genéticas) y se almacenan dentro del mismo archivo. Las lecturas genéticas se alinean con al menos un genoma de referencia. En una siguiente etapa, la información de alineación así como las secuencias retiradas, por ejemplo, secuencias de códigos de barras, se usan para formar grupos de lecturas que comparten la misma posición genómica y secuencia de código de barras. Dicho grupo también puede denominarse "familia". En una etapa adicional, se crean secuencias consenso, por ejemplo, para cada grupo. En la formación de secuencias consenso, se consulta nucleótido por nucleótido. En particular, una representación del nucleótido en una posición solo se establece en la secuencia consenso si el nucleótido se produce con al menos un nivel de umbral particular entre las lecturas de un grupo, por ejemplo, al menos el 76 %. Esto significa que una representación del nucleótido más abundante (por ejemplo, T, A, C, G o U) en la posición solo se establece si dicho umbral se cumple (por encima o igual al umbral). Si no se cumple esta condición, se coloca una etiqueta "N" en dicha posición. Las "variaciones genéticas" o "variaciones" (mutaciones) pueden detectarse si se identifica una diferencia entre la secuencia consenso y el genoma de referencia. En los ejemplos adjuntos se documenta que un procedimiento tal puede, por ejemplo, proporcionar una sensibilidad y especificidad para la detección de variaciones en las cuatro diluciones de las líneas celulares del 86,5 % y el 65,5 %, respectivamente.

Como se demuestra en los ejemplos adjuntos, se implementó una segunda etapa de filtrado ("Filtro N") en el método descrito anteriormente; véanse, por ejemplo, las Figuras 4 y 5. Se demostró inesperadamente en los ejemplos adjuntos que la segunda etapa de filtrado mejora además la especificidad y sensibilidad para detectar variaciones genéticas; véase, por ejemplo, la Tabla 5. En los Ejemplos adjuntos, se observó que las regiones genómicas que contienen una sustitución introducida erróneamente durante el flujo de trabajo de secuenciación exhiben una tasa menor de nucleótidos consenso y por lo tanto un número más alto de la etiqueta "N" en comparación con las regiones que contienen variaciones genéticas "verdaderas". Estas regiones a menudo se definen mediante repeticiones de secuencia (por ejemplo, Nakamura K, Sequence- specific error profile of Illumina sequencers. Nucleic Acids Res. julio de 2011;39(13):e90. doi: 10.1093/nar/gkr344, o Schirmer M, Illumina error profiles:resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics. 11 de marzo de 2016;17:125. doi: 10.1186/s12859-016-0976-y). Tales repeticiones frecuentemente conducen a la incorporación de nucleótidos incorrectos, aumentando por lo tanto la probabilidad de una incorporación de "N" en esta posición. Por lo tanto, siendo la abundancia de "N" un sustituto de las posiciones secuenciadas que contienen alto ruido de fondo. Por lo tanto, los métodos y sistemas proporcionados en el presente documento incluyen el filtro adicional que usa la abundancia de "N" en una posición de variación para discriminar la mutación verdadera de una sustitución errónea. En consecuencia, se determina el número de secuencias consenso que tienen la variación genética o que tienen N en cada posición específica de una pluralidad de posiciones (o una posición específica) en las secuencias consenso. Posteriormente, se calcula la relación entre el número de secuencias consenso que contienen "N" y las que contienen la variación genética; véanse, por ejemplo, las Figuras 4 y 5. En consecuencia, las llamadas verdaderas pueden distinguirse de las llamadas falsas positivas. Un efecto tan inesperado se documenta en los ejemplos adjuntos. Este segundo filtro también se denomina en el presente documento el "Filtro N". En los Ejemplos adjuntos, el análisis de las cuatro diluciones de las líneas celulares por el método que incluye el filtro N dio como resultado una sensibilidad y especificidad de, por ejemplo, el 82,5 % y el 80,5 %, respectivamente; véase, por ejemplo, la Tabla 5. En consecuencia, un método tal permite retirar una amplia gama de llamadas de variación de falso positivo. De esta manera, la especificidad se mejora considerablemente. Por lo tanto, la secuenciación de ácidos nucleicos y, en particular, la identificación de variaciones genéticas (llamada de mutación) se mejora mediante los métodos y sistemas proporcionados en el presente documento.

Como se indica anteriormente, el filtro N emplea una relación entre el número de secuencias consenso que comprenden la variación en una posición específica y el número de secuencias consenso que comprenden N en la posición específica. El umbral de esta relación (r*) mejora la especificidad y la sensibilidad. Por ejemplo, si se aplica un umbral de 2 para la relación (r*), se consigue una sensibilidad del 82,5 % y una especificidad del 80,5 % como se demuestra en el Ejemplo ilustrativo 1. Si se aplica un umbral de 4 para la relación (r*), se proporciona una sensibilidad del 87,9 % y una especificidad del 77,4 % como se muestra en el Ejemplo ilustrativo 2.

Kennedy et al., 2014 (Nature Protocols, Vol. 9 N.º 11) incorporaron "N" en la secuencia consenso. Sin embargo, la información "N" no se empleó activamente en la llamada de la variación genética. Por lo tanto, un método tal de la técnica anterior no incluye el beneficioso Filtro N de la invención y por lo tanto es comparable a un análisis sin emplear el Filtro N. En consecuencia, los métodos proporcionados en el presente documento se mejoran con respecto a la técnica anterior.

En resumen, la presente invención tiene, entre otros, las siguientes ventajas sobre los métodos y sistemas de la técnica anterior. Los métodos y sistemas ventajosos proporcionan una secuenciación de ácidos nucleicos con una alta especificidad y sensibilidad. En consecuencia, se reduce el número de falsos positivos en la secuenciación de ácidos nucleicos.
En consecuencia, los métodos y sistemas proporcionados en el presente documento proporcionan una secuenciación de ácidos nucleicos fiable.

Los métodos y sistemas de la invención se refieren a la secuenciación de ácidos nucleicos. En consecuencia, los métodos de la invención y los sistemas empleados en un método tal incluyen opcionalmente la etapa de realizar la secuenciación de ácido nucleico de una muestra de ácido nucleico. El método y los sistemas de la invención también pueden realizarse con lecturas genéticas obtenidas. La secuenciación de ácidos nucleicos es un método para obtener la pluralidad de lecturas genéticas para un ácido nucleico. Por lo tanto, el método de la invención puede comprender la secuenciación de ácidos nucleicos para obtener la pluralidad de lecturas genéticas de un ácido nucleico comprendido en una muestra de ácido nucleico. La persona experta sabe cómo realizar la secuenciación de ácidos nucleicos, es decir, cómo determinar la secuencia de ácido nucleico de una muestra de ácido nucleico.

En consecuencia, el método de la invención puede comprender:

(a) obtener una pluralidad de lecturas genéticas mediante la secuenciación de una muestra de ácido nucleico;
(b) alinear la pluralidad de lecturas genéticas con al menos una secuencia genética de referencia;
(c) agrupar las lecturas genéticas que comparten una posición genética en una secuencia genética de referencia de la al menos una secuencia genética de referencia en una pluralidad de grupos;
(d) crear una secuencia consenso para cada grupo de la pluralidad de grupos, en donde se crea una secuencia consenso correspondiente determinando un nucleótido más abundante *man_p* en cada posición específica *p* de una pluralidad de posiciones dentro del grupo correspondiente de lecturas genéticas y

(i) establecer una representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación r entre el número de lecturas genéticas dentro del grupo correspondiente que tiene nucleótido más abundante *man_p* en la posición específica *p* y el número de lecturas genéticas es superior o igual a un umbral predeterminado t; y
(ii) establecer una etiqueta *N* si la relación *r* está por debajo del umbral predeterminado *t*;

(e) comparar las secuencias consenso de la pluralidad de grupos con la secuencia genética de referencia en cada posición específica *p* de una pluralidad de posiciones de las secuencias consenso y en donde una diferencia en una posición específica entre las secuencias consenso y la secuencia genética de referencia indica una variación en la posición específica;
(f) determinar el número de secuencias consenso que comprenden la variación en cada posición específica *p* de una pluralidad de posiciones y determinar el número de secuencias consenso que comprenden la etiqueta *N* en cada posición específica *p* de una pluralidad de posiciones; y
(g) identificar la variación en cada posición específica *p* de una pluralidad de posiciones como una verdadera variación si una relación *r** entre el número de secuencias consenso que comprenden la etiqueta *N* en la posición específica *p* y el número de secuencias consenso que comprenden la variación en la posición específica *p* está por debajo de un umbral *t**.

Como se usa en el presente documento, la expresión "obtener la pluralidad de lecturas genéticas de una muestra de ácido nucleico" significa que se determina la secuencia de ácido nucleico de los ácidos nucleicos/moléculas de ácido nucleico comprendidas en la muestra y por lo tanto se determina una pluralidad de las lecturas genéticas.

En particular, la pluralidad de lecturas genéticas se obtiene o determina mediante secuenciación de una muestra de ácido nucleico. La muestra de ácido nucleico como se usa en el presente documento comprende una o más moléculas de ácido nucleico. Antes de obtener una pluralidad de lecturas genéticas, las moléculas de ácido nucleico comprendidas en la muestra podrían amplificarse. La pluralidad de lecturas genéticas podría determinarse mediante secuenciación de ácidos nucleicos.

Como se usa en el presente documento, los términos y expresiones "lectura genética", "lectura" o "lectura de secuenciación", como se usan en el presente documento de manera intercambiable, o una variante gramatical de los mismos se refieren a información sobre una secuencia de ácido nucleico que se obtiene de una molécula de ácido nucleico comprendida en la muestra de ácido nucleico. Por lo tanto la lectura genética significa un tramo de nucleótidos contiguos que es una representación de un tramo correspondiente en la muestra de ácido nucleico. La lectura genética puede estar representada simbólicamente por la secuencia de nucleótidos (por ejemplo, A, T, C, G o U). La lectura genética puede obtenerse como la salida de un secuenciador de ácido nucleico (unidad de obtención). En consecuencia, las lecturas genéticas se obtienen mediante el secuenciador de ácidos nucleicos. Por lo tanto, la lectura genética puede entenderse como la señal recibida. En otras palabras, la lectura genética puede ser la secuencia de datos sin procesar. Por lo tanto, la lectura genética puede comprender un error que se introduce artificialmente por la unidad de obtención que lee la muestra de ácido nucleico o durante la preparación del ADN para permitir la secuenciación del mismo. Puede almacenarse en un dispositivo de memoria y procesarse según sea apropiado para determinar si coincide con una secuencia de referencia o si cumple con otros criterios. Puede obtenerse una lectura directamente de un aparato de secuenciación o indirectamente de la información de secuencia almacenada relativa a la muestra. En algunos casos, una lectura es una secuencia de longitud suficiente (por ejemplo, al menos aproximadamente 30 nucleótidos) que puede usarse para identificar una secuencia o región más grande, por ejemplo, que puede alinearse y asignarse específicamente a un cromosoma o región genómica o gen.

La pluralidad de lecturas genéticas podría determinarse mediante secuenciación de ácidos nucleicos. En consecuencia, la invención puede referirse particularmente a un método que comprende realizar la secuenciación de ácido nucleico de una o más moléculas de ácido nucleico comprendidas en una muestra de ácido nucleico y determinar de esta manera una pluralidad de lecturas genéticas.

Antes de realizar la secuenciación de ácidos nucleicos de una muestra de ácido nucleico y, por lo tanto, determinar una pluralidad de lecturas genéticas, las moléculas de ácido nucleico pueden procesarse. Particularmente, las moléculas de ácido nucleico pueden amplificarse.

En la secuenciación de ácidos nucleicos y la posterior alineación y agrupación como se proporciona en el presente documento, pueden emplearse adaptador o adaptadores. El uso de adaptadores es una forma de ejemplo de facilitar la alineación y agrupación. Sin embargo, también podrían usarse otros medios en tales etapas.
El (a) adaptador puede estar unido a las moléculas de ácido nucleico que van a secuenciarse y que están comprendidas en la muestra de ácido nucleico. En consecuencia, los métodos y sistemas de la invención podrían comprender: (i) unir al menos un adaptador a una molécula de ácido nucleico para generar moléculas de ácido nucleico etiquetadas, (ii) amplificar las moléculas de ácido nucleico etiquetadas para producir amplicones etiquetados; y (iii) determinar la secuencia de ácido nucleico de los amplicones etiquetados, por ejemplo, mediante secuenciación de ácidos nucleicos. Por lo tanto, la determinación de la pluralidad de lecturas genéticas y, por lo tanto, la secuenciación de ácidos nucleicos que podría realizarse en los métodos y sistemas de la invención puede comprender: (i) unir al

menos un adaptador a una molécula de ácido nucleico para generar moléculas de ácido nucleico etiquetadas, (ii) amplificar las moléculas de ácido nucleico etiquetadas para producir amplicones etiquetados; y (iii) determinar la secuencia de ácido nucleico de los amplicones etiquetados mediante secuenciación de ácido nucleico y determinar de esta manera una pluralidad de lecturas genéticas. En consecuencia, el método de la invención podría comprender:

(a) obtener una pluralidad de lecturas genéticas mediante la secuenciación de ácidos nucleicos de una muestra de ácido nucleico que comprende:

(i) unir una pluralidad de adaptadores a moléculas de ácido nucleico comprendidas en la muestra para generar moléculas de ácido nucleico etiquetadas,
(ii) amplificar las moléculas de ácido nucleico etiquetadas para producir amplicones etiquetados; y
(iii) determinar la secuencia de ácido nucleico de los amplicones etiquetados mediante secuenciación de ácido nucleico y determinar de esta manera una pluralidad de lecturas genéticas;

(b) alinear la pluralidad de lecturas genéticas con al menos una secuencia genética de referencia;
(c) agrupar las lecturas genéticas que comparten una posición genética en una secuencia genética de referencia de la al menos una secuencia genética de referencia en una pluralidad de grupos;
(d) crear una secuencia consenso para cada grupo de la pluralidad de grupos, en donde se crea una secuencia consenso correspondiente determinando un nucleótido más abundante *man_p* en cada posición específica *p* de una pluralidad de posiciones dentro del grupo correspondiente de lecturas genéticas y

(i) establecer una representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación r entre el número de lecturas genéticas dentro del grupo correspondiente que tiene el nucleótido más abundante *man_p* en la posición específica *p* y el número de lecturas genéticas dentro del un grupo correspondiente es superior o igual a un umbral predeterminado *t;* y
(ii) establecer una etiqueta *N* si la relación *r* está por debajo del umbral predeterminado *t;*

(e) comparar las secuencias consenso de la pluralidad de grupos con la secuencia genética de referencia en cada posición específica *p* de una pluralidad de posiciones de las secuencias consenso y en donde una diferencia en una posición específica entre las secuencias consenso y la secuencia genética de referencia indica una variación en la posición específica;
(f) determinar el número de secuencias consenso que comprenden la variación en cada posición específica *p* de una pluralidad de posiciones y determinar el número de secuencias consenso que comprenden la etiqueta *N* en cada posición específica *p* de una pluralidad de posiciones; y
(g) identificar la variación en cada posición específica *p* de una pluralidad de posiciones como una verdadera variación si una relación *r** entre el número de secuencias consenso que comprenden la etiqueta *N* en la posición específica *p* y el número de secuencias consenso que comprenden la variación en la posición específica *p* está por debajo de un umbral *t**.

En los métodos y sistemas de la invención, los adaptadores predefinidos pueden ligarse a las moléculas de ácido nucleico. Los términos "adaptor" y "adaptador" se usan indistintamente en este documento. El adaptador o la molécula adaptadora empleados en los métodos y sistemas de la presente invención puede incluir al menos una secuencia de código de barras. Además, el adaptador puede incluir un adaptador de ligamiento. Opcionalmente, el adaptador incluye además al menos uno, preferentemente al menos dos sitios de unión del cebador de PCR, al menos uno, preferentemente al menos dos sitios de unión del cebador de secuenciación, o ambos. En la Figura 1 se ilustra una molécula adaptadora de ejemplo. En los métodos y sistemas de la invención, al menos un adaptador puede unirse/ligarse a la molécula de ácido nucleico. En particular, pueden unirse uno o dos adaptador o adaptadores a una molécula de ácido nucleico para generar una molécula de ácido nucleico etiquetada. Por ejemplo, pueden unirse dos adaptadores a la molécula de ácido nucleico, por ejemplo, en cualquier extremo de la molécula de ácido nucleico (extremo 5' y 3').
Como se usa en el presente documento, unir el adaptador a la molécula de ácido nucleico significa que el adaptador o adaptadores está/están ligados a la molécula de ácido nucleico.

La persona experta conoce los medios para ligar dos moléculas de ácido nucleico juntas. Por ejemplo, los Kits de reactivos TruSeq Exome (Illumina) o SureSelectXT (Agilent Technologies). Como se detalla a continuación, pueden usarse adaptadores de ligación para unir el adaptador o adaptadores a la molécula de ácido nucleico.

Una molécula de ácido nucleico que comprende la molécula o moléculas adaptadoras, es decir, que se une al adaptador o adaptadores, se designa en el presente documento "molécula de ácido nucleico etiquetada". Por lo tanto, para determinar la secuencia de ácido nucleico de las moléculas de ácido nucleico comprendidas en la muestra, la unión del al menos un adaptador a la molécula de ácido nucleico podría generar moléculas de ácido nucleico etiquetadas.

Como se ha descrito anteriormente, la molécula adaptadora puede comprender al menos una secuencia de código de

barras. Particularmente, la molécula adaptadora puede comprender una secuencia de código de barras. Una secuencia de código de barras también se denomina en el presente documento secuencia de "identificador de molécula única" (SMI). La secuencia de código de barras significa una secuencia individual. En los métodos proporcionados en el presente documento, podría unirse un código de barras a la molécula de ácido nucleico comprendida en la muestra de ácido nucleico y de la cual se obtiene la pluralidad de lecturas genéticas. En consecuencia, las lecturas genéticas detectadas, por ejemplo, en la secuenciación de ácidos nucleicos, puede comprender tal secuencia o secuencias de código de barras y las lecturas detectadas pueden por lo tanto etiquetarse con secuencia o secuencias de ácido nucleico individuales/únicas. Por lo tanto, las lecturas pueden comprender las secuencias de códigos de barras además de su información genética. Por lo tanto, las lecturas genéticas pueden distinguirse y clasificarse durante el análisis de datos en función de su secuencia o secuencias de códigos de barras.

La secuencia de código de barras como se usa en el presente documento también puede ser un identificador de molécula única que se produce de forma natural en la molécula de ácido nucleico. Dicho código de barras de origen natural también puede emplearse en la agrupación de lecturas genéticas como se describe a continuación.

La secuencia de código de barras puede ser una secuencia bicatenaria, complementaria o una secuencia monocatenaria. La secuencia de código de barras puede estar degenerada o semidegenerada. La secuencia de código de barras degenerada o semidegenerada también puede ser una secuencia degenerada aleatoria. Una secuencia de código de barras bicatenaria incluye una primera secuencia de código de barras degenerada o semidegenerada y una segunda secuencia de código de barras que es complementaria a la primera secuencia de código de barras degenerada o semidegenerada, mientras que una secuencia de código de barras monocatenaria incluye una primera secuencia de código de barras de nucleótidos degenerada o semidegenerada. La primera y/o segunda secuencias de códigos de barras degeneradas o semidegeneradas pueden tener cualquier longitud adecuada para producir un número suficientemente grande de etiquetas únicas para marcar/etiquetar las moléculas de ácido nucleico de la muestra de ácido nucleico, que puede procesarse adicionalmente, por ejemplo, fragmentarse como se describió anteriormente.

La secuencia de código de barras es un tramo corto de nucleótidos de aproximadamente 3 a 20 nucleótidos de longitud. Por lo tanto, la secuencia del código de barras puede tener aproximadamente 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 nucleótidos de longitud. En principio, puede usarse cualquier longitud siempre que la información de la secuencia de código de barras sea suficiente para distinguirse y clasificarse durante el análisis de las secuencias.

La secuencia de la secuencia de código de barras puede comprender cualquier nucleótido de origen natural, por ejemplo, adenina (A), citosina (C), guanina (G), timina (T) o uracilo (U), o nucleótidos de ADN o ARN no naturales o sustancias similares a nucleótidos o análogos con propiedades de emparejamiento de bases (por ejemplo, xantosina, inosina, hipoxantina, xantina, 7-metilguanina, 7-metilguanosina, 5,6-dihidrouracilo, 5-metilcitosina, dihidouridina, isocitosina, isoguanina, desoxinucleósidos, nucleósidos, ácidos nucleicos peptídicos, ácidos nucleicos bloqueados, ácidos nucleicos de glicol y ácidos nucleicos de treosa). Las secuencias de códigos de barras pueden generarse mediante un método mediado por polimerasa o pueden generarse preparando e hibridando una biblioteca de oligonucleótidos individuales de secuencia conocida.

Como se usa en el presente documento, unir una pluralidad de adaptadores a las moléculas de ácido nucleico significa que se une más de un adaptador a la molécula o moléculas de ácido nucleico comprendidas en la muestra y/o que se emplean más de un adaptador diferente en la etapa de unión. Por ejemplo, los adaptadores pueden diferir en la secuencia del código de barras que se incluye en los adaptadores adjuntos.

Por ejemplo, la secuencia de código de barras puede ser una secuencia de nucleótidos degenerada aleatoria que tiene una longitud de 12 nucleótidos. Una secuencia de código de barras de 12 nucleótidos que se adjunta/liga a cada extremo de la molécula de ácido nucleico, como se describe en el Ejemplo a continuación, resulta en la generación de hasta 4^{24} (es decir, 2,8 x 10^{14}) secuencias de etiquetas distintas. En aspectos particulares de la invención, cada molécula de ácido nucleico está unida o marcada con dos adaptadores, en donde cada adaptador comprende una secuencia de código de barras. Las secuencias de códigos de barras pueden ser complementarias. Una realización de ejemplo tal se representa como α y β en la Figura 1 A que muestra una ilustración esquemática de fragmentos de ADN cortados marcados con secuencias de nucleótidos predefinidas bicatenarias. Por lo tanto, en esta figura, los códigos de barras están representados por "α" y "β". Los números 1 y 2 que se muestran en la Figura 1A representan las colas compatibles con la célula de flujo de Illumina (véase, por ejemplo, http://www.illumina.com/documents/products/techspotlights/tec hspotlight_sequencing.pdf para obtener información sobre el flujo de células Illumina).

El adaptador puede unirse/ligarse a un extremo o a ambos extremos de la molécula de ácido nucleico (diana). No es necesario incluir códigos de barras en ambos extremos del adaptador siempre que pueda determinarse qué cadena es cuál.
Como se usa en el presente documento, el "adaptador de ligamiento" puede ser cualquier adaptador de ligamiento adecuado que sea complementario a un adaptador de ligación añadido a una secuencia de ácido nucleico diana bicatenaria que incluye, pero no limitado a un saliente T, un saliente A, un saliente CG, un extremo romo o cualquier

otra secuencia ligable. El adaptador de ligamiento puede fabricarse usando un método para cola A o cola T con extensión de polimerasa; creando un saliente con una enzima diferente; usando una enzima de restricción para crear un saliente de nucleótido único o múltiples, escisión basada en transposones seguida de etiquetado con adaptador o cualquier otro método conocido en la técnica.

Como se ha descrito anteriormente, la molécula adaptadora puede incluir al menos dos sitios de unión del cebador de PCR, tales como sitios de unión de "célula de flujo". Por ejemplo, un sitio de unión de cebador de PCR directo (o un sitio de unión de "célula de flujo 1" (FC1)) y un sitio de unión de cebador de PCR inversa (o un sitio de unión de "célula de flujo 2" (FC2)). La molécula adaptadora puede incluir también al menos dos sitios de unión del cebador de secuenciación, correspondiendo cada uno a una lectura de secuenciación. Alternativamente, los sitios de unión de los cebadores de secuenciación pueden añadirse en una etapa separada mediante la inclusión de las secuencias necesarias como colas a los cebadores de PCR, o por ligamiento de las secuencias necesarias. Por lo tanto, si una molécula de ácido nucleico diana bicatenario tiene una molécula adaptadora SMI ligada a cada extremo, cada cadena secuenciada tendrá dos lecturas - una lectura directa y otra inversa.

Posteriormente a la etapa de unir la pluralidad de adaptadores a las moléculas de ácido nucleico, las moléculas etiquetadas de ácido nucleico pueden amplificarse. Esta amplificación genera un conjunto de productos de ácidos nucleicos amplificados marcados/etiquetados de forma única, es decir, amplicones. Los métodos de amplificación son conocidos en la técnica (por ejemplo, un método de PCR o distinto de PCR).

Como se muestra en los ejemplos adjuntos, la amplificación por PCR de las moléculas de ácido nucleico ligadas al adaptador o adaptadores da como resultado dos productos de PCR diferentes (amplicones). Las moléculas obtenidas de la amplificación de la cadena superior tienen, por ejemplo, un código de barras "α" cerca de la secuencia de la célula de flujo número 1 (FC1) y un código de barras "β" al lado de la secuencia de la célula de flujo número 2 (FC2) - denominado "familias αβ" en lo sucesivo en el presente documento. Los amplicones resultantes de la amplificación de la cadena inferior tienen las dos cadenas complementarias etiquetadas recíprocamente. Dichos amplicones se denominan "familias βα" en lo sucesivo en el presente documento, como también puede verse en la Figura ilustrativa 1B, que muestra dos moléculas bicatenarias diferentes resultantes de la amplificación por PCR de las moléculas de ácido nucleico etiquetadas; mostradas en la Figura 1A. Los amplicones resultantes de la cadena superior tienen el código de barras α cerca de la secuencia de la célula de flujo número 1 y el código de barras β cerca de la secuencia de la célula de flujo número 2, es decir, las "familias αβ". Los amplicones resultantes de la amplificación de la cadena inferior tienen las dos cadenas complementarias etiquetadas recíprocamente, denominadas en el presente documento las "familias βα".

Las moléculas amplificadas, es decir, los amplicones o la progenie, pueden secuenciarse después usando cualquier método adecuado conocido en la técnica. La secuencia de ácido nucleico de la muestra de ácido nucleico puede determinarse mediante la plataforma de secuenciación de lllumina, plataforma de secuenciación ABI SOliD, plataforma de secuenciación de Pacific Biosciences, plataforma de secuenciación 454 Life Sciences, plataforma de secuenciación Ion Torrent, plataforma de secuenciación Helicos y tecnología de secuenciación de nanoporos. Por ejemplo, se incorporan nucleótidos marcados con fluorescencia. Pueden obtenerse imágenes de la célula de flujo y puede registrarse la emisión de cada grupo. La longitud de onda y la intensidad de la emisión pueden usarse para identificar los nucleótidos incorporados. El ciclo se repite "n" veces para crear una longitud de lectura genética de "n" nucleótidos. En consecuencia, en la etapa de secuenciación de ácidos nucleicos, las lecturas genéticas se determinan. Las lecturas genéticas representan la secuencia de ácido nucleico de las moléculas de ácido nucleico que están comprendidas en la muestra de ácido nucleico.

Como se usa en el presente documento, "determinar" u "obtener" las lecturas genéticas significa que se determina la secuencia de ácido nucleico de las moléculas de ácido nucleico. Una pluralidad de lecturas genéticas significa que se determina más de una lectura genética. El método puede comprender secuenciar un subconjunto del conjunto de moléculas de ácido nucleico suficiente para producir lecturas genéticas para al menos una progenie de cada una de al menos el 20 %, al menos el 30 %, al menos el 40 %, al menos el 50 %, al menos el 60 %, al menos el 70 %, al menos el 80 %, al menos el 90 %, al menos el 95 %, al menos el 98 %, al menos el 99 %, al menos el 99,9 % o al menos el 99,99% de las moléculas de ácido nucleico comprendidas en la muestra de ácido nucleico. La al menos una progenie puede ser una pluralidad de progenie, por ejemplo, al menos 2, al menos 5 o al menos 10 descendientes. Como se indica anteriormente, los métodos y sistemas de la invención también pueden usar lecturas genéticas que se han determinado anteriormente. Esto es, los métodos y sistemas de la invención también pueden usar una pluralidad determinada u obtenida de secuencias genéticas.

Se conocen en la técnica diversas moléculas adaptadoras, por ejemplo, el documento WO 2013/142389. Por ejemplo, una molécula adaptadora usada en el método de la invención puede formar una "forma de Y" o una "forma de horquilla". El adaptador "en forma de Y" permite que ambas cadenas se amplifiquen independientemente mediante un método de PCR antes de la secuenciación porque tanto la cadena superior como la inferior tienen sitios de unión para los cebadores de PCR FC1 y FC2.

Además, Se prevé en este documento que las mutaciones introducidas artificialmente, por ejemplo, durante la

amplificación puede reducirse mediante el uso de diferentes polimerasas durante la primera o primeras rondas de PCR. Además de las polimerasas, antes de la amplificación podrían usarse otras enzimas modificadoras/reparadoras de ADN para convertir el daño de un tipo que no da una firma mutagénica específica en otro tipo que lo haga con cualquier polimerasa que se use. Alternativamente, podrían usarse enzimas modificadoras/reparadoras de ADN para eliminar las bases dañadas y podrían secuenciarse ambas cadenas de ADN con y sin el tratamiento enzimático. Por lo tanto puede inferirse que las mutaciones en el ADN monocatenario que van a retirarse mediante el tratamiento enzimático se deben al daño del ADN. Esto podría ser útil en ADNmt o nuclear humano pero también podría ser útil con organismos modelo (ratones, levaduras, bacterias, etc.), tratados con diferentes agentes dañinos nuevos, facilitando una pantalla de compuestos que dañan el ADN que sería análoga a la prueba de Ames ampliamente usada.

Como se usa en el presente documento, las expresiones "secuencia genética de referencia", "genoma de referencia" y "secuencia de referencia" pueden usarse indistintamente en el presente documento y se refiere a cualquier secuencia del genoma conocida particular, ya sea parcial o completa, de cualquier organismo que pueda usarse para hacer referencia a secuencias identificadas de un sujeto. Por ejemplo, un genoma de referencia usado para sujetos humanos así como muchos otros organismos se encuentra en el National Center for Biotechnology Information en www.ncbi.nlm.nih.gov. Un "genoma" se refiere a la información genética conocida de un organismo.

Como se usa en el presente documento, los términos "alineado", "alineamiento" o "alineación" o variantes gramaticales de las mismas se refieren al proceso de comparar una secuencia de ácido nucleico, por ejemplo, una lectura o etiqueta genética, con una secuencia genética de referencia y, por tanto, determinar si la secuencia genética de referencia contiene la secuencia de ácido nucleico, por ejemplo, la lectura genética o partes de la misma. Además, si la secuencia genética de referencia comparte información genética con la lectura genética, la lectura puede mapearse y asignarse a la posición particular que comparte la totalidad o subpartes de la información genética sobre la secuencia genética de referencia. La lectura genética comparte información genética con la secuencia genética de referencia si existe similitud/identidad entre las secuencias.
Como se usa en el presente documento, la expresión "alinear una pluralidad de lecturas genéticas con al menos una secuencia genética de referencia" significa que las lecturas genéticas que se determinan en la secuenciación de ácido nucleico o un subconjunto de las mismas están alineadas con la al menos una secuencia genética de referencia. De esta manera, las lecturas genéticas están alineadas con áreas de similitud que pueden estar asociadas con características específicas que están más altamente conservadas que otras regiones que se producen en la al menos una secuencia de referencia genética. Como se usa en el presente documento, el término "similitud" puede referirse a la identidad de secuencia. Las áreas de similitud tienen un porcentaje específico de nucleótidos que son iguales, cuando se compara y alinea para una correspondencia máxima sobre una región designada (que se define a continuación, por ejemplo, 72 nucleótidos) según se mide usando un algoritmo de comparación de secuencias como se conoce en la técnica, o por alineación manual e inspección visual. Las secuencias que tienen, por ejemplo, identidad de secuencia del 70 % al 90 % o más pueden considerarse ser sustancialmente idénticas. Particularmente, la identidad/similitud descrita existe en un tramo de al menos aproximadamente 10 nucleótidos, preferentemente sobre un tramo de al menos aproximadamente 20 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 30 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 40 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 50 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 60 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 70 nucleótidos y lo más preferentemente en un tramo de 72 nucleótidos. La similitud entre la secuencia de la lectura genética y la secuencia genética de referencia también puede existir en tramos más largos, por ejemplo, en un tramo de al menos aproximadamente 80 nucleótidos, en un tramo de al menos aproximadamente 100 nucleótidos, o en toda la longitud de la lectura genética. En consecuencia, una lectura genética puede estar alineada con la secuencia genética de referencia si la lectura genética es idéntica/similar en un tramo de al menos aproximadamente 5 nucleótidos, preferentemente de al menos aproximadamente 10 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 15 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 20 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 30 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 40 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 50 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 60 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 70 nucleótidos y lo más preferentemente en un tramo de 72 nucleótidos. Además, una lectura genética puede estar alineada con la secuencia genética de referencia si al menos aproximadamente el 50 %, preferentemente al menos aproximadamente el 60 %, más preferentemente al menos aproximadamente el 70 %, más preferentemente al menos aproximadamente el 80 %, más preferentemente al menos aproximadamente el 90 %, más preferentemente al menos aproximadamente el 95 %, más preferentemente al menos aproximadamente el 99 % o lo más preferentemente al menos aproximadamente el 100 % de los nucleótidos son idénticos entre la lectura genética y la secuencia genética de referencia. Las lecturas genéticas que no se asignan a la secuencia genética de referencia de acuerdo con los criterios definidos anteriormente se descartan.

La alineación puede realizarse manualmente, aunque normalmente se implementa mediante un algoritmo informático, ya que sería imposible alinear las lecturas en un período de tiempo razonable para implementar los métodos desvelados en el presente documento. Por ejemplo, podría usarse el programa informático Efficient Local Alignment of Nucleotide Data (ELAND) distribuido como parte del proceso de análisis genómico de Illumina o el alineador

Burrows-Wheeler (BWA) (véase, por ejemplo, http://bio-bwa.source-forge.net/ bwa.shtml; o Li H. y Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60). Alternativamente, puede emplearse un filtro Bloom o un comprobador de pertenencia de un conjunto similar para alinear las lecturas con los genomas de referencia. Véase la Solicitud de Patente de EE.UU. N.º 61/552.374 enviada el 27 de octubre de 2011. En aspectos particulares de la invención, la pluralidad de lecturas genéticas se alinean mediante el alineador de Burrows-Wheeler (BWA).

Como se usa en el presente documento, la expresión "alinear la pluralidad de lecturas genéticas con al menos una secuencia genética de referencia" significa que las lecturas genéticas, las cuales se determinan, por ejemplo, por la secuenciación de ácidos nucleicos, están alineados con la al menos una secuencia genética de referencia. En consecuencia, al menos una lectura genética está alineada con la al menos una secuencia genética de referencia.

Como se usa en el presente documento, la expresión "al menos una secuencia genética de referencia" se refiere a al menos una secuencia de ácido nucleico (por ejemplo, al menos un genoma) de un vertebrado o humano. Si se quiere decir más de una secuencia de ácido nucleico (genomas), la etapa de alineación se puede realizar de manera que primero las lecturas genéticas se alineen con una secuencia genética de referencia y en la siguiente etapa las lecturas genéticas se alineen con una secuencia genética de referencia adicional, es decir, la alineación se realiza una tras otra. Alternativamente, se crea un consenso de más de una secuencia genética de referencia y, posteriormente, las lecturas genéticas se alinean con la secuencia consenso de más de una secuencia genética de referencia. La expresión "al menos una secuencia genética de referencia" puede referirse a una, dos, tres, cuatro, cinco, seis, siete, ocho, nueve, diez, quince, veinte o al menos treinta secuencia o secuencias genéticas de referencia.
Además, la expresión "al menos una secuencia genética de referencia" también podría referirse a al menos una secuencia de ácido nucleico de al menos un vertebrado, por ejemplo, más de un individuo de una especie o uno o más individuo o individuos de más de una especie. La al menos una secuencia genética de referencia se refiere particularmente a al menos una secuencia de ácido nucleico de ser humano o animal, particularmente un mamífero. Por lo tanto, los métodos proporcionados en el presente documento son aplicables a ácidos nucleicos de humanos y animales. En consecuencia, la al menos una secuencia genética de referencia puede ser al menos una secuencia de ácido nucleico de un animal, tales como un ratón, rata, hámster, conejo, cobaya, hurón, gato, perro, pollo, oveja, especie bovina, caballo, camello o primate. Particularmente, la al menos una secuencia genética de referencia se refiere a una secuencia de ácido nucleico (genoma) de un individuo. Además, la al menos una secuencia genética de referencia se refiere a las secuencias de ácido nucleico (genomas) de al menos un individuo humano. Particularmente, la al menos una secuencia genética de referencia se refiere a la secuencia de ácido nucleico (genoma) de un individuo o sujeto humano. La al menos una secuencia genética de referencia puede referirse a una secuencia de ácido nucleico (genoma) de un sujeto humano sano o puede referirse a (una) secuencia o secuencias de ácido nucleico (genoma o genomas) de al menos un sujeto humano sano, es decir, una cohorte de sujetos sanos. Un sujeto sano no tiene ninguna enfermedad y/o trastorno diagnosticado por un médico. En determinados aspectos, la al menos una secuencia genética de referencia puede referirse a (una) secuencia o secuencias de ácido nucleico (genoma o genomas) de al menos un individuo humano que padece una enfermedad y/o trastorno. La enfermedad y/o trastorno puede ser, por ejemplo, un trastorno genético. Dicho trastorno puede estar provocado por una o más anomalías en el genoma, particularmente una afección que está presente desde el nacimiento (congénita). Los trastornos genéticos pueden ser hereditarios, transmitidos de los genes de los padres. En consecuencia, la al menos una secuencia genética de referencia puede referirse a al menos una secuencia de ácido nucleico de uno o más sujetos que padecen una enfermedad y/o trastorno que se selecciona del grupo que consiste en un trastorno de un solo gen, trastorno autosómico dominante, trastorno autosómico recesivo, trastorno dominante ligado al cromosoma X, trastorno recesivo ligado al cromosoma X, trastorno ligado al cromosoma Y, enfermedad mitocondrial y un trastorno genético asociado a múltiples genes. La al menos una secuencia genética de referencia puede referirse a (una) secuencia o secuencias de ácido nucleico (genoma o genomas) de uno o más sujetos que padecen cáncer y/o enfermedad o enfermedades tumorales.

En determinados aspectos, la etapa de alinear la pluralidad de lecturas genéticas con al menos una secuencia genética de referencia significa alinear la pluralidad de lecturas genéticas con una única secuencia genética de referencia.

Como se usa en el presente documento, la expresión "agrupar las lecturas genéticas que comparten una posición genética en una secuencia genética de referencia de la al menos una secuencia genética de referencia en una pluralidad de grupos" significa que las lecturas genéticas que se alinean con una posición genética se agrupan en una pluralidad de grupos. Como se usa en el presente documento, la expresión "posición genética" puede referirse a un 1 nucleótido, preferentemente al menos 1 nucleótido, más preferentemente al menos aproximadamente 2 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 3 nucleótidos, sobre un tramo de al menos aproximadamente 4 nucleótidos, un tramo de al menos aproximadamente 5 nucleótidos, preferentemente sobre un tramo de al menos aproximadamente 10 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 15 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 20 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 30 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 40 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 50 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 60 nucleótidos, más preferentemente en un tramo de al menos aproximadamente 70 nucleótidos y lo más preferentemente en un tramo de

72 nucleótidos. Por lo tanto, las lecturas genéticas que comparten la misma ubicación genética se agrupan en una pluralidad de grupos.

Como se usa en el presente documento, una pluralidad de grupos puede referirse a al menos 2, 3, 5, 10, 50, 100, 1000, 100000, $1x10^6$, $1x10^9$ grupos.

Como se usa en el presente documento, el tamaño del grupo también podría ser de al menos 4, al menos 5, al menos 6, al menos 7, al menos 8, al menos 9, al menos 10, al menos 20, al menos 30, 100, al menos 1000 lecturas genéticas. Particularmente, el tamaño del grupo puede ser de al menos tres lecturas genéticas. En consecuencia, un grupo correspondiente puede comprender al menos 3 lecturas genéticas.

Como se usa en el presente documento, el término "compartir" significa que la lectura genética es de al menos aproximadamente el 50 %, preferentemente al menos aproximadamente el 60 %, más preferentemente al menos aproximadamente el 70 %, más preferentemente al menos aproximadamente el 80 %, más preferentemente al menos aproximadamente el 90 %, más preferentemente al menos aproximadamente el 95 %, más preferentemente al menos aproximadamente el 99 % o lo más preferentemente al menos aproximadamente el 100 % idéntica a la secuencia genética de referencia. La posición genética puede identificarse mediante la alineación de las lecturas genéticas entre sí y/o con la secuencia genética de referencia. En consecuencia, incluso si las lecturas genéticas comprenden (un) nucleótido o nucleótidos erróneos que pueden incorporarse mediante la amplificación de la muestra de ácido nucleico, la agrupación clasifica las lecturas genéticas de modo que las lecturas genéticas deriven de la misma posición genética. En consecuencia, en la agrupación, las lecturas genéticas que derivan de la misma posición genética se clasifican en grupos/familias. En consecuencia, los grupos respectivos comprenden lecturas genéticas que son amplicones (productos de amplificación) de la misma posición/región genética correspondiente.

Las lecturas genéticas se agrupan compartiendo una posición genética en una secuencia genética de referencia de la al menos una secuencia genética de referencia. En consecuencia, la agrupación puede basarse en una secuencia genética de referencia que se utiliza en la etapa de alineación. La agrupación también puede basarse en secuencias genéticas de referencia adicionales.

En aspectos particulares de la invención, la etapa de agrupación agrupa las lecturas genéticas que comparten una posición genética en una secuencia genética de referencia de la al menos una secuencia genética de referencia en una pluralidad de grupos y en donde cada lectura genética en un grupo correspondiente de la pluralidad de grupos comprende al menos una secuencia de ácido nucleico particular adicional, particularmente al menos una secuencia de código de barras.

En particular, si hay información adicional sobre la secuencia, por ejemplo, secuencias artificiales, tales como (una) molécula o moléculas adaptadoras o código o códigos de barras, están unidos a las moléculas de ácido nucleico, la alineación con la al menos una secuencia genética de referencia puede no considerar la información de secuencia adicional. En aspectos particulares, la molécula o moléculas adaptadoras o el código o códigos de barras no están alineados con la secuencia o secuencias de referencia. En una etapa tal, la información de secuencia adicional puede cortarse/retirarse.

En consecuencia, la agrupación de las lecturas genéticas puede basarse en (una) característica o características particulares, por ejemplo, al menos una secuencia de ácido nucleico (particular), en la secuencia de las lecturas genéticas. Para agrupar las lecturas genéticas, las lecturas genéticas pueden alinearse con tales (una) característica o características particulares, por ejemplo, (una) secuencia o secuencias de ácido nucleico (particulares). Tales (una) característica o características particulares pueden predeterminarse y/o pueden seleccionarse y las lecturas genéticas pueden entonces alinearse con tales (una) característica o características. Un grupo (correspondiente) de los grupos construidos puede entonces comprender o consistir en lecturas genéticas que comparten o comprenden la característica o características particulares, por ejemplo, al menos una secuencia de ácido nucleico particular. Por lo tanto, las lecturas genéticas de este grupo comprenden la característica o características particulares idénticas, por ejemplo, secuencias de ácido nucleico particulares idénticas.

La secuencia de ácido nucleico particular como se usa en el presente documento puede ser un tramo corto de nucleótidos, una secuencia tal de código de barras o identificador de molécula única.

En consecuencia, en la etapa de agrupación, las lecturas genéticas pueden alinearse con la al menos una secuencia genética de referencia y además se pueden alinear con al menos una secuencia de ácido nucleico particular. Una secuencia tal de ácido nucleico particular puede no ser de origen natural en el genoma del individuo cuya secuencia de ácido nucleico se analiza (característica que no es de origen natural). Por ejemplo, una característica tal puede ser una secuencia de ácido nucleico que está unida a la molécula de ácido nucleico comprendida en la muestra de ácido nucleico. En consecuencia, la al menos una secuencia de ácido nucleico particular podría unirse a la molécula de ácido nucleico y en donde la molécula o moléculas de ácido nucleico comprendidas en la muestra de ácido nucleico se analizan mediante secuenciación de ácido nucleico y por lo tanto se determina la pluralidad de lecturas genéticas.

Por ejemplo, la al menos una secuencia de ácido nucleico particular puede ser información genética que está comprendida en el adaptador o adaptadores unidos a la molécula de ácido nucleico, por ejemplo, una o más secuencia o secuencias de códigos de barras que están incluidas en el adaptador adaptadores, como se ha descrito anteriormente.

En aspectos particulares de la invención, el método proporcionado en el presente documento puede comprender la agrupación y la unidad de cálculo puede configurarse para agrupar las lecturas genéticas que comparten una posición genética en una secuencia genética de referencia de la al menos una secuencia genética de referencia en una pluralidad de grupos alineando la al menos una secuencia de código de barras, por ejemplo, de los amplicones etiquetados, entre sí. La agrupación basada en las secuencias de códigos de barras proporciona "familias" o "grupos" que comparten la misma posición genómica y la secuencia o secuencias de códigos de barras idénticas. En consecuencia, en aspectos particulares, la agrupación de las lecturas genéticas es una alineación con la secuencia o secuencias genéticas de referencia y con la secuencia o secuencias de código de barras. Por ejemplo, la agrupación puede comprender la alineación de al menos una secuencia de código de barras particular. La agrupación puede agrupar las lecturas genéticas de modo que las lecturas genéticas de un grupo comprendan o compartan una secuencia de código de barras idéntica particular, o que las lecturas genéticas de un grupo comprendan o compartan al menos una secuencia de código de barras particular en un grupo, por ejemplo, dos secuencias de códigos de barras individuales. En otras palabras, las lecturas genéticas de un grupo comprenden o comparten las secuencias de códigos de barras idénticas. Por lo tanto, las lecturas genéticas de un grupo pueden compartir la al menos una secuencia de código de barras. Particularmente, las lecturas genéticas de un grupo comparten dos secuencias de códigos de barras. Una agrupación tal se ilustra a continuación en los ejemplos adjuntos.

En aspectos adicionales de la invención, la característica o características particulares empleadas en el agrupamiento de las lecturas genéticas pueden estar comprendidas en el genoma del que deriva la lectura genética. En consecuencia, la característica o características particulares empleadas en la agrupación de las lecturas genéticas también pueden ser al menos una característica de origen natural, por ejemplo, (una) secuencia o secuencias de ácido nucleico (particulares) que se producen en el genoma del individuo cuya secuencia de ácido nucleico se analiza. Por ejemplo, la al menos una secuencia de ácido nucleico particular corresponde al extremo o extremos de las moléculas de ácido nucleico que están comprendidas en la muestra de ácido nucleico. Por ejemplo, los extremos cortados aleatoriamente (por ejemplo, los extremos 5' y 3') de las moléculas de ácido nucleico pueden emplearse como secuencias de ácido nucleico particulares, por ejemplo, como secuencias de códigos de barras.

Además, la característica o características particulares empleadas en el agrupamiento de las lecturas genéticas pueden ser una combinación de al menos una característica de origen no natural y al menos una característica de origen natural.

Las lecturas genéticas dentro del grupo podrían tener una longitud diferente. En aspectos preferidos particulares, las lecturas genéticas del grupo tienen la misma longitud.

La invención puede referirse a un método para la secuenciación de ácidos nucleicos que comprende las siguientes etapas:

(a) opcionalmente, obtener una pluralidad de lecturas genéticas mediante la secuenciación de una muestra de ácido nucleico;
(b) alinear la pluralidad de lecturas genéticas con al menos una secuencia genética de referencia;
(c) agrupar las lecturas genéticas que comparten una posición genética en una secuencia genética de referencia de la al menos una secuencia genética de referencia en una pluralidad de grupos, en donde cada lectura genética en un grupo correspondiente de la pluralidad de grupos comprende al menos una secuencia de código de barras;
(d) crear una secuencia consenso para cada grupo de la pluralidad de grupos, en donde se crea una secuencia consenso correspondiente determinando el nucleótido más abundante *man_p* en cada posición específica *p* de una pluralidad de posiciones dentro del grupo correspondiente de lecturas genéticas y

(i) establecer una representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación r entre el número de lecturas genéticas dentro del grupo correspondiente que tiene el nucleótido más abundante *man_p* en la posición específica *p* y el número de lecturas genéticas dentro del un grupo correspondiente es superior o igual a un umbral predeterminado *t;* y
(ii) establecer una etiqueta *N* si la relación *r* está por debajo del umbral predeterminado *t;*

(e) comparar las secuencias consenso de la pluralidad de grupos con la secuencia genética de referencia en cada posición específica *p* de una pluralidad de posiciones de las secuencias consenso y en donde una diferencia en una posición específica entre las secuencias consenso y la secuencia genética de referencia indica una variación en la posición específica;
(f) determinar el número de secuencias consenso que comprenden la variación en cada posición específica *p* de una pluralidad de posiciones y determinar el número de secuencias consenso que comprenden la etiqueta *N* en cada posición específica *p* de una pluralidad de posiciones; y

(g) identificar la variación en cada posición específica *p* de una pluralidad de posiciones como una verdadera variación si una relación *r** entre el número de secuencias consenso que comprenden la etiqueta *N* en la posición específica *p* y el número de secuencias consenso que comprenden la variación en la posición específica *p* está por debajo de un umbral *t**.

Además, la invención puede referirse a un método para la secuenciación de ácidos nucleicos que comprende las siguientes etapas:

(a) obtener una pluralidad de lecturas genéticas mediante la secuenciación de una muestra de ácido nucleico que comprende:

(i) unir una pluralidad de adaptadores a moléculas de ácido nucleico comprendidas en la muestra para generar moléculas de ácido nucleico etiquetadas, en donde la pluralidad de adaptadores comprende secuencias de códigos de barras,
(ii) amplificar las moléculas de ácido nucleico etiquetadas para producir amplicones etiquetados; y
(iii) determinar la secuencia de ácido nucleico de los amplicones etiquetados mediante secuenciación de ácido nucleico y determinar de esta manera una pluralidad de lecturas genéticas; y

(b) alinear la pluralidad de lecturas genéticas con al menos una secuencia genética de referencia;
(c) agrupar las lecturas genéticas que comparten una posición genética en una secuencia genética de referencia de la al menos una secuencia genética de referencia en una pluralidad de grupos, en donde cada lectura genética en un grupo correspondiente de la pluralidad de grupos comprende al menos una secuencia de código de barras;
(d) crear una secuencia consenso para cada grupo de la pluralidad de grupos, en donde se crea una secuencia consenso correspondiente determinando el nucleótido más abundante *man_p* en cada posición específica *p* de una pluralidad de posiciones dentro del grupo correspondiente de lecturas genéticas y

(i) establecer una representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación r entre el número de lecturas genéticas dentro del grupo correspondiente que tiene el nucleótido más abundante *man_p* en la posición específica *p* y el número de lecturas genéticas dentro del un grupo correspondiente es superior o igual a un umbral predeterminado *t;* y
(ii) establecer una etiqueta *N* si la relación r está por debajo del umbral predeterminado *t;*

(e) comparar las secuencias consenso de la pluralidad de grupos con la secuencia genética de referencia en cada posición específica *p* de una pluralidad de posiciones de las secuencias consenso y en donde una diferencia en una posición específica entre las secuencias consenso y la secuencia genética de referencia indica una variación en la posición específica;
(f) determinar el número de secuencias consenso que comprenden la variación en cada posición específica *p* de una pluralidad de posiciones y determinar el número de secuencias consenso que comprenden la etiqueta *N* en cada posición específica *p* de una pluralidad de posiciones; y
(g) identificar la variación en cada posición específica *p* de una pluralidad de posiciones como una verdadera variación si una relación *r** entre el número de secuencias consenso que comprenden la etiqueta *N* en la posición específica *p* y el número de secuencias consenso que comprenden la variación en la posición específica *p* está por debajo de un umbral *t**.

Además, el orden de las etapas podría cambiarse siempre que el método sea adecuado para identificar una verdadera variante genética. Por ejemplo, la invención puede referirse a un método para la secuenciación de ácidos nucleicos que comprende las siguientes etapas:

(a) opcionalmente, obtener una pluralidad de lecturas genéticas mediante la secuenciación de una muestra de ácido nucleico;
(b) agrupar una pluralidad de lecturas genéticas en una pluralidad de grupos, en donde las lecturas genéticas comprendidas en un grupo correspondiente comparten al menos una secuencia de código de barras particular,
(c) crear una secuencia consenso para cada grupo de la pluralidad de grupos, en donde se crea una secuencia consenso correspondiente determinando el nucleótido más abundante *man_p* en cada posición específica *p* de una pluralidad de posiciones dentro del grupo correspondiente de lecturas genéticas y

(i) establecer una representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación r entre el número de lecturas genéticas dentro del grupo correspondiente que tiene el nucleótido más abundante *man_p* en la posición específica *p* y el número de lecturas genéticas dentro del un grupo correspondiente es superior o igual a un umbral predeterminado *t;* y
(ii) establecer una etiqueta *N* si la relación r está por debajo del umbral predeterminado;

(d) alinear las secuencias consenso de la pluralidad de grupos con al menos una secuencia genética de referencia;
(e) comparar las secuencias consenso de la pluralidad de grupos con la secuencia genética de referencia en cada

posición específica *p* de una pluralidad de posiciones de las secuencias consenso y en donde una diferencia en una posición específica entre las secuencias consenso y la secuencia genética de referencia indica una variación en la posición específica;
(f) determinar el número de secuencias consenso que comprenden la variación en cada posición específica *p* de una pluralidad de posiciones y determinar el número de secuencias consenso que comprenden la etiqueta *N* en cada posición específica *p* de una pluralidad de posiciones; y
(g) identificar la variación en cada posición específica *p* de una pluralidad de posiciones como una verdadera variación si una relación *r** entre el número de secuencias consenso que comprenden la etiqueta *N* en la posición específica *p* y el número de secuencias consenso que comprenden la variación en la posición específica *p* está por encima de un umbral *t**.

Como se usa en el presente documento, el "grupo correspondiente", o el "un grupo correspondiente" se refiere a un grupo entre la pluralidad de grupos que comprende lecturas genéticas que comparten la característica o características particulares, es decir, las lecturas de un grupo tal comprenden la característica o características idénticas.

Como se indica anteriormente, las lecturas genéticas se agrupan preferentemente que comparten una posición genética en una secuencia genética de referencia de la al menos una secuencia genética de referencia en una pluralidad de grupos, en donde cada lectura genética en un grupo correspondiente de la pluralidad de grupos comprende al menos una secuencia de ácido nucleico particular adicional, tal como al menos una secuencia de código de barras. Esto significa que las lecturas genéticas se clasifican en una pluralidad de grupos de acuerdo con la posición genética (como se describe anteriormente) y la secuencia o secuencias de ácido nucleico particulares adicionales que están comprendidas en la secuencia de ácido nucleico de las lecturas genéticas. Por lo tanto, las lecturas genéticas del un grupo correspondiente entre la pluralidad de grupos comprenden o comparten la misma posición genética y la secuencia o secuencias nucleicas particulares adicionales idénticas, por ejemplo, la al menos una secuencia de código de barras. En consecuencia, las lecturas genéticas en el un grupo correspondiente pueden clasificarse de acuerdo con su secuencia genética y al menos una secuencia de ácido nucleico predeterminada y/o seleccionada, o al menos una secuencia de código de barras. En consecuencia, en aspectos particulares, las lecturas genéticas en el agrupamiento se agrupan en función de su posición genética y su secuencia de código de barras.

Como se indica anteriormente, se forma una pluralidad de grupos. La pluralidad de grupos puede comprender un grupo' de ejemplo y un grupo" de ejemplo adicional. La pluralidad de grupos puede comprender grupos adicionales, por ejemplo, del grupo"' al grupo^{n}. Las lecturas genéticas del grupo" pueden superponerse, al menos parcialmente, con las lecturas genéticas del grupo'. Las lecturas genéticas del grupo" pueden no superponerse con las lecturas genéticas del grupo'. Las lecturas genéticas del grupo" pueden superponerse completamente con las lecturas genéticas del grupo'.

Como se usa en el presente documento, la expresión "superposición parcial" puede significar que las lecturas genéticas de un grupo correspondiente (grupo') son idénticas a otro grupo (por ejemplo, grupo"), por ejemplo, al menos aproximadamente el 50 %, al menos aproximadamente el 60 %, al menos aproximadamente el 70 %, al menos aproximadamente el 80 %, al menos aproximadamente el 90 %, al menos aproximadamente el 95 %, al menos aproximadamente el 99 % o al menos aproximadamente el 100 % idénticas a la secuencia genética de referencia. La identidad puede producirse, por ejemplo, en 1 nucleótido, al menos 1 nucleótido, al menos aproximadamente 2 nucleótidos, en un tramo de al menos aproximadamente 3 nucleótidos, al menos aproximadamente 4 nucleótidos, al menos aproximadamente 5 nucleótidos, al menos aproximadamente 10 nucleótidos, al menos aproximadamente 15 nucleótidos, al menos aproximadamente 20 nucleótidos, al menos aproximadamente 30 nucleótidos, al menos aproximadamente 40 nucleótidos, al menos aproximadamente 50 nucleótidos, al menos aproximadamente 60 nucleótidos, al menos aproximadamente 70 nucleótidos o al menos aproximadamente o igual a 72 nucleótidos.

En consecuencia, el agrupamiento agrupa lecturas genéticas que comparten una posición genética en una secuencia genética de referencia de la al menos una secuencia genética de referencia en una pluralidad de grupos. Por ejemplo, cada lectura genética en un grupo correspondiente de la pluralidad de grupos comprende al menos una secuencia de código de barras. Por lo tanto, las lecturas genéticas comprendidas en la pluralidad de grupos comparten una posición genética y cada grupo de la pluralidad de grupos comparte al menos una secuencia de código de barras.

La agrupación también puede realizarse para una segunda pluralidad de grupos que comparten una segunda posición genética en la secuencia genética de referencia de la al menos una secuencia genética de referencia. En consecuencia, los métodos de la invención pueden comprender la etapa y el sistema de la invención puede configurarse para agrupar/hacer grupos de las lecturas genéticas que comparten una segunda posición genética en la secuencia genética de referencia de la al menos una secuencia genética de referencia en una segunda pluralidad de grupos. Cada lectura genética en un grupo correspondiente de la segunda pluralidad de grupos puede comprender particularmente al menos una secuencia de código de barras.
La agrupación también puede realizarse para al menos una pluralidad adicional de grupos que comparten al menos una posición genética adicional en la secuencia genética de referencia de la al menos una secuencia genética de referencia. En consecuencia, los métodos de la invención pueden comprender la etapa y el sistema de la invención puede configurarse para agrupar/hacer grupos de las lecturas genéticas que comparten al menos una posición

genética adicional en la secuencia genética de referencia de la al menos una secuencia genética de referencia en al menos una pluralidad adicional de grupos. Cada lectura genética en un grupo correspondiente de la al menos una pluralidad adicional de grupos puede comprender particularmente al menos una secuencia de código de barras.

Como se indica en los ejemplos, la cadena complementaria puede comprender el complemento inverso de la secuencia del código de barras. La cadena complementaria puede identificarse mediante el marcador/la etiqueta recíproco. Por lo tanto, las lecturas genéticas del grupo' pueden corresponder a una primera cadena de un ácido nucleico bicatenario y las lecturas genéticas del grupo" o grupo adicional pueden corresponder a la segunda cadena complementaria de dicho ácido nucleico bicatenario.

El método proporcionado en el presente documento comprende además la creación y la unidad de cálculo puede configurarse para crear una secuencia consenso para cada grupo de la pluralidad de grupos, en donde se crea una secuencia consenso correspondiente determinando el nucleótido más abundante *man_p* en cada posición específica *p* de una pluralidad de posiciones dentro del grupo correspondiente de lecturas genéticas. En consecuencia, la secuencia en cada posición específica *p* de una pluralidad de posiciones de las lecturas genéticas en el grupo o cada grupo (o un subconjunto del mismo) se colapsa en una secuencia consenso respectiva.

Como se usa en el presente documento, la expresión "cada posición específica *p* de una pluralidad de posiciones" significa que cada posición es consultada de una pluralidad de posiciones paso a paso. Por ejemplo, cada posición de las lecturas genéticas dentro de un grupo correspondiente se consulta base por base para una pluralidad de posiciones.

Como se usa en el presente documento, la expresión "pluralidad de posiciones" significa al menos aproximadamente el 50 %, preferentemente al menos aproximadamente el 60 %, más preferentemente al menos aproximadamente el 70 %, más preferentemente al menos aproximadamente el 80 %, más preferentemente al menos aproximadamente el 90 %, más preferentemente al menos aproximadamente el 99 % o más preferiblemente el 100 % de las posiciones de las lecturas genéticas o las secuencias consenso se consultan.

La pluralidad de posiciones dentro del grupo correspondiente de lecturas genéticas significa que la creación de la secuencia consenso se realiza durante al menos aproximadamente el 50 %, preferentemente al menos aproximadamente el 60 %, más preferentemente al menos aproximadamente el 70 %, más preferentemente al menos aproximadamente el 80 %, más preferentemente al menos aproximadamente el 90 %, más preferentemente al menos aproximadamente el 99 % de las posiciones de las lecturas genéticas dentro del grupo. Particularmente, la creación de la secuencia consenso se realiza para todas las posiciones respectivas de las lecturas genéticas dentro del grupo correspondiente.

Como se usa en el presente documento, la pluralidad de grupos puede referirse a al menos aproximadamente el 50 %, preferentemente al menos aproximadamente el 60 %, más preferentemente al menos aproximadamente el 70 %, más preferentemente al menos aproximadamente el 80 %, más preferentemente al menos aproximadamente el 90 %, más preferentemente al menos aproximadamente el 99 % o lo más preferentemente el 100 % de los grupos, por ejemplo, que están formados por la agrupación de las lecturas genéticas. Lo más preferentemente, se considera que todas las lecturas genéticas forman la secuencia consenso dentro de un grupo.

Como se usa en el presente documento, cada grupo de la pluralidad de grupos significa que se consulta cada grupo de la pluralidad de grupos, por ejemplo, comparando, determinando o identificando, paso a paso. Por ejemplo, la secuencia consenso se crea para cada grupo que se forma en la etapa de agrupación.

En determinados aspectos, el método proporcionado en el presente documento puede comprender además la creación y la unidad de cálculo puede configurarse para crear una secuencia consenso para cada grupo de una pluralidad de grupos, en donde la pluralidad de grupos puede referirse a al menos aproximadamente el 50 %, preferentemente al menos aproximadamente el 60 %, más preferentemente al menos aproximadamente el 70 %, más preferentemente al menos aproximadamente el 80 %, más preferentemente al menos aproximadamente el 90 %, más preferentemente al menos aproximadamente el 99 % o lo más preferentemente el 100 % de los grupos que están formados por la agrupación de las lecturas genéticas que comparten una posición genética en una secuencia genética de referencia.

En aspectos particulares, las secuencias consenso se crean para cada grupo de la pluralidad de grupos, en donde la pluralidad de grupos corresponde al 100 % de los grupos que se forman por la agrupación de las lecturas genéticas.

En aspectos particulares, el nucleótido más abundante *man_p* se determina en cada posición específica *p* de una pluralidad de posiciones y la representación del nucleótido más abundante *man_p* o la etiqueta N se establece en cada posición específica *p* de una pluralidad de posiciones, en donde la pluralidad de posiciones se refiere a todas las posiciones respectivas de las lecturas genéticas dentro del grupo correspondiente.

En los métodos y sistemas de la invención, la secuencia consenso para cada grupo de la pluralidad de grupos puede

crearse determinando el nucleótido más abundante *man_p* en cada puesto específico *p* de una pluralidad de posiciones dentro del grupo correspondiente de lecturas genéticas y estableciendo una representación del nucleótido man_p más abundante o una etiqueta N basada en un umbral t.

Como se usa en el presente documento, la "secuencia consenso" es el orden calculado de los nucleótidos más frecuentes que se encuentran en las posiciones correspondientes en las lecturas genéticas que componen el grupo, en donde la secuencia consenso comprende una representación del nucleótido (man_p) o una etiqueta N como se define a continuación.

El "nucleótido más frecuente" o "nucleótido más abundante" en una posición específica p se denomina en el presente documento *"man_p".* La (representación del) nucleótido más abundante en la posición específica *p* se refiere a cualquiera de adenina ("A"), guanina ("G"), citosina ("C"), timina ("T") o uracilo ("U") o cualquier otra nucleobase presente en la secuencia de ácido nucleico de las lecturas genéticas. Por lo tanto, como se usa en el presente documento, la "representación del nucleótido más abundante *man_p"* puede referirse a la nucleobase de origen predominante adenina ("A"), guanina ("G"), citosina ("C"), timina ("T") o uracilo ("U") en las lecturas genéticas dentro de un grupo correspondiente.

Las secuencias consenso pueden crearse comparando las secuencias de las lecturas genéticas entre sí, que están alineados en los grupos, por ejemplo, las lecturas genéticas comparten la posición genética. La secuencia de ácido nucleico de las lecturas genéticas de un grupo puede puntuarse nucleótido por nucleótido. Pueden puntuarse los nucleótidos y puede determinarse que el nucleótido en una posición específica p es el "nucleótido más abundante" o *"man_p"* en las lecturas genéticas dentro de un grupo correspondiente. Por ejemplo, se determina que el nucleótido en una posición específica p es el nucleótido más abundante y se establece en la secuencia consenso que se produce en esta posición particular p dentro de las lecturas genéticas de este grupo particular con un umbral particular (por ejemplo, denominado t), por ejemplo, al menos aproximadamente el 76 %.

La representación del nucleótido más abundante *man_p* se establece en la secuencia consenso si una relación r entre el número de lecturas genéticas dentro del grupo correspondiente que tiene el nucleótido más abundante *man_p* en la posición específica *p* y el número de lecturas genéticas dentro de la correspondiente es superior o igual a un umbral predeterminado *t.* Si la relación *r* está por debajo del umbral predeterminado *t,* el nucleótido en la posición específica se establece *"N".* Por lo tanto, el radio *r* determina si se establece la representación del nucleótido más abundante o si se establece la etiqueta "N" en la secuencia consenso. La relación r se calcula entre el número de lecturas genéticas dentro del grupo correspondiente que tiene el nucleótido más abundante *man_p* en la posición específica *p* y el número de lecturas genéticas dentro del grupo correspondiente.

En consecuencia, la expresión de los métodos y sistemas proporcionados en el presente documento de "crear una secuencia consenso para cada grupo de la pluralidad de grupos" puede referirse a la creación de una secuencia consenso para cada grupo de la pluralidad de grupos, en donde una secuencia consenso correspondiente se crea estableciendo una representación del nucleótido más abundante *man_p* en cada posición específica *p* de una pluralidad de posiciones de la secuencia consenso si una relación r entre el número de lecturas genéticas dentro del grupo correspondiente que tiene el nucleótido más abundante *man_p* en la posición específica *p* y el número de lecturas genéticas dentro del un grupo correspondiente es superior o igual a un umbral predeterminado *t;* y establecer una etiqueta *N* si la relación *r* está por debajo del umbral predeterminado *t*.

En consecuencia en los métodos y sistemas proporcionados en el presente documento, la secuencia consenso en cada posición específica *p* de una pluralidad de posiciones dentro del grupo de lecturas genéticas, en donde se establece una representación del nucleótido más abundante en la secuencia consenso si este nucleótido se produce en esta posición específica con una aparición particular, en donde la aparición es el umbral t, y en donde se establece una etiqueta N en la secuencia consenso si este nucleótido no aparece en esta posición específica con la aparición particular.

Como se usa en el presente documento, la expresión "número de lecturas genéticas" se refiere al número de lecturas genéticas que están comprendidas en un grupo correspondiente de la pluralidad de grupos. Por ejemplo, este término puede referirse al número de lecturas genéticas dentro de un grupo de la pluralidad de grupos, en donde en este grupo las lecturas genéticas comparten una posición genética y opcionalmente comparten al menos una secuencia de código de barras.

Como se usa en el presente documento, la expresión el "número de lecturas genéticas dentro del grupo correspondiente que tiene el nucleótido más abundante *man_p* en la posición específica *p"* se refiere al número de las lecturas genéticas, que están comprendidas en un grupo correspondiente de la pluralidad de grupos, y que tienen el nucleótido más abundante *man_p* en una posición particular *p* de la pluralidad de posiciones. Por lo tanto, se refiere al número de lecturas genéticas dentro del grupo que tienen en esta posición particular un nucleótido idéntico más abundante.

La invención comprende en los métodos y sistemas crear una secuencia consenso determinando el nucleótido más abundante *man_p* en cada posición específica *p* de una pluralidad de posiciones dentro del grupo correspondiente de lecturas genéticas y

(i) establecer una representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación r entre el número de lecturas genéticas dentro del grupo correspondiente que tiene el nucleótido más abundante *man_p* en la posición específica *p* y el número de lecturas genéticas dentro del un grupo correspondiente es superior o igual a un umbral predeterminado t; y

(ii) establecer una etiqueta *N* si la relación r entre el número de lecturas genéticas dentro del un grupo correspondiente que tiene el nucleótido más abundante *man_p* en la posición específica *p* y el número de lecturas genéticas dentro del un grupo correspondiente está por debajo del umbral predeterminado *t.*

En otras palabras, los métodos y sistemas proporcionados en el presente documento crean una secuencia consenso en cada posición específica p de una pluralidad de posiciones para cada grupo de la pluralidad de grupos,
en donde una representación de un nucleótido en una posición específica de la pluralidad de posiciones se establece en una secuencia consenso correspondiente, si una relación entre el número de nucleótidos en la posición específica y el número de lecturas genéticas dentro del grupo correspondiente es superior o igual a un umbral predeterminado t y
en donde una etiqueta N se establece en la secuencia consenso correspondiente en la posición específica de la pluralidad de posiciones si la relación está por debajo del umbral predeterminado t.

En aspectos particulares, los métodos y sistemas proporcionados en el presente documento crean una secuencia consenso en cada posición específica p de una pluralidad de posiciones para cada grupo de la pluralidad de grupos, en donde el nucleótido o nucleótidos se determinan en cada una de las posiciones específicas de la pluralidad de posiciones y
en donde una representación de un nucleótido del nucleótido o nucleótidos en una posición específica de la pluralidad de posiciones se establece en una secuencia consenso correspondiente, si una relación entre el número de este nucleótido en la posición específica y el número de lecturas genéticas dentro del grupo correspondiente es superior o igual a un umbral predeterminado t y
en donde una etiqueta N se establece en la secuencia consenso correspondiente en esta posición específica si la relación está por debajo del umbral predeterminado t.
En consecuencia, para crear la secuencia consenso, no es necesario determinar el nucleótido man_p más abundante siempre que el nucleótido del cual se establece la representación en la secuencia consenso cumpla el umbral t, es decir, el valor particular de la relación r.

Como se usa en el presente documento, el umbral *"f"* o "t" se refiere a cualquier número que se use como límite. El umbral t puede identificarse analizando conjuntos de entrenamiento. El umbral en el presente documento se refiere a un valor en porcentaje. Por lo tanto, el valor determinado por la relación r puede multiplicarse por 100 para obtener un valor que corresponda al valor en %. Como se describe en los ejemplos adjuntos, las muestras establecidas, por ejemplo, líneas celulares, tales como las líneas celulares normales HapMap (por ejemplo, como se da en la Tabla 1), pueden usarse para determinar un valor umbral apropiado. Se conoce la secuencia del genoma de tales líneas celulares. Por lo tanto, puede determinarse el umbral t apropiado que dé como resultado el resultado más favorable. En particular, el siguiente procedimiento puede usarse para identificar un umbral t apropiado:

Con un error de secuenciación del 1 %, la probabilidad de acumular una base incorrecta en la misma posición es 0,01 (1 % de error) / 3 (por ejemplo, 3 posibles bases incorrectas si A es correcto, G, T y C pueden ser artefactos) para cada lectura individual. Véase la tabla ilustrativa a continuación para ver los diferentes tamaños de familias de ejemplo (por ejemplo, 3 - 6) y las respectivas probabilidades de errores de secuenciación recurrentes (anotados con una probabilidad individual de 0,0033).

lectura 1	**lectura 2**	**lectura 3**	**lectura 4**	**lectura 5**	**lectura 6**	**límite**	**límite**	**límite**	**probabilidad**	**probabilidad en 1 de 10000 lecturas**	**probabilidad en %**
0,0033	0,0033	0,0033				**SÍ**	**SÍ**	**SÍ**	3,5937E-08	0,00035937	0,035937
0,0033	0,0033	0,0033	0,99			**SÍ**	**SÍ**	**NO**	3,55776E-08	0,000355776	0,03557763
0,0033	0,0033	0,0033	0,99	0,99		**SÍ**	NO	NO	3,52219E-08	0,000352219	0,035221854
0,0033	0,0033	0,0033	0,99	0,99	0,99	NO	NO	NO	3,48696E-08	0,000348696	0,034869635
0,0033	0,0033	0,0033	0,0033	0,99		**SÍ**	**SÍ**	**SÍ**	1,17406E-10	1,17406E-06	0,000117406
0,0033	0,0033	0,0033	0,0033	0,99	0,99	**SÍ**	NO	NO	1,16232E-10	1,16232E-06	0,000116232

Como se indica en la tabla anterior, dependiendo del corte elegido, la probabilidad de que los artefactos se consideren verdaderas mutaciones puede variar.

Por lo tanto, si el umbral se define como el 76 %, al menos 4 bases incorrectas e idénticas podrían producirse en la posición particular con una probabilidad de (0,0033*0,0033*0,0033*0,0033) 1,2E-10). Los umbrales más bajos pueden aumentar la probabilidad respectiva en consecuencia. Como consecuencia, los umbrales más bajos pueden conducir a secuencias consenso que contienen errores de secuenciación y, por lo tanto, pueden conducir a llamadas de mutación artificiales. Este riesgo aumenta en un entorno en el cual se secuencia la misma región de interesados con una profundidad de secuenciación de varios miles. Por ejemplo, a una profundidad de secuenciación de 10 000X, un límite del 75 % o más conduciría a una probabilidad de 0,0033*0,0033*0,0033*0,99 = 3,55776E-08 * 10 000X = 0,000355776 o ~ 0,04 % de un consenso artificial en una familia de lectura con 4 miembros. Considerando las bases generales que podrían secuenciarse en el ensayo respectivo, esta probabilidad puede considerarse demasiado alta. Por el contrario, a menos que las tres lecturas de una familia de ejemplo de 3 contengan artefactos de secuenciación, un límite del 76 % conduce a una probabilidad de 0,0033*0,0033*0,0033*0,0033= 1,18592E-10 * 10 000X = 1,18592E-06 o ~ 0,0001 % de un consenso artificial.

La elección del umbral depende del nivel de confianza que el usuario desee tener para realizar la clasificación.

Puede ser ventajoso usar conjuntos más grandes de muestras calificadas para mejorar la utilidad de los valores umbral.

En aspectos particulares, el umbral t predeterminado es aproximadamente el 76 %.

El umbral *t* predeterminado puede ser aproximadamente el 99 %, preferentemente el umbral *t* predeterminado es aproximadamente el 95 %, más preferentemente el umbral *t* predeterminado es aproximadamente el 90 %, más preferentemente el umbral *t* predeterminado es aproximadamente el 85 %, más preferentemente el umbral *t* predeterminado es aproximadamente el 80 %, más preferentemente el umbral *t* predeterminado es aproximadamente el 79 %, más preferentemente el umbral *t* predeterminado es aproximadamente el 78 %, más preferentemente el umbral *t* predeterminado es aproximadamente el 77 % o lo más preferentemente el umbral t predeterminado es aproximadamente el 76 %.

Sin desear quedar ligados a teoría alguna, puede usarse un tamaño de grupo de ejemplo de entre 3 y 8 lecturas genéticas para obtener una proporción óptima de "consenso" y "costes de secuenciación". En consecuencia, el experto en la materia puede determinar cuántas lecturas dentro de un grupo, por ejemplo, de 3, 4, 5, 6, 7 u 8 lecturas genéticas, pueden ser necesarias para formar una secuencia consenso. Como se muestra en los ejemplos adjuntos, 4 de cada 5 lecturas genéticas mejoraron el análisis. El 76 % representa la situación en la que 4 de cada 5 lecturas genéticas dentro de un grupo comparten la misma secuencia de nucleótidos en una posición determinada. También pueden emplearse dos de cada tres, tres de cada cuatro lecturas genéticas.

El método y sistema proporcionados en el presente documento comprenden en la creación de la secuencia consenso establecer la representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación r entre el número de lecturas genéticas dentro del un grupo correspondiente que tiene el nucleótido más abundante *man_p* en la posición específica *p* y el número de lecturas genéticas dentro del un grupo correspondiente es superior o igual a aproximadamente el 55 %, y el establecimiento de una etiqueta *N* si la relación está por debajo de aproximadamente el 55 %.

Preferentemente, el método y el sistema comprenden en la creación de la secuencia consenso establecer la representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación *r* es superior o igual a aproximadamente el 60 % y el establecimiento de una etiqueta *N* si la relación *r* está por debajo de aproximadamente el 60 %.

Más preferentemente, el método y el sistema comprenden en la creación de la secuencia consenso establecer la representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación *r* es superior o igual a aproximadamente el 65 % y el establecimiento de una etiqueta N si la relación r está por debajo de aproximadamente el 65 %.

Más preferentemente, el método y el sistema comprenden en la creación de la secuencia consenso establecer la representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación *r* es superior o igual a aproximadamente el 68 % y el establecimiento de una etiqueta N si la relación r está por debajo de aproximadamente el 68 %.

Más preferentemente, el método y el sistema comprenden en la creación de la secuencia consenso establecer la representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación r es superior o igual a aproximadamente el 70 % y el establecimiento de una etiqueta N si la relación r está por debajo de aproximadamente el 70 %.

Más preferentemente, el método y el sistema comprenden en la creación de la secuencia consenso establecer la representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación *r* es superior o igual a aproximadamente el 71 % y el establecimiento de una etiqueta *N* si la relación r está por debajo de aproximadamente el 71 %.

Más preferentemente, el método y el sistema comprenden en la creación de la secuencia consenso establecer la representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación *r* es superior o igual a aproximadamente el 72 % y el establecimiento de una etiqueta *N* si la relación *r* está por debajo de aproximadamente el 72 %.
Más preferentemente, el método y el sistema comprenden en la creación de la secuencia consenso establecer la representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación *r* es superior o igual a aproximadamente el 73 % y el establecimiento de una etiqueta *N* si la relación está por debajo de aproximadamente el 73 %.
Más preferentemente, el método y el sistema comprenden en la creación de la secuencia consenso establecer la representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación *r* es superior o igual a aproximadamente el 74 % y el establecimiento de una etiqueta *N* si la relación *r* está por debajo de aproximadamente el 74 %.

Más preferentemente, el método y el sistema comprenden en la creación de la secuencia consenso establecer la representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación *r* es superior o igual a aproximadamente el 75 % y el establecimiento de una etiqueta *N* si la relación *r* está por debajo de aproximadamente el 75 %.
Más preferentemente, el método y el sistema comprenden en la creación de la secuencia consenso establecer la representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación *r* es superior o igual a aproximadamente el 76 % y el establecimiento de una etiqueta *N* si la relación r está por debajo de aproximadamente el 76 %.
Además, el método y el sistema comprenden en la creación de la secuencia consenso establecer la representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación *r* es superior o igual a aproximadamente el 99 % y el establecimiento de una etiqueta *N* si la relación *r* está por debajo de aproximadamente el 99 %.
Preferentemente, el método y el sistema comprenden en la creación de la secuencia consenso establecer la representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación r es superior o igual a aproximadamente el 95 % y el establecimiento de una etiqueta *N* si la relación r está por debajo de aproximadamente el 95 %.
Más preferentemente, el método y el sistema comprenden en la creación de la secuencia consenso establecer la representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación es superior o igual a aproximadamente el 90 % y el establecimiento de una etiqueta *N* si la relación r está por debajo de aproximadamente el 90 %.
Más preferentemente, el método y el sistema comprenden en la creación de la secuencia consenso establecer la representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación *r* es superior o igual a aproximadamente el 85 % y el establecimiento de una etiqueta *N* si la relación está por debajo de aproximadamente el 85 %.
Más preferentemente, el método y el sistema comprenden en la creación de la secuencia consenso establecer la representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación *r* es superior o igual a aproximadamente el 80 % y el establecimiento de una etiqueta N si la relación *r* está por debajo de aproximadamente el 80 %.
Más preferentemente, el método y el sistema comprenden en la creación de la secuencia consenso establecer la representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación *r* es superior o igual a aproximadamente el 79 % y el establecimiento de una etiqueta *N* si la relación *r* está por debajo de aproximadamente el 79 %.
Más preferentemente, el método comprende en la creación de la secuencia consenso establecer la representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación *r* es superior o igual a aproximadamente el 78 % y el establecimiento de una etiqueta N si la relación *r* está por debajo de aproximadamente el 78 %.
Más preferentemente, el método y el sistema comprenden en la creación de la secuencia consenso establecer la representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación *r* es superior o igual a aproximadamente el 77 % y el establecimiento de una etiqueta *N* si la relación *r* está por debajo de aproximadamente el 77 %.

En determinados aspectos, la creación de la secuencia consenso también puede realizarse para cada grupo de la segunda pluralidad de grupos. En determinados aspectos, la creación de la secuencia consenso también puede realizarse para cada grupo de al menos una pluralidad adicional de grupos o una segunda pluralidad de grupos.

La creación de una secuencia consenso de una segunda pluralidad de grupos o al menos una pluralidad adicional de grupos puede implicar un alineamiento adicional de la pluralidad de lecturas genéticas con al menos una secuencia genética de referencia y/o agrupar las lecturas genéticas.
Como se ha descrito anteriormente, las lecturas genéticas del grupo' de la pluralidad de grupos pueden corresponder a una primera cadena de un ácido nucleico bicatenario y las lecturas genéticas del grupo" pueden corresponder a la segunda cadena complementaria de dicho ácido nucleico bicatenario. Los métodos y sistemas de la invención pueden realizar la etapa de crear una secuencia consenso como se define en el presente documento para al menos un

segundo grupo o un grupo adicional.

El grupo" de ejemplo o grupo adicional puede corresponder a la segunda cadena complementaria, por lo tanto puede crearse la secuencia consenso de un ácido nucleico bicatenario. En el presente documento, dicha secuencia consenso también se denomina secuencia consenso bicatenaria o secuencias consenso dúplex. La secuencia consenso bicatenaria correspondiente puede crearse determinando el nucleótido más abundante *man_p* en cada posición específica *p* de una pluralidad de posiciones dentro de las lecturas genéticas del correspondiente grupo' y grupo" y en donde cada posición corresponde a un par de bases.

La secuencia consenso bicatenaria podría crearse mediante varios procedimientos, no limitados a aquellos proporcionados en lo siguiente.
El método para la secuenciación de ácidos nucleicos, particularmente para reducir el número de falsos positivos en la secuenciación de ácidos nucleicos, puede comprender las siguientes etapas:

(a) (opcionalmente obtener una pluralidad de lecturas genéticas mediante la secuenciación de una muestra de ácido nucleico;)
(b) alinear la pluralidad de lecturas genéticas con al menos una secuencia genética de referencia;
(c) agrupar las lecturas genéticas que comparten una posición genética en una secuencia genética de referencia de la al menos una secuencia genética de referencia en una pluralidad de grupos;
(d) crear una secuencia consenso para cada grupo de la pluralidad de grupos, en donde se crea una secuencia consenso correspondiente determinando el nucleótido más abundante *man_p* en cada posición específica *p* de una pluralidad de posiciones dentro del grupo correspondiente de lecturas genéticas y

(i) establecer una representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación r entre el número de lecturas genéticas dentro del grupo correspondiente que tiene el nucleótido más abundante *man_p* en la posición específica *p* y el número de lecturas genéticas dentro del un grupo correspondiente es superior o igual a un umbral predeterminado *t;* y
(ii) establecer una etiqueta *N* si la relación *r* está por debajo del umbral predeterminado *t,*

en donde el grupo correspondiente es un grupo' y en donde se crea una secuencia consenso para un grupo" realizando las etapas (i) y (ii) para un grupo adicional.

En dichos aspectos, las lecturas genéticas del grupo' pueden corresponder a una primera cadena de un ácido nucleico bicatenario y las lecturas genéticas del grupo" pueden corresponder a la segunda cadena complementaria de dicho ácido nucleico bicatenario.

También puede crearse la secuencia consenso bicatenaria, en donde la representación del nucleótido más abundante *man_p* en la posición específica *p* se establece en la secuencia consenso bicatenaria si la representación está presente en la secuencia consenso para el grupo' correspondiente a la primera cadena y si la representación está presente en la secuencia consenso para el grupo" correspondiente a la segunda cadena complementaria de dicho ácido nucleico bicatenario.

En particular, la secuencia consenso bicatenaria puede crearse al

(i) establecer una representación del nucleótido más abundante *man_p* en cada posición específica *p* de una pluralidad de posiciones en la secuencia consenso bicatenaria si la representación o la etiqueta en la posición específica *p* está presente respectivamente en ambas secuencias consenso monocatenarias del grupo' y del grupo"; y
(ii) establecer la etiqueta "N" en cada posición específica *p* de una pluralidad de posiciones en la secuencia consenso bicatenaria si la etiqueta "N" está presente en la posición específica *p* en una de las secuencias consenso monocatenarias del grupo' o del grupo" o si la representación del nucleótido más abundante *man_p* no es idéntica en la posición específica en ambas secuencias consenso monocatenarias del grupo' o del grupo".

La etiqueta "N" en la posición específica *p* puede establecerse en la secuencia consenso bicatenaria

(i) si la etiqueta N o la representación del nucleótido más abundante *man_p* en la posición específica *p* está presente en la secuencia consenso para el grupo' correspondiente a la primera cadena y si la etiqueta N o la representación del nucleótido más abundante *man_p* en la posición específica *p* no está presente en la secuencia consenso para el grupo" correspondiente a la segunda cadena complementaria de dicho ácido nucleico bicatenario; o
(ii) si la etiqueta N en la posición específica *p* está presente en la secuencia consenso para el grupo' correspondiente a la primera cadena y si la etiqueta N en la posición específica *p* está presente en la secuencia consenso para el grupo" correspondiente a la segunda cadena complementaria de dicho ácido nucleico bicatenario.

Por lo tanto, el método puede comprender la agrupación y la unidad de cálculo puede configurarse para agrupar las correspondientes secuencias consenso monocatenarias y colapsar las correspondientes secuencias consenso monocatenarias en una secuencia consenso bicatenaria, en donde la representación del nucleótido más abundante *man_p* se establece o la etiqueta "N" se establece en la posición específica *p* en la secuencia consenso bicatenaria si la representación o la etiqueta en la posición específica *p* está presente respectivamente en ambas de las dos secuencias consenso monocatenarias correspondientes, o en donde la etiqueta "N" se establece en la posición específica *p* en la secuencia consenso bicatenaria si la etiqueta "N" está presente en la posición específica *p* en una de las dos secuencias consenso monocatenarias correspondientes.

En particular, el método puede comprender la agrupación y la unidad de cálculo puede configurarse para agrupar las correspondientes secuencias consenso monocatenarias basadas en secuencias de código de barras y colapsar las correspondientes secuencias consenso monocatenarias en una secuencia consenso bicatenaria, en donde la representación del nucleótido más abundante *man_p* se establece o la etiqueta "N" se establece en la posición específica *p* en la secuencia consenso bicatenaria si la representación o la etiqueta en la posición específica *p* está presente respectivamente en ambas de las dos secuencias consenso monocatenarias correspondientes, o en donde la etiqueta "N" se establece en la posición específica *p* en la secuencia consenso bicatenaria si la etiqueta "N" está presente en la posición específica *p* en una de las dos secuencias consenso monocatenarias correspondientes.

En consecuencia, el método puede comprender además: crear una secuencia consenso bicatenaria al

(i) agrupar la pluralidad de grupos que comparten la posición genética en una secuencia genética de referencia por una característica o características particulares adicionales, por ejemplo, secuencia o secuencias de código de barras, en donde la pluralidad de grupos comprende el grupo' y el grupo'', en donde el grupo' comprende lecturas genéticas correspondientes al complemento inverso de las lecturas genéticas comprendidas en el grupo'',
(ii) crear una secuencia consenso bicatenaria al

(i) establecer una representación del nucleótido más abundante *man_p* o la etiqueta "N" en cada posición específica *p* de una pluralidad de posiciones en la secuencia consenso bicatenaria si la representación o la etiqueta en la posición específica *p* está presente respectivamente en ambas secuencias consenso monocatenarias del grupo' y del grupo''; y
(ii) establecer la etiqueta "N" en cada posición específica *p* de una pluralidad de posiciones en la secuencia consenso bicatenaria si la etiqueta "N" está presente en la posición específica *p* en una de las secuencias consenso monocatenarias del grupo' o del grupo'' o si la representación del nucleótido más abundante *man_p* no es idéntica en la posición específica en ambas secuencias consenso monocatenarias del grupo' o del grupo''.

El método de la invención es particularmente adecuado para identificar (una) variación o variaciones (genéticas) en la secuencia de ácido nucleico de un sujeto.

En particular, para identificar la variación o variaciones, el método de la invención compara y el sistema de la invención está configurado para comparar las secuencias consenso de la pluralidad de grupos con la secuencia genética de referencia o la al menos una secuencia genética de referencia en cada posición específica *p* de una pluralidad de posiciones de las secuencias consenso y en donde una diferencia en una posición específica entre las secuencias consenso de la pluralidad de grupos y la secuencia genética de referencia o la al menos una secuencia genética de referencia indica una variación (genética) en la posición específica. Esta etapa también se ejemplifica en las Figuras 4 y 5.

En este contexto, las secuencias consenso de la pluralidad de grupos se comparan con la secuencia genética de referencia, en donde la pluralidad de grupos puede referirse a al menos aproximadamente el 50 %, preferentemente al menos aproximadamente el 60 %, más preferentemente al menos aproximadamente el 70 %, más preferentemente al menos aproximadamente el 80 %, más preferentemente al menos aproximadamente el 90 %, más preferentemente al menos aproximadamente el 99 % o lo más preferentemente el 100 % de los grupos que están formados por la agrupación de las lecturas genéticas que comparten una posición genética en una secuencia genética de referencia.

Como se usa en el presente documento, la expresión "comparar las secuencias consenso de la pluralidad de grupos con la secuencia genética de referencia en cada posición específica *p* de una pluralidad de posiciones de las secuencias consenso" o variantes gramaticales de la misma significa que cada posición específica nucleótido por nucleótido de al menos aproximadamente el 50 %, preferentemente al menos aproximadamente el 60 %, más preferentemente al menos aproximadamente el 70 %, más preferentemente al menos aproximadamente el 80 %, más preferentemente al menos aproximadamente el 90 %, más preferentemente al menos aproximadamente el 99 % o lo más preferentemente el 100 % de las posiciones de las secuencias consenso se comparan con la secuencia genética de referencia.

En aspectos particulares, las secuencias consenso de la pluralidad de grupos se comparan con la secuencia genética de referencia en cada posición específica *p* de una pluralidad de posiciones de las secuencias consenso, en donde se

comparan todas las posiciones respectivas y en donde una diferencia en una posición específica entre las secuencias consenso y la secuencia genética de referencia indica una variación (genética) en la posición específica.

En aspectos particulares, las secuencias consenso de la pluralidad de grupos se comparan con la secuencia genética de referencia en cada posición específica *p* de una pluralidad de posiciones de las secuencias consenso, en donde todas las posiciones respectivas y en donde todas las secuencias consenso se comparan con la secuencia genética de referencia y en donde una diferencia en una posición específica entre las secuencias consenso y la secuencia genética de referencia indica una variación genética en la posición específica.

Como se usa en el presente documento, el término "comparar" o variantes gramaticales del mismo en este contexto significa que las secuencias consenso están alineadas con la secuencia genética de referencia. Una diferencia en una posición específica entre las secuencias consenso creadas y la secuencia genética de referencia indica una variación (genética) en una posición específica. En esta etapa, se compara el nucleótido (la representación de los nucleótidos) de las secuencias consenso y la secuencia genética de referencia. En particular, la variación (genética) no está indicada en la posición específica si la etiqueta N se fija en la posición específica de la secuencia consenso. En consecuencia, se indica una variación (genética) en una posición específica si se incluye una diferencia entre una representación del conjunto de nucleótidos en la posición específica en una secuencia consenso correspondiente y el nucleótido de la secuencia genética de referencia. En consecuencia, la diferencia como se usa en este contexto no se refiere a la etiqueta N.

En aspectos particulares, las secuencias consenso de la pluralidad de grupos se comparan con la secuencia genética de referencia en cada posición específica *p* de una pluralidad de posiciones de las secuencias consenso y en donde una diferencia en una posición específica entre el nucleótido más abundante *man_p* de una secuencia consenso correspondiente y la secuencia genética de referencia indica una variación en la posición específica.

En consecuencia, el método de la invención compara y el sistema de la invención está configurado para comparar las secuencias consenso de la pluralidad de grupos con la secuencia genética de referencia en cada posición específica p de una pluralidad de posiciones de las secuencias consenso y en donde una diferencia en un posición específica entre el nucleótido más abundante *man_p* de las secuencias consenso y la representación de la secuencia genética de referencia indica una variación genética en la posición específica.

En otras palabras, el método de la invención compara y el sistema de la invención está configurado para comparar las secuencias consenso de la pluralidad de grupos con la secuencia genética de referencia en cada posición específica *p* de una pluralidad de posiciones de las secuencias consenso, si está comprendida una diferencia en una posición específica entre la representación del nucleótido o nucleótidos de las secuencias consenso y el nucleótido de la secuencia genética de referencia.

En determinados aspectos, las secuencias consenso de la segunda pluralidad de grupos o de al menos una pluralidad adicional de grupos pueden compararse con la secuencia genética de referencia.

Además, el método de la invención determina y el sistema de la invención está configurado para determinar el número de secuencias consenso que comprenden la variación en cada posición específica *p* de una pluralidad de posiciones y el método de la invención determina y el sistema de la invención se configura para determinar el número de secuencias consenso que comprenden la etiqueta *N* en cada posición específica *p* de una pluralidad de posiciones. Esta etapa también se ejemplifica en las Figuras 4 y 5. Como se usa en el presente documento, la expresión "determinar el número de secuencias consenso que comprenden la variación en cada posición específica *p* de una pluralidad de posiciones" o variantes gramaticales de la misma se refiere a determinar el número de secuencias consenso de la pluralidad de grupos. En este contexto la pluralidad de grupos puede referirse a al menos aproximadamente el 50 %, preferentemente al menos aproximadamente el 60 %, más preferentemente al menos aproximadamente el 70 %, más preferentemente al menos aproximadamente el 80 %, más preferentemente al menos aproximadamente el 90 %, más preferentemente al menos aproximadamente el 99 % o lo más preferentemente el 100 % de los grupos que están formados por la agrupación de las lecturas genéticas que comparten una posición genética en una secuencia genética de referencia. En particular, la pluralidad de grupos (en la etapa de comparación) puede corresponder a la pluralidad de grupos consultados en la etapa de comparación. En consecuencia, las secuencias consenso de las que se determina el número pueden referirse a las secuencias consenso de la pluralidad de grupos que se comparan con la secuencia genética de referencia.
Además, como se usa en el presente documento, la expresión "el número de secuencias consenso que comprenden la variación en cada posición específica *p* de una pluralidad de posiciones" se refiere al número de secuencias consenso que comprenden la variación en cada posición específica *p* de una pluralidad de posiciones. En consecuencia, el número de variaciones en cada posición específica de al menos aproximadamente el 50 %, preferentemente al menos aproximadamente el 60 %, más preferentemente al menos aproximadamente el 70 %, más preferentemente al menos aproximadamente el 80 %, más preferentemente al menos aproximadamente el 90 %, más preferentemente al menos aproximadamente el 99 % o lo más preferentemente el 100 % de las posiciones, se determina.
En aspectos particulares, el número de secuencias consenso que comprenden la variación en cada posición específica

p de una pluralidad de posiciones se determina, en donde se determina el número de variaciones de todas las posiciones respectivas.

En aspectos particulares, el número de secuencias consenso que comprenden la variación en cada posición específica *p* de una pluralidad de posiciones se determina, en donde se determina el número de variaciones de todas las secuencias consenso de la pluralidad de grupos.

En aspectos particulares, el número de secuencias consenso que comprenden la variación en cada posición específica *p* de una pluralidad de posiciones se determina, en donde se determina el número de variaciones de todas las posiciones respectivas y todas las secuencias consenso de la pluralidad de grupos.

En determinados aspectos, el número de secuencias consenso de la segunda pluralidad de grupos o de al menos una pluralidad adicional de grupos que comprenden la variación en cada posición específica *p* de una pluralidad de posiciones puede determinarse y el número de las secuencias consenso de la segunda pluralidad de grupos o de al menos una pluralidad adicional de grupos que comprenden la etiqueta *N* en cada posición específica *p* de una pluralidad de posiciones puede determinarse.

Además, el método de la invención determina y el sistema de la invención se configura para determinar el número de secuencias consenso que comprenden la etiqueta *N* en cada posición específica *p* de una pluralidad de posiciones. Las definiciones y explicaciones proporcionadas anteriormente también se aplican cambiando lo que se tenga que cambiar a la determinación del número de secuencias consenso que comprenden N.

Además, el método de la invención identifica y el sistema de la invención está configurado para identificar la variación (genética) en cada posición específica *p* de una pluralidad de posiciones como una verdadera variación (genética) si una relación *r** entre el número de secuencias consenso que comprenden la etiqueta *N* en la posición específica *p* y el número de secuencias consenso que comprenden la variación (genética) en la posición específica *p* está por debajo de un umbral *t**.
En consecuencia, esta etapa usa los números determinados de las secuencias consenso que comprenden la variación (genética) o la etiqueta N en cada posición específica *p* de una pluralidad de posiciones para decidir si la variación (genética) es una variación genética verdadera o falsa. Esta etapa también se ejemplifica en las Figuras 4 y 5.
La "variación" se refiere en el presente documento a una variación en un nucleótido que se produce en una posición específica del genoma, en donde un nucleótido tal no se encuentra en una posición homóloga o correspondiente en la al menos una secuencia genética de referencia. La variación también puede entenderse como mutación, polimorfismo de un solo nucleótido (SNP) o alelo variante. Las variaciones pueden incluir una sustitución o sustituciones, deleción o deleciones o adición o adiciones en cualquier posición en la secuencia de ácido nucleico en comparación con la al menos una secuencia genética de referencia. El término "variación" puede significar una "variación genética".

Como se usa en el presente documento, la expresión "variación verdadera" se refiere a una variación determinada en las secuencias consenso que cumple los criterios del filtro N. En consecuencia, la variación determinada en la secuencia consenso es una verdadera variación si una relación *r** entre el número de secuencias consenso que comprenden la etiqueta *N* en la posición específica *p* y el número de secuencias consenso que comprenden la variación en la posición específica *p* está por debajo de un umbral *t**. En consecuencia, la variación verdadera tiene una probabilidad reducida de ser una variación falsa positiva, por ejemplo, una llamada de mutación falsa. Por lo tanto, la probabilidad verdadera puede tener una alta probabilidad de ser una llamada de mutación verdadera.

Como se usa en el presente documento, el umbral *"t**"* o" t * "se refiere a cualquier número que se use como límite para la relación *r** entre el número de secuencias consenso que comprenden la etiqueta *N* en la posición específica *p* y el número de secuencias consenso que comprenden la variación en la posición específica *p* está por debajo de un umbral *t**.

Como se usa en el presente documento, la relación *"r**"* o "*r**"se refiere a la relación entre el número de secuencias consenso que comprenden la etiqueta N en la posición específica p y el número de secuencias consenso que comprenden la variación en la posición específica p.

En determinados aspectos, la variación se identifica en cada posición específica *p* de una pluralidad de posiciones como una verdadera variación, en donde la pluralidad de grupos se refiere a al menos aproximadamente el 50 %, preferentemente al menos aproximadamente el 60 %, más preferentemente al menos aproximadamente el 70 %, más preferentemente al menos aproximadamente el 80 %, más preferentemente al menos aproximadamente el 90 %, más preferentemente al menos aproximadamente el 99 % o lo más preferentemente el 100 % de las posiciones de la secuencia o secuencias consenso.

En aspectos particulares, la variación se identifica como una variación verdadera en cada posición específica *p* de una pluralidad de posiciones de las secuencias consenso de la pluralidad de grupos. En este contexto, la pluralidad de

grupos puede referirse a al menos aproximadamente el 50 %, preferentemente al menos aproximadamente el 60 %, más preferentemente al menos aproximadamente el 70 %, más preferentemente al menos aproximadamente el 80 %, más preferentemente al menos aproximadamente el 90 %, más preferentemente al menos aproximadamente el 99 % o lo más preferentemente el 100 % de los grupos que se forman mediante la agrupación de las lecturas genéticas. En particular, la pluralidad de grupos (en la etapa de identificación) puede corresponder a la pluralidad de grupos consultados en la etapa de comparación y/o etapa de determinación.

En aspectos particulares, se identifica la variación en cada posición específica *p* de una pluralidad de posiciones como una verdadera variación, en donde la verdadera variación debe identificarse en todas las posiciones respectivas.

En aspectos particulares, se identifica la variación en cada posición específica *p* de una pluralidad de posiciones como una verdadera variación, en donde la verdadera variación debe identificarse en todas las respectivas secuencias consenso de la pluralidad de grupos.

En aspectos particulares, se identifica la variación en cada posición específica *p* de una pluralidad de posiciones como una verdadera variación, en donde la verdadera variación debe identificarse en todas las posiciones respectivas y en todas las respectivas secuencias consenso de la pluralidad de grupos.

En determinados aspectos, puede identificarse la verdadera variación de las secuencias consenso de la segunda pluralidad de grupos o de al menos una pluralidad adicional de grupos.

Además, la variación en cada posición específica *p* de una pluralidad de posiciones puede identificarse como una verdadera variación si una relación *r** entre el número de secuencias consenso que comprenden la etiqueta *N* en la posición específica *p* y el número de secuencias consenso que comprenden la variación en la posición específica *p* es igual o inferior a aproximadamente 1, 1,1, 1,2, 1,3, 1,4, 1,5, 1,6, 1,7, 1,8.

En aspectos particulares, la variación en cada posición específica *p* de una pluralidad de posiciones se identifica como una verdadera variación si una relación *r** entre el número de secuencias consenso que comprenden la etiqueta *N* en la posición específica *p* y el número de secuencias consenso que comprenden la variación en la posición específica *p* es igual a o por debajo de 1,8.

En particular, la relación r es igual o por encima del 76 % y la relación *r** es igual a o por debajo de 1,8.

En particular, el método y los sistemas de la invención son particularmente adecuados para identificar variaciones con una frecuencia alélica de menos del 5 %, el 4 %, el 3 %, el 2 % o particularmente el 1 %. el método y los sistemas de la invención son particularmente adecuados para identificar variaciones con una frecuencia alélica de menos del 0,01 %, el 0,05 %, el 0,1 %, el 0,2 % o particularmente el 1 %.

El número de las lecturas genéticas que abarcan una posición específica de una pluralidad de posiciones es una consecuencia de la profundidad de secuenciación aplicada. En consecuencia, cuanto mayor es la profundidad de secuenciación, más lecturas cubren una posición definida. Una sub fracción de estas lecturas puede contener una variante y otra sub fracción puede contener "N". La relación de N y la variante en dicha posición se usa para determinar la proporción de la variante "N" que se usa para filtrar nuestras llamadas de mutación falsa positiva.

En determinados aspectos, la secuenciación de la muestra de ácido nucleico tiene una profundidad/cobertura de secuenciación de 100 veces a 50000 veces. En determinados aspectos, la secuenciación de la muestra de ácido nucleico tiene una profundidad/cobertura de secuenciación de al menos 100 veces a menos de 50000 veces.

La variación o variaciones verdaderas también pueden identificarse empleando la secuencia consenso bicatenaria. La invención se refiere a un método y sistema que se configura para

(a) opcionalmente obtener una pluralidad de lecturas genéticas mediante la secuenciación de una muestra de ácido nucleico;
(b) alinear la pluralidad de lecturas genéticas con al menos una secuencia genética de referencia;
(c) agrupar las lecturas genéticas que comparten una posición genética en una secuencia genética de referencia de la al menos una secuencia genética de referencia en una pluralidad de grupos;
(d) crear una secuencia consenso para cada grupo de la pluralidad de grupos, en donde se crea una secuencia consenso correspondiente determinando el nucleótido más abundante *man_p* en cada posición específica *p* de una pluralidad de posiciones dentro del grupo correspondiente de lecturas genéticas y

(i) establecer una representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación r entre el número de lecturas genéticas dentro del grupo correspondiente que tiene el nucleótido más abundante *man_p* en la posición específica *p* y el número de lecturas genéticas dentro del un grupo correspondiente es superior o igual a un umbral predeterminado *t;* y

(ii) establecer una etiqueta *N* si la relación *r* está por debajo del umbral predeterminado *t*, en donde dicho umbral *t* es al menos aproximadamente el 76 %, en donde el grupo correspondiente es un grupo' y en donde se crea una secuencia consenso para un grupo'' realizando las etapas (i) y (ii) para un grupo adicional y
crear una secuencia consenso bicatenaria al
(iii) establecer una representación del nucleótido más abundante *man_p* o la etiqueta "N" en cada posición específica *p* de una pluralidad de posiciones en la secuencia consenso bicatenaria si la representación o la etiqueta en la posición específica *p* está presente respectivamente en ambas secuencias consenso monocatenarias del grupo' y del grupo''; y
(iv) establecer la etiqueta "N" en cada posición específica *p* de una pluralidad de posiciones en la secuencia consenso bicatenaria si la etiqueta "N" está presente en la posición específica *p* en una de las secuencias consenso monocatenarias del grupo' o del grupo'' o si la representación del nucleótido más abundante *man_p* no es idéntica en la posición específica en ambas secuencias consenso monocatenarias del grupo' o del grupo'';

(e) comparar la secuencia o secuencias consenso bicatenarias y/o las secuencias consenso de la pluralidad de grupos con la secuencia genética de referencia en cada posición específica *p* de una pluralidad de posiciones de las secuencias consenso y en donde una diferencia en una posición específica entre las secuencias consenso y la secuencia genética de referencia indica una variación en la posición específica;
(f) determinar el número de secuencias consenso que comprenden la variación en cada posición específica *p* de una pluralidad de posiciones y determinar el número de secuencias consenso que comprenden la etiqueta *N* en cada posición específica *p* de una pluralidad de posiciones; y
(g) identificar la variación en cada posición específica *p* de una pluralidad de posiciones como una verdadera variación si una relación *r** entre el número de secuencias consenso que comprenden la etiqueta *N* en la posición específica *p* y el número de secuencias consenso que comprenden la variación en la posición específica *p* está por debajo de un umbral *t**, en donde dicho umbral *t** es igual o por debajo de 1,8.

Como se ha descrito anteriormente, puede emplearse la amplificación de los ácidos nucleicos/moléculas de ácido nucleico. La amplificación es un método para generar grandes cantidades de una secuencia diana. En general, uno o más cebadores de amplificación se hibridan con una secuencia de ácido nucleico. Usando enzimas apropiadas, las secuencias que se encuentran adyacentes o entre los cebadores se amplifican. Antes o simultáneamente con el análisis, la muestra puede amplificarse mediante una diversidad de mecanismos. En algunos aspectos, los métodos de amplificación de ácidos nucleicos tales como la PCR pueden combinarse con los métodos y sistemas desvelados. Véase, por ejemplo, PCR Technology: Principles and Applications for DNA Amplification (Ed. H.A. Erlich, Freeman Press, NY, NY, 1992); PCR Protocols: A Guide to Methods and Applications (Eds. Innis, et al., Academic Press, San Diego, CA, 1990); Mattila et al., Nucleic Acids Res. 19, 4967 (1991); Eckert et al., PCR Methods and Applications 1, 17 (1991); PCR (Eds. McPherson et al., IRL Press, Oxford).

Como se usa en el presente documento, el "sujeto" (o "paciente") puede ser un vertebrado. En el contexto de la presente invención, el término "sujeto" incluye tanto a humanos como a animales, particularmente mamíferos y otros organismos. Por lo tanto, los métodos proporcionados en el presente documento son aplicables tanto a sujetos humanos como animales. En consecuencia, dicho sujeto puede ser un animal tales como un ratón, rata, hámster, conejo, cobaya, hurón, gato, perro, pollo, oveja, especie bovina, caballo, camello o primate. Preferentemente, el sujeto es un mamífero. Lo más preferentemente, el sujeto es un ser humano. El método, los sistemas y kits proporcionados en el presente documento pueden usarse en cualquier sujeto que sea un sujeto sano o un sujeto que padezca cualquier enfermedad o trastorno. En aspectos preferidos, el sujeto padece una enfermedad, trastorno o afección médica.

La muestra de ácido nucleico, el ácido nucleico o la molécula de ácido nucleico puede ser, por ejemplo, o puede comprender ADN, ARN, amplicones, ADNc, ADNbc, ADNmc, ADN plasmídico, ADN cósmido, ADN de alto peso molecular (MW), ADN cromosómico, ADN genómico, ADN vírico, ADN bacteriano, ADNmt (ADN mitocondrial), ARNm, ARNr, ARNt, ARNn, ARNip, ARNnp, ARNnop, ARNcap, microARN, ARNbc, ribozima, ribointerruptor, híbrido de ADN-ARN y ARN vírico (por ejemplo, ARN retrovírico).

Como se usa en el presente documento, el término "muestra" es una muestra biológica que se obtiene del sujeto. "Muestra" como se usa en el presente documento puede, por ejemplo, referirse a una muestra de líquido o tejido corporal obtenida con fines de diagnóstico, pronóstico o evaluación de un sujeto de interés, tal como un paciente. Preferentemente en el presente documento, la muestra es una muestra de un fluido corporal, tales como sangre, suero, plasma, pleura, líquido cefalorraquídeo, orina, saliva, esputo y derrames pleurales. Particularmente, la muestra es sangre, plasma sanguíneo, suero sanguíneo u orina. La muestra puede estar pre-tratada, por ejemplo, mediante procedimientos de purificación, por ejemplo, separación de sangre completa en componentes de suero o plasma. Dichos pretratamientos también pueden incluir, pero no se limitan a dilución, filtración, centrifugación, concentración, sedimentación, precipitación o diálisis. Los pretratamientos también pueden incluir la adición de sustancias químicas o bioquímicas a la solución, tales como ácidos, bases, tampones, sales, disolventes, colorantes reactivos, detergentes, emulsionantes, quelantes.

La expresión "muestra de ácido nucleico" se refiere a una muestra que comprende uno o más ácidos nucleicos o una o más moléculas de ácido nucleico. La muestra puede por lo tanto procesarse para obtener una muestra de ácido

nucleico. En consecuencia, después de obtener una muestra, por ejemplo, una muestra de plasma, los ácidos nucleicos o moléculas de ácido nucleico pueden aislarse y/o extraerse de la muestra para obtener una muestra de ácido nucleico. Los ácidos nucleicos o moléculas de ácido nucleico pueden aislarse/extraerse de la muestra mediante cualquier método físico y químico conocido por la persona experta. Por ejemplo, los ácidos nucleicos pueden extraerse de una muestra mediante extracción con fenol-cloroformo, precipitación diferencial, precipitación en etanol, separación en gel o separación en fase sólida.

Pueden usarse etapas de limpieza adicionales tales como columnas basadas en sílice para retirar contaminantes o sales. Las etapas generales pueden optimizarse para aplicaciones específicas. Los polinucleótidos vehículo en masa no específicos, por ejemplo, puede añadirse a lo largo de la reacción para optimizar ciertos aspectos del procedimiento tales como el rendimiento. El aislamiento y la purificación de moléculas de ácido nucleico pueden lograrse usando cualquier medio, incluyendo, pero no limitado a, el uso de kits y protocolos comerciales proporcionados por empresas tales como Sigma Aldrich, Life Technologies, Promega, Affymetrix, IBI o similares. Los kits y protocolos también pueden estar disponibles no comercialmente. Después del aislamiento, en algunos casos, la muestra de ácido nucleico puede pre-mezclarse con uno o más materiales adicionales, tales como uno o más reactivos (por ejemplo, ligasa, proteasa, polimerasa) antes de la secuenciación.

La molécula o moléculas de ácido nucleico o el ácido o ácidos nucleicos de la que se determina la secuencia de ácido nucleico y que se encuentran en la muestra de ácido nucleico pueden someterse a fragmentación para obtener una longitud de las moléculas de ácido nucleico que sea adecuada para secuenciación ácida. Las moléculas de ácido nucleico que tienen más de 1000 nucleótidos habitualmente se fragmentan. La muestra de ácido nucleico puede comprender moléculas de ácido nucleico con una longitud de menos de 1000 nucleótidos.

Las moléculas de ácido nucleico que van a secuenciarse pueden obtenerse/aislarse primero de una muestra biológica. Estas moléculas de ácido nucleico pueden fragmentarse después hasta una longitud óptima. La longitud óptima puede determinarse por la plataforma aguas abajo. El experto en la materia conoce técnicas para fragmentar el ácido nucleico obtenido de una muestra. Las técnicas de ejemplo son el cizallamiento acústico, aplicación de fuerzas de nebulización, sometimiento a ultrasonidos, cizallamiento de agujas, tratamientos de presión francesa y/o basados en enzimas mediante la escisión simultánea de ambas cadenas o mediante la generación de mellas en cada cadena de ADNbc para producir roturas de ADNbc.

La presente invención también se refiere a un producto de programa informático que comprende uno o más medios legibles por ordenador que tienen ejecutables por ordenador para instrucciones para realizar las etapas de cualquiera de los métodos proporcionados en el presente documento.

La invención se refiere además a un sistema que puede emplearse en los métodos proporcionados en el presente documento. En particular, la presente invención se refiere a un sistema para la secuenciación de ácidos nucleicos, particularmente para reducir el número de falsos positivos en la secuenciación de ácidos nucleicos. Las definiciones y explicaciones proporcionadas en el presente documento en relación con los métodos y las etapas de los métodos (por ejemplo, obtener la pluralidad de lecturas genéticas, alinear, agrupar y crear la secuencia consenso) también se aplican cambiando lo que se tenga que cambiar al sistema de la invención. En consecuencia, el sistema proporcionado en el presente documento está configurado para emplearse en cualquier etapa de los métodos proporcionados en el presente documento. Por ejemplo, el sistema proporcionado en el presente documento comprende una unidad de cálculo que se configura para crear una secuencia consenso.

Por ejemplo, la unidad de cálculo puede configurarse para alinear la pluralidad de lecturas genéticas con al menos una secuencia genética de referencia.
Además, la unidad de cálculo puede configurarse para agrupar las lecturas genéticas que comparten una posición genética en una secuencia genética de referencia de la al menos una secuencia genética de referencia en una pluralidad de grupos.
Además, estando la unidad de cálculo configurada para crear una secuencia consenso para cada grupo de la pluralidad de grupos, en donde se crea una secuencia consenso correspondiente determinando un nucleótido más abundante *man_p* en cada posición específica *p* de una pluralidad de posiciones dentro del grupo correspondiente de lecturas genéticas y

(i) establecer una representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación r entre el número de lecturas genéticas dentro del grupo correspondiente que tiene el nucleótido más abundante *man_p* en la posición específica *p* y el número de lecturas genéticas dentro del un grupo correspondiente es superior o igual a un umbral predeterminado *t;* y
(ii) establecer una etiqueta *N* si la relación r está por debajo del umbral predeterminado *t*.

Además, estando la unidad de cálculo configurada para comparar las secuencias consenso de la pluralidad de grupos con la secuencia genética de referencia en cada posición específica *p* de una pluralidad de posiciones de las secuencias consenso y en donde una diferencia en una posición específica entre las secuencias consenso y la secuencia genética de referencia indica una variación en la posición específica.

Además, estando la unidad de cálculo configurada para determinar el número de secuencias consenso que comprenden la variación en cada posición específica *p* de una pluralidad de posiciones y determinar el número de secuencias consenso que comprenden la etiqueta *N* en cada posición específica *p* de una pluralidad de posiciones.
Además, estando la unidad de cálculo configurada para identificar la variación en cada posición específica *p* de una pluralidad de posiciones como una verdadera variación si una relación *r** entre el número de secuencias consenso que comprenden la etiqueta *N* en la posición específica *p* y el número de secuencias consenso que comprenden la variación en la posición específica *p* está por debajo de un umbral *t**.
Además, la unidad de cálculo puede configurarse para crear una secuencia consenso en cada posición específica p de una pluralidad de posiciones dentro del grupo de lecturas genéticas, en donde una representación de un nucleótido en una posición específica de la pluralidad de posiciones se establece en la secuencia consenso, si una relación entre el número de este nucleótido en esta posición específica y el número de las lecturas genéticas dentro de este grupo está por encima o es igual a un umbral predeterminado t, y en donde una etiqueta N se establece en la secuencia consenso en esta posición específica si la relación está por debajo del umbral predeterminado t.

Además, la unidad de cálculo puede configurarse para crear una secuencia consenso en cada posición específica p de una pluralidad de posiciones dentro del grupo de lecturas genéticas, en donde el nucleótido o nucleótidos se determinan en cada una de las posiciones específicas de la pluralidad de posiciones y
en donde una representación de un nucleótido del nucleótido o nucleótidos en una posición específica de la pluralidad de posiciones se establece en la secuencia consenso, si una relación entre el número de este nucleótido en esta posición específica y el número de las lecturas genéticas dentro de este grupo está por encima o es igual a un umbral predeterminado t, y en donde una etiqueta N se establece en la secuencia consenso en esta posición específica si la relación está por debajo del umbral predeterminado t.

La unidad de cálculo puede configurarse además para alinear la pluralidad de lecturas genéticas con al menos una secuencia genética de referencia y/o agrupar las lecturas genéticas antes de que la unidad de cálculo cree la secuencia consenso. En particular, la unidad de cálculo está configurada para agrupar las lecturas genéticas que pertenecen a la misma secuencia genética de referencia en un grupo correspondiente. Antes de que la unidad de cálculo cree la secuencia consenso, la unidad de cálculo también puede configurarse para alinear la pluralidad de lecturas genéticas con al menos una secuencia genética de referencia y/o para agrupar las lecturas genéticas en uno o más grupos, en donde las lecturas genéticas comprendidas en un grupo correspondiente comparten al menos una característica particular, tal como al menos una secuencia de ácido nucleico particular, por ejemplo, al menos una secuencia de código de barras.

Además de la unidad de cálculo el sistema puede comprender una unidad de obtención para obtener la pluralidad de lecturas genéticas. Preferentemente, el sistema comprende además una unidad de obtención estando configurada para obtener la pluralidad de lecturas genéticas de un ácido nucleico a secuenciar. Además, el sistema puede comprender una unidad de obtención configurada para obtener la pluralidad de lecturas genéticas mediante secuenciación de ácidos nucleicos de una muestra de ácidos nucleicos. El sistema puede comprender una unidad de obtención configurada para determinar la secuencia de ácido nucleico de los amplicones marcados mediante secuenciación de ácido nucleico y por lo tanto determinar una pluralidad de lecturas genéticas.

La unidad de cálculo puede configurarse para crear secuencias consenso respectivas y está configurada para establecer una representación respectiva o una etiqueta N respectiva para todas las posiciones respectivas dentro de uno o más grupos. Preferentemente, la unidad de cálculo está configurada para crear una secuencia consenso respectiva y está configurada para establecer una representación respectiva o una etiqueta *N* respectiva para todas las posiciones respectivas dentro del grupo. Además, la unidad de cálculo está configurada para crear una secuencia consenso respectiva de un segundo grupo o al menos un grupo adicional.

Como se describió anteriormente un segundo grupo tal puede corresponder a un grupo que comprende las lecturas genéticas de la cadena complementaria.

La invención se refiere además a kits, el uso de los kits, por ejemplo, en los métodos proporcionados anteriormente en el presente documento. Los kits pueden comprender el sistema que comprende una unidad de cálculo y el programa informático como se proporciona anteriormente. El kit puede comprender además reactivos de detección para determinar la secuencia de ácido nucleico de la muestra de ácido nucleico y determinar de esta manera la pluralidad de lecturas genéticas. Por ejemplo, los Kits NextSeq 500/550 v2 (Illumina), el Kit MiSeq Reagent v3 (Illumina), los Kits de reactivos TruSeq Exome (Illumina) o SureSelectXT (Agilent Technologies).

Dichos reactivos de detección pueden incluir adaptadores como se describió anteriormente.

El experto en la materia entenderá que la presente invención es adecuada para reducir el número de falsos positivos en una sola cadena y considerar que una única cadena en la presente invención es suficiente.

Sin embargo, la presente invención también brinda la oportunidad de explotar (i) el uso de las características

particulares, por ejemplo, secuencias de códigos de barras, para discriminar los duplicados de la reacción en cadena de la polimerasa (PCR) de las réplicas biológicas (tales como las moléculas de ADN individuales) y/o (ii) la naturaleza complementaria de las dos cadenas de una molécula de ADN. Por lo tanto, en primer lugar, mediante la identificación de duplicados de PCR, pueden formarse secuencias consenso descartando la información de la secuencia por debajo de una abundancia predefinida dentro del conjunto de duplicados de PCR de la misma molécula de ácido nucleico. En segundo lugar, se espera que las mutaciones verdaderas estén presentes en ambas cadenas de ácido nucleico, mientras que las variantes introducidas erróneamente durante el flujo de trabajo de secuenciación aparecen solo en una cadena de ácido nucleico.

Los sistemas, kits y métodos de la invención también pueden usarse en el diagnóstico, el seguimiento y/o el pronóstico de enfermedad y/o trastorno. Se conocen varias variaciones que pueden indicar que un sujeto padece o corre el riesgo de padecer (una) enfermedad o enfermedades y/o trastorno o trastornos. Por lo tanto tales variaciones podrían emplearse en el diagnóstico, el seguimiento y/o el pronóstico.

Una enfermedad y/o trastorno tales que puede seleccionarse del grupo que consiste en un trastorno de un solo gen, trastorno autosómico dominante, trastorno autosómico recesivo, trastorno dominante ligado al cromosoma X, trastorno recesivo ligado al cromosoma X, trastorno ligado al cromosoma Y, enfermedad mitocondrial y un trastorno genético asociado a múltiples genes.

Los sistemas, kits y métodos de la invención también pueden usarse en el diagnóstico, el seguimiento y/o el pronóstico de cáncer y/o enfermedad o enfermedades tumorales. Particularmente, los sistemas, kits y métodos de la invención también pueden usarse en el diagnóstico, el seguimiento y/o el pronóstico de cáncer y/o enfermedad o enfermedades tumorales, en donde la variación o variaciones en la secuencia de ácido nucleico se identifican.

"Cáncer", de acuerdo con la presente invención, se refiere a una clase de enfermedades o trastornos caracterizados por la división incontrolada de células y la capacidad de estas para propagarse, ya sea por crecimiento directo en el tejido adyacente por invasión, o por implantación en sitios distantes por metástasis, donde las células cancerosas se transportan a través del torrente sanguíneo o del sistema linfático.

La "enfermedad tumoral" puede ser cualquier forma de cáncer, un tumor o se elige de cáncer de páncreas, cáncer de mama, cáncer epitelial, carcinoma hepatocelular, cáncer colangiocelular, cáncer de estómago, cáncer de colon, cáncer de próstata, cáncer de vejiga, cáncer de lengua, cáncer de cabeza y cuello, cáncer de piel (melanoma), un cáncer del tracto urogenital, por ejemplo, cáncer de ovario, cáncer de endometrio, cáncer de cuello uterino y cáncer de riñón; cáncer de pulmón, cáncer gástrico, un cáncer del intestino delgado, cáncer de hígado, cáncer de vesícula biliar, un cáncer de los conductos biliares, cáncer de esófago, un cáncer de las glándulas salivales o un cáncer de la glándula tiroides.

Los sistemas, kits y métodos de la invención también pueden usarse en el diagnóstico, el seguimiento y/o el pronóstico de cáncer y/o (una) enfermedad o enfermedades tumorales, en donde la variación o variaciones en la secuencia de ácido nucleico se identifican.

Por ejemplo, puede obtenerse y prepararse sangre de un sujeto con riesgo de cáncer para obtener una muestra de ácido nucleico. Los sistemas y métodos proporcionados en el presente documento pueden emplearse para identificar variación o variaciones/mutación o mutaciones raras que pueden existir en ciertos cánceres. El método puede ayudar a detectar la presencia de células cancerosas en el cuerpo, a pesar de la ausencia de síntomas u otras características de la enfermedad.

Los métodos, los kits y los sistemas proporcionados en el presente documento también pueden usarse en la detección temprana de una enfermedad y/o trastorno, por ejemplo, cáncer. El sistema y los métodos proporcionados en el presente documento pueden usarse para detectar cualquier número de variación o variaciones que puedan provocar o resultar de cánceres. Estos pueden incluir pero no se limitan a mutaciones, mutaciones raras, indeles, variaciones del número de copias, transversiones, translocaciones, inversión, deleciones, aneuploidía, aneuploidía parcial, poliploidía, inestabilidad cromosómica, alteraciones de la estructura cromosómica, fusiones de genes, fusiones de cromosomas, truncamientos de genes, amplificación de genes, duplicaciones de genes, lesiones cromosómicas, lesiones del ADN, cambios anormales en las modificaciones químicas del ácido nucleico, cambios anormales en los patrones epigenéticos, cambios anormales en la metilación de ácidos nucleicos, infección y cáncer.

Los sistemas, Los kits y métodos proporcionados en el presente documento también pueden usarse para caracterizar determinados cánceres. Los datos genéticos producidos a partir del sistema y los métodos de esta divulgación pueden permitir a los médicos ayudar a caracterizar mejor una forma específica de cáncer. Muchas veces, los cánceres son heterogéneos tanto en composición como en estadios. Los datos del perfil genético pueden permitir la caracterización de subtipos específicos de cáncer que pueden ser importantes en el diagnóstico o tratamiento de ese subtipo específico. Esta información también puede proporcionar pistas al sujeto o al médico con respecto al pronóstico de un tipo específico de cáncer.

Los sistemas, Los kits y métodos proporcionados en el presente documento pueden usarse para controlar enfermedad o enfermedades y/o trastorno o trastornos ya diagnosticados. Esto puede permitir que un sujeto o un médico adapten las opciones de tratamiento de acuerdo con el progreso de la enfermedad y/o trastorno. Por ejemplo, los sistemas y métodos descritos en este documento pueden usarse para construir perfiles genéticos de un sujeto particular del curso de la enfermedad y/o trastorno. En algunos casos, los cánceres pueden progresar, volverse más agresivos y genéticamente inestables. En otros ejemplos, los cánceres pueden permanecer benignos, inactivos, latentes o en remisión. El sistema y los métodos de esta divulgación pueden ser útiles para determinar la progresión de la enfermedad y/o trastorno, la remisión o la recurrencia.

Además, los sistemas, los kits y los métodos proporcionados en el presente documento pueden ser útiles para determinar la eficacia de una opción de tratamiento particular. En un ejemplo, las opciones de tratamiento exitosas pueden en realidad aumentar o disminuir la cantidad de (una) variación o variaciones o mutación o mutaciones raras identificadas en la muestra de ácido nucleico en respuesta al tratamiento. Por ejemplo, si el tratamiento es exitoso, más células cancerosas y/o tumorales pueden morir y perder ADN. En otros ejemplos, esto puede no ocurrir. En otro ejemplo, quizás determinadas opciones de tratamiento puedan estar correlacionadas con los perfiles genéticos de los cánceres a lo largo del tiempo. Esta correlación puede resultar útil para seleccionar una terapia. Adicionalmente, si se observa que un cáncer está en remisión después del tratamiento, los sistemas, los kits y los métodos descritos en el presente documento pueden ser útiles para controlar la enfermedad residual y/o el trastorno o la recurrencia de la enfermedad.

Los métodos, los kits y los sistemas proporcionados en el presente documento pueden no estar limitados a la detección de variación o variaciones asociadas a cánceres. Diversas otras enfermedades e infecciones pueden dar lugar a otros tipos de afecciones que pueden ser adecuadas para la detección y el seguimiento tempranos. Por ejemplo, en determinados casos, los trastornos genéticos o las enfermedades infecciosas pueden provocar cierto mosaicismo genético en un sujeto. Este mosaicismo genético puede causar variaciones en el número de copias y variación o variaciones (por ejemplo, mutaciones raras) que podrían observarse.
En otro ejemplo, el sistema, los kits y los métodos de la divulgación también pueden usarse para controlar los genomas de las células inmunitarias dentro del cuerpo. Las células inmunitarias, tales como células, puede experimentar una rápida expansión clonal ante la presencia de determinadas enfermedades. Las expansiones clonales pueden monitorizarse usando la detección de variación del número de copias y pueden monitorizarse determinados estados inmunitarios.

Además, los sistemas, los kits y los métodos proporcionados en el presente documento también pueden usarse para monitorizar las propias infecciones sistémicas, ya que pueden estar provocadas por un patógeno tal como una bacteria o un virus. La identificación de la variación o variaciones o de la variación del número de copias puede usarse para determinar cómo está cambiando una población de patógenos durante el transcurso de la infección. Esto puede ser particularmente importante durante las infecciones crónicas, tales como infecciones por HTV/SIDA o hepatitis, por lo que los virus pueden cambiar el estado del ciclo de vida y/o mutar en formas más virulentas durante el transcurso de la infección.

Otro ejemplo más de que el sistema, los kits y los métodos proporcionados en el presente documento pueden usarse para el seguimiento de sujetos trasplantados. Generalmente, el tejido trasplantado se somete a un cierto grado de rechazo por parte del cuerpo tras el trasplante.

Los sistemas, los kits y los métodos de la invención también pueden usarse en el genotipado de un sujeto, en donde se identifican la variación o variaciones en la secuencia de ácido nucleico del sujeto. Como se usa en el presente documento, el término "genotipado" se refiere a determinar la diferencia o diferencias en el genotipo de un sujeto analizando una muestra de ácido nucleico del sujeto y comparándola con la secuencia de otro sujeto o al menos una secuencia de referencia. El genotipado puede revelar los alelos que el sujeto ha heredado de sus padres.

Los sistemas, los kits y métodos de la invención también pueden usarse para detectar y contar alteraciones somáticas que pueden definir neoepítopos y, por lo tanto, pueden iniciar respuestas inmunitarias hacia las células tumorales. Por lo tanto, la carga mutacional específica de la muestra, así como los neoepítopos identificados, pueden funcionar como biomarcadores para estratificar a los pacientes para el tratamiento con inhibidores de puntos de control inmunitarios.
El término "nucleótido" como se describe en el presente documento, se refiere a todos y cada uno de los nucleótidos o cualquier nucleótido o ARN natural o no natural adecuado o sustancia similar a un nucleótido o análogo con propiedades de emparejamiento de bases como se describe anteriormente

Como se usa en el presente documento, las expresiones "que comprende" y "que incluye" o variantes gramaticales de las mismas deben considerarse como que especifican al menos las características, integrantes, etapas o componentes citados, pero no excluye la presencia o adición de una o más características, integrantes, etapas, componentes o grupos de los mismos. Esta expresión abarca las expresiones "que consiste en" y "que consiste esencialmente en" que se entiende que especifican solo la característica, integrantes, etapas o componentes citados con exclusión de cualquier característica adicional.

Por lo tanto, las expresiones "que comprende"/"que incluye"/"que tiene" significan que cualquier componente adicional (o características, integrantes, etapas similares y similares) pueden estar presentes.

La expresión "que consiste en" significa que ningún componente adicional (o características, integrantes, etapas similares y similares) está presente.

La expresión "que consiste esencialmente en" o variantes gramaticales de la misma cuando se usa en el presente documento debe tomarse como una especificación de las características. integrantes, etapas o componentes citados, pero no excluye la presencia o adición de una o más características, integrantes, etapas, componentes o grupos adicionales de los mismos, pero solo si las características, integrantes, etapas, componentes o grupos adicionales de los mismos no alteran materialmente las características básicas y novedosas de la composición, dispositivo o método reivindicados.

Por lo tanto, la expresión "que consiste esencialmente en" significa aquellos componentes adicionales específicos (o características, integrantes, etapas similares y similares) pueden estar presentes, a saber, aquellos que no afecten materialmente a las características esenciales de la composición, el dispositivo o el método. En otras palabras, la expresión "que consiste esencialmente en" (que puede usarse indistintamente en el presente documento con la expresión "que comprende sustancialmente"), permite la presencia de otros componentes en la composición, el dispositivo o el método además de los componentes obligatorios (o características, integrantes, etapas similares y similares), siempre que las características esenciales del dispositivo o método no se vean afectadas materialmente por la presencia de otros componentes.

El término "método" se refiere a las maneras, medios, técnicas y procedimientos para realizar una tarea dada, incluyendo, pero no limitado a, aquellas maneras, medios, técnicas y procedimientos conocidos o desarrollados fácilmente a partir de las maneras, medios, técnicas y procedimientos conocidos por profesionales de las artes químicas, biológicas y biofísicas.

El término "aproximadamente" se refiere preferentemente a ± 10 % del valor numérico indicado, más preferentemente a ± 5 % del valor numérico indicado, y en particular al valor numérico exacto indicado.

Como se usa en el presente documento, el término "aproximadamente" se refiere a ± 10 % del valor numérico indicado y en particular a ± 5 % del valor numérico indicado. Siempre que se utilice el término "aproximadamente", se incluye también una referencia específica al valor numérico exacto indicado. Si el término "aproximadamente" se usa en relación con un parámetro que se cuantifica en números enteros, tales como el número de nucleótidos en un ácido nucleico dado, los números correspondientes a ± 10 % o ± 5 % del valor numérico indicado deben redondearse al número entero más próximo. Por ejemplo, la expresión "aproximadamente 25 nucleótidos" se refiere al intervalo de 23 a 28 nucleótidos, en particular el intervalo de 24 a 26 nucleótidos y preferentemente se refiere al valor específico de 25 nucleótidos.

A menos que se indique lo contrario, los métodos establecidos de tecnología de genes recombinantes se usaron como se describe, por ejemplo, en Sambrook, Russell "Molecular Cloning, A Laboratory Manual", Cold Spring Harbor Laboratory, N.Y. (2001)). La presente invención se describe adicionalmente con referencia a las siguientes figuras y ejemplos no limitantes.

Breve descripción de las figuras

Figura 1: (A) Ilustración esquemática de fragmentos de ADN cortados marcados con secuencias de nucleótidos predefinidas bicatenarias ("códigos de barras") (aquí representadas con α y β). Los números 1 y 2 mostrados en gris oscuro y claro, respectivamente, representan las colas compatibles con la célula de flujo de Illumina. (B) Dos moléculas de ADN bicatenarias diferentes resultan de la amplificación por PCR de los fragmentos de ADN con código de barras que se muestran en (A). Aquellos resultantes de la amplificación de la cadena superior tienen el código de barras α cerca de la secuencia de la célula de flujo número 1 y el código de barras β cerca de la secuencia de la célula de flujo número 2 ("familias αβ"). Los duplicados de PCR resultantes de la amplificación de la cadena inferior tienen las dos cadenas complementarias de ADN marcadas recíprocamente ("familias βα"). Adaptado de Kennedy SR et al. Nature Protocols (2014).

Figura 2: Ilustración esquemática del análisis computacional realizado para generar lecturas consenso monocatenarias sintéticas (sscs). Las mutaciones/variaciones se resaltan con círculos. A continuación se presentan tres escenarios posibles diferentes. Las lecturas que pertenecen a una familia se comparan en cada posición genómica y se escribe un nucleótido en la lectura de sscs solo cuando al menos el 76 % (4 lecturas de 5) de los miembros de la familia muestran la misma base en la posición investigada. Si no se alcanza el consenso, la posición evaluada en la lectura sscs se rellena con una "N". Adaptado de Schmitt MW et al. Proc Natl Acad Sci U S A. (2012).

Figura 3: Ilustración esquemática del análisis computacional realizado para generar lecturas consenso bicatenarias sintéticas (dscs). Las mutaciones/variaciones se resaltan con círculos. Se comparan lecturas scsc alineadas en la misma ubicación genómica y caracterizadas por códigos de barras recíprocos (familias αβ y βα). Un nucleótido se escribe en una lectura de consenso bicatenaria (dscs) solo si está presente en ambas lecturas de sscs, de lo contrario la posición evaluada en la lectura dscs se rellena con una "N". Adaptado de Schmitt MW et al. Proc Natl Acad Sci U S A. (2012).

Figura 4: Ejemplo para una posición dada en la cual se detecta un nucleótido diferente del nucleótido anotado en el genoma de referencia, si en esa posición hay más del doble de "N" presentes (en comparación con la variante), no se llama mutación.

Figura 5: Ejemplo para una posición dada en la cual se detecta un nucleótido diferente del nucleótido anotado en el genoma de referencia, si en esa posición hay más de cuatro veces de "N" presentes (en comparación con la variante), no se llama mutación.

Descripción de realizaciones de ejemplo

La Figura 2 ilustra un método de acuerdo con una realización de ejemplo de la presente invención que muestra una canalización computacional desarrollada específicamente para el análisis de datos de secuenciación sin procesar con código de barras.

En el método de la invención como se muestra en la Figura 2, solo se crea la información de una sola cadena de una secuencia de ácido nucleico, un llamado consenso monocatenario, en lo sucesivo en el presente documento denominado sscs. La primera etapa del flujo computacional del método como se presenta en la Figura 2 es alinear los datos de secuenciación sin procesar con el genoma de referencia humano. En otras palabras, la primera etapa del flujo computacional del método es alinear una pluralidad de lecturas genéticas con al menos una secuencia genética de referencia. Durante este proceso, los primeros nucleótidos de cada lectura, que representa el código de barras predefinido, se retiran y se colocan en el nombre de lectura para que estén disponibles para la siguiente etapa del análisis computacional.

Después de la alineación, los pares de lectura que comparten la misma posición genómica y los códigos de barras se agrupan en "familias" como se muestra en la Figura 2. En otras palabras, la segunda etapa se refiere a las lecturas genéticas que pertenecen a la misma secuencia genética de referencia en un grupo correspondiente.

El objeto es generar una lectura consenso sintética monocatenaria (sscs) usando la información contenida en todos los miembros de la familia, es decir, considerando verdadera la información de secuenciación más abundante. Por lo tanto, las lecturas que pertenecen a la misma familia se comparan nucleótido por nucleótido y el nucleótido más abundante *man_p* se determina en cada posición.
Un nucleótido más abundante *man_p* respectivo se escribe en la lectura sintética sscs en la posición respectiva solo si el nucleótido aparece, por ejemplo, en al menos el 76 %, por ejemplo > 3 de cada 4, de los miembros de la familia en la posición genómica investigada. En otras palabras, el nucleótido más abundante *man_p* se determina en cada posición específica p de una pluralidad de posiciones dentro de la familia respectiva. Una representación del nucleótido más abundante *man_p* en cada posición específica p de la secuencia consenso solo se establece si la relación r entre el número de nucleótidos de la familia en la posición específica es el nucleótido más abundante *man_p* y el número total de lecturas de la familia, es decir, el número de miembros de la familia, está por encima o es igual al umbral t predeterminado, es decir, el 76 % en la presente realización inventiva. De otro modo, se establece una etiqueta N, es decir, si la relación r entre el número de nucleótidos es el nucleótido más abundante *man_p* en la posición específica p y el número de lecturas está por debajo del 76 %.

La parte superior de la Figura 2A muestra la situación en la familia αβ donde solo una lectura contiene una variación/mutación (mostrada como un círculo) en una posición investigada.
Dicha variación solo está contenida en una de las lecturas genéticas. Por lo tanto, este nucleótido no es el nucleótido más abundante y el umbral predeterminado, por ejemplo al menos el 76 %, no se cumple. Por lo tanto, el nucleótido de tipo silvestre más abundante (al menos el 76 %) comprendido en las otras lecturas genéticas en esta posición se establece en la secuencia consenso.
En la parte inferior de la Figura 2A, se muestra la familia βα respectiva. Los miembros de la familia βα no muestran la variación que se encuentra en la familia superior.

La Figura 2B muestra el caso donde para la familia αβ puede encontrarse un consenso para una posición de investigación como la variación, que es el nucleótido más abundante *man_p,* se produce en cada lectura en la familia. Por lo tanto, una representación del nucleótido más abundante *man_p,* que es la variante, se establece en la posición de investigación. Por el contrario, la respectiva familia βα no muestra dicho nucleótido más abundante *man_p* para la familia αβ en absoluto.
La Figura 2C muestra el caso en el cual para la familia αβ puede encontrarse un consenso en una determinada posición

de investigación, pero otras posiciones no cumplen el requisito de que el nucleótido respectivo tiene que estar presente en la posición de investigación en, por ejemplo, el 76 % de los miembros de la familia. Por lo tanto, se dice una etiqueta "N" para los casos donde ninguno de los nucleótidos respectivos está presente en la posición de investigación respectiva en ≥ 76 % de los miembros de la familia.

La Figura 3 ilustra un método de acuerdo con una realización de ejemplo preferida de la presente invención.

Después de que se haya determinado la secuencia consenso de una sola cadena (sscs) como se muestra en la Figura 2, las lecturas sscs se alinean en la misma ubicación genómica y se caracterizan por códigos de barras recíprocos, es decir, las familias αβ y βα, se comparan.

Un nucleótido se escribe en una lectura consenso bicatenaria solo si este nucleótido es el nucleótido más abundante *man_p* en ambas lecturas sscs, de lo contrario la posición evaluada en la lectura sscs se establece con una etiqueta "N".

La Figura 3A muestra la situación donde las sscs de ambas cadenas, es decir, la familia αβ y la familia βα, están de acuerdo entre sí. Por lo tanto, la secuencia dscs es idéntica a la secuencia de ambas lecturas sscs y no se establece ninguna representación/etiqueta en el consenso bicatenario.

La Figura 3B muestra la situación donde un determinado nucleótido más abundante *man_p* puede encontrarse solo en una de las familias, es decir, la familia αβ, pero no en la familia βα. Por lo tanto, como este nucleótido solo está presente en una de las familias pero no en la familia recíproca, la etiqueta "N" se establece en el consenso bicatenario.

En la Figura 3C, un determinado nucleótido más abundante *man_p* se encuentra para la familia αβ en una determinada posición p, pero también para la familia βα recíproca en la misma posición *p* respectiva. Por lo tanto, una representación de dicho nucleótido más abundante *man_p* se escribe en el consenso bicatenario (círculo). Por el contrario, en otras dos posiciones distintas, se estableció una etiqueta "N" en la familia αβ o en la familia βα. Como la etiqueta "N" solo se estableció en una de las familias recíprocas, se establece una etiqueta "N" respectiva en el consenso bicatenario en la posición respectiva.

Ejemplos

Para evaluar la sensibilidad y especificidad (PPV) del método de acuerdo con la presente invención en la detección de variantes con MAF bajo (frecuencia de alelos menores ≤ 1 %), se generan cuatro diluciones de tres líneas celulares normales HapMap. Los detalles pueden tomarse de la Tabla 1.

El ADN de estas cuatro diluciones se analiza usando el método de crear un consenso monocatenario como se describe anteriormente.

Tabla 1: Detalles de las diluciones de las líneas celulares normales de HapMap generadas en el laboratorio para evaluar la sensibilidad y la especificidad del enfoque de los presentes inventores en la detección de variantes con MAF bajo

Nombre normal de HapMap	Dilución 1	Dilución 2	Dilución 3	Dilución 4
GM19194B (%)	99,4	99,4	99,4	99,4
GM19153B (%)	0,2	0,4	nd	nd
GM12144C (%)	0,4	0,2	nd	nd
GM19137B (%)	nd	nd	0,2	0,4
GM19142B (%)	nd	nd	0,4	0,2

Un conjunto de polimorfismos de un solo nucleótido, SNP específicos para las líneas celulares individuales de HapMap en uso, se determinan a partir de cada línea celular no diluida. Estos se subdividieron en un conjunto de SNP únicos, en lo sucesivo en el presente documento denominados "SNP privados", que son específicos de una de las líneas celulares HapMap en uso. Los SNP presentes en más de una de las líneas celulares en uso se denominan SNP no privados. La suma de los SNP privados y no privados se denomina SNP totales.

Las diluciones de la línea celular HapMap se usan para determinar el límite de detección del ensayo de los presentes inventores. Debido a que solo se consideraron SNP privados heterocigotos, todas las diluciones se caracterizaron por SNP privados con un 0,1 %=<MAF<0,2 % esperado y MAF ~ 50 % como puede verse en la tabla 1.

Debido a un error experimental, el MAF detectado es en algunos casos diferente al esperado. Teniendo en cuenta este error experimental intrínseco, la sensibilidad y la especificidad (PPV) se calcularon de la siguiente manera:

Sensibilidad = SNP privados verdaderos positivos/(SNP privados verdaderos positivos + falso negativo)

Especificidad (PPV) = Total verdadero positivo/(Total verdadero positivo + Falso positivo)

Donde verdadero positivo, falso negativo y falso positivo se definen como:

- SNP privados verdaderos positivos: SNP privados con 0,1 % =< MAF < 0,2 % detectado
- SNP no privados verdaderos positivos: SNP no privados con 0,1 % =< MAF < 0,2 % detectado
- Verdadero positivo total: SNP privados verdaderos positivos + SNP verdaderos no privados positivos
- Falso negativo: SNP privados caracterizados por un 0,1 % =< MAF <0,2 % detectado en los datos brutos de lectura de secuenciación alineada pero no llamada por el algoritmo de análisis de los presentes inventores.
- Falso positivo: SNV con 0,1 % =< MAF < 0,2 % detectado clasificado como "mutaciones verdaderas" por el algoritmo de análisis de los presentes inventores pero no presente en ninguna de las líneas celulares normales HapMap no diluidas.

En la Tabla 2 los detalles del número de verdaderos positivos, falsos negativos y falsos positivos con 0,1 % =< MAF < 0,2 % detectado se informan para las cuatro diluciones de líneas celulares normales HapMap presentadas en la Tabla 1 con una cobertura de secuenciación media de 3000x.

Tabla 2: Número de SNP verdaderos positivos, falsos negativos, falsos positivos detectados usando el método presentado en este documento.

	Cobertura promedio	**Número de SNP privados falsos negativos**	**Número de variantes falsos positivos**	**Número de SNP privados verdaderos positivos**	**SNP verdaderos positivos totales**
Dilución 1	2943,56	6	16	25	91
Dilución 2	3042,1	2	12	27	49
Dilución 3	2995,51	4	18	30	58
Dilución 4	2934,2	4	25	34	66
	total	22	172	141	326

En la presente invención se desarrolla un filtro que usa el nivel de fiabilidad para cada llamada de nucleótido, aprovecha de esta manera la naturaleza de las lecturas sscs y/o dscs. Como se ha mencionado anteriormente, las lecturas sscs están compuestas únicamente por los nucleótidos presentes en al menos el 76 % de las lecturas de una familia en la posición investigada.

Si no se cumple esta condición, se coloca una "N" en la posición investigada en la lectura sscs para indicar que no se alcanzó el consenso. El experto en la materia entenderá que la presente invención también funciona solo en el caso en el que se considere solo una sscs de consenso monocatenario.

Sin embargo, con el fin de mejorar aún más el método de la presente invención, también se consideró el consenso bicatenario. Por lo tanto, de manera similar, se escribe un nucleótido en la lectura de dscs sintéticos solo si está presente en ambas lecturas de sscs alineándose en la misma posición genómica y mostrando códigos de barras complementarios, se coloca una "N" en la posición investigada en la lectura dscs para indicar que no se alcanzó el consenso. Por lo tanto, las posiciones con "N" representan regiones para las que no se alcanzó un consenso y se desconoce la verdadera naturaleza de la secuencia.

Se observó que las lecturas sscs y dscs que se alinean en regiones genómicas que contienen sustituciones introducidas erróneamente durante el flujo de trabajo de secuenciación exhiben una tasa más baja de nucleótidos consenso y, por lo tanto, un mayor número de N en comparación con las regiones que contienen variantes verdaderas. Estas regiones pueden definirse mediante repeticiones de secuencia, que con frecuencia conducen a la incorporación de nucleótidos incorrectos y, por lo tanto, no son eliminados por el límite de consenso del 76 % aplicado en la etapa anterior.

Basándose en esta información, la abundancia de "N" en una posición variante en todas las lecturas que cubren la posición respectiva (que representa la profundidad de secuenciación en esta posición) puede usarse para discriminar

una mutación verdadera de una sustitución errónea. En particular, el enfoque que se ha implementado de acuerdo con la presente invención emplea la relación entre el número de lecturas que contienen "N" y las que contienen la variante para discernir una llamada verdadera de las llamadas positivas falsas (en lo sucesivo en el presente documento, "filtro N").

Ejemplo 1:

Para evaluar la validez de la presente invención, los datos obtenidos de la dilución de las líneas celulares normales de HapMap analizadas previamente con la canalización computacional como se usó antes de la invención se procesaron adicionalmente usando el filtro N con una proporción requerida de n.º (número) de lecturas que contienen una "N" en una posición definida dividida por el n.º (número) de lecturas que contienen la variante en la posición definida es >2. Por lo tanto, si en una posición definida hay más del doble de "N" presentes (en comparación con la variante), no se llama mutación; véase también la Figura 4.

El número de llamadas verdaderas positivas, falsas negativas y falsas positivas detectadas al 0,1 % =< MAF < 0,2 % usando este análisis mejorado adicional se presentan en la Tabla 3.

Tabla 3: Número de SNP verdaderos positivos, falsos negativos, falsos positivos detectados usando la canalización computacional optimizada que incluye el filtro N.

	Cobertura promedio	**Número de SNP privados falsos negativos**	**Número de variantes falsos positivos**	**Número de SNP privados verdaderos positivos**	**Total de SNP positivos verdaderos (privados + no privados)**
Dilución 1	2943,56	10	9	21	42
Dilución 2	3042,1	5	9	24	35
Dilución 3	2995,51	4	10	30	54
Dilución 4	2934,2	4	18	34	60
	total	23	46	109	191

La sensibilidad y especificidad (PPV) estimadas para el flujo de trabajo presentado en los detalles anteriores fueron 82,5 % y 80 %, respectivamente con y sin el filtro N con una relación > 2 (véase la Tabla 5).

Ejemplo 2:

Para evaluar la validez de la presente invención, los datos obtenidos de la dilución de las líneas celulares normales de HapMap analizadas previamente con la canalización computacional como se usó antes de la invención se procesaron adicionalmente usando el filtro N con una proporción requerida de n.º (número) de lecturas que contienen una "N" en una posición definida dividida por el n.º (número) de lecturas que contienen la variante en la posición definida es >4. Por lo tanto, si en una posición definida hay más de cuatro veces de "N" presentes (en comparación con la variante), no se llama mutación; véase la Figura 5.

El número de llamadas verdaderas positivas, falsas negativas y falsas positivas detectadas al 0,1 % =< MAF < 0,2 % usando este análisis mejorado adicional se presentan en la Tabla 3.

Tabla 4: Número de SNP verdaderos positivos, falsos negativos, falsos positivos detectados usando la canalización computacional optimizada que incluye el filtro N.

	Cobertura promedio	**Número de SNP privados falsos negativos**	**Número de variantes falsos positivos**	**Número de SNP privados verdaderos positivos**	**Total de SNP positivos verdaderos (privados + no privados)**
Dilución 1	2943,56	6	15	25	47
Dilución 2	3042,1	2	11	27	44
Dilución 3	2995,51	4	14	30	54

(continuación)

	Cobertura promedio	**Número de SNP privados falsos negativos**	**Número de variantes falsos positivos**	**Número de SNP privados verdaderos positivos**	**Total de SNP positivos verdaderos (privados + no privados)**
Dilución 4	2934,2	4	20	34	60
	total	16	60	116	205

La sensibilidad y especificidad (PPV) estimadas para el flujo de trabajo presentado en los detalles anteriores fueron 82,5 % y 87,9 %, respectivamente con y sin el filtro N con una relación > 4 (véase la Tabla 5).

Tabla 5: Sensibilidad (PPV) obtenida para las 4 diluciones de líneas celulares normales HapMap analizadas usando la canalización computacional con dos proporciones de filtro N diferentes, así como sin el filtro N

	Sensibilidad	Especificidad (PPV)
sin filtro N	86,5 %	65,5 %
con filtro N > 4	87,9 %	77,4 %
con filtro N > 2	82,5 %	80,5 %

Por lo tanto, los presentes ejemplos muestran que el uso del filtro N descrito anteriormente muestra una reducción significativa de falsos positivos.

Como la presente invención puede realizarse de varias formas sin apartarse del alcance o características esenciales de la misma, debe entenderse que las realizaciones descritas anteriormente no están limitadas por ninguno de los detalles de las descripciones anteriores, a menos que se especifique de otro modo, sino que debe interpretarse ampliamente dentro del alcance como se define en las reivindicaciones adjuntas y, por lo tanto, todos los cambios y modificaciones que caen dentro de la presente invención pretenden estar incluidos en las reivindicaciones adjuntas.

Referencias

[1] - Schmitt MW et al. Proc Natl Acad Sci U S A. (2012) y Kennedy SR et al. Nature Protocols (2014)

[2]-http://web.archive.org/web/20151107034222/http:/www.illumina.
com/documents/products/techspotlights/techspotlight_sequencin g.pdf

REIVINDICACIONES

1. Método de secuenciación de ácidos nucleicos que comprende las siguientes etapas:

(a) obtener una pluralidad de lecturas genéticas mediante la secuenciación de una muestra de ácido nucleico;
(b) alinear la pluralidad de lecturas genéticas con al menos una secuencia genética de referencia;
(c) agrupar las lecturas genéticas que comparten una posición genética en una secuencia genética de referencia de la al menos una secuencia genética de referencia en una pluralidad de grupos;
(d) crear una secuencia consenso para cada grupo de la pluralidad de grupos, en donde se crea una secuencia consenso correspondiente determinando un nucleótido más abundante *man_p* en cada posición específica p de una pluralidad de posiciones dentro del grupo correspondiente de lecturas genéticas y

(i) establecer una representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación r entre el número de lecturas genéticas dentro del grupo correspondiente que tiene el nucleótido más abundante *man_p* en la posición específica *p* y el número de lecturas genéticas dentro del un grupo correspondiente es superior o igual a un umbral predeterminado *t;* y
(ii) establecer una etiqueta *N* si la relación *r* está por debajo del umbral predeterminado *t*, en donde dicho umbral *t* es al menos aproximadamente el 76 %;

(e) comparar las secuencias consenso de la pluralidad de grupos con la secuencia genética de referencia en cada posición específica *p* de una pluralidad de posiciones de las secuencias consenso y en donde una diferencia en una posición específica entre las secuencias consenso y la secuencia genética de referencia indica una variación genética en la posición específica;
(f) determinar el número de secuencias consenso que comprenden la variación en cada posición específica *p* de una pluralidad de posiciones y determinar el número de secuencias consenso que comprenden la etiqueta *N* en cada posición específica *p* de una pluralidad de posiciones; y
(g) identificar la variación genética en cada posición específica *p* de una pluralidad de posiciones como una verdadera variación genética si una relación *r** entre el número de secuencias consenso que comprenden la etiqueta *N* en la posición específica *p* y el número de secuencias consenso que comprenden la variación genética en la posición específica *p* está por debajo de un umbral *t**, en donde dicho umbral *t** es igual o por debajo de 1,8.

2. Método de acuerdo con la reivindicación 1, en donde en la etapa (c) cada lectura genética en un grupo correspondiente de la pluralidad de grupos comprende al menos una secuencia de ácido nucleico particular, tal como al menos una secuencia de código de barras.

3. Método de acuerdo con la reivindicación 2, en donde cada secuencia de ácido nucleico particular, tal como cada secuencia de código de barras, corresponde a una molécula respectiva.

4. Método de acuerdo con una cualquiera de las reivindicaciones anteriores, en donde las lecturas genéticas de la etapa (c) se agrupan en función de su posición genética y su secuencia de código de barras.

5. Método de acuerdo con una cualquiera de las reivindicaciones anteriores, en donde en la etapa (d) un grupo correspondiente de la pluralidad de grupos comparte al menos una secuencia de ácido nucleico particular, tal como al menos una secuencia de código de barras.

6. Método de acuerdo con una cualquiera de las reivindicaciones anteriores, en donde la etapa (d) se realiza para todas las posiciones respectivas dentro del grupo, en donde la etapa (e) se realiza para todas las posiciones respectivas dentro del grupo, en donde la etapa (f) se realiza para todas las posiciones respectivas dentro del grupo y/o en donde la etapa (g) se realiza para todas las posiciones respectivas dentro del grupo.

7. Método de acuerdo con cualquiera de las reivindicaciones anteriores, en donde el número de posiciones en las lecturas genéticas es 72.

8. Método de acuerdo con una cualquiera de las reivindicaciones anteriores, en donde un grupo correspondiente comprende al menos 3 lecturas genéticas.

9. Método de acuerdo con una cualquiera de las reivindicaciones anteriores, en donde la pluralidad de grupos es al menos dos grupos.

10. Método de acuerdo con una cualquiera de las reivindicaciones anteriores, en donde la pluralidad de grupos comprende un grupo' y un grupo" y en donde las lecturas genéticas del grupo" se superponen al menos parcialmente con las lecturas genéticas del grupo'.

11. Método de acuerdo con una cualquiera de las reivindicaciones 1 a 10, en donde la pluralidad de grupos comprende

un grupo' y un grupo" y en donde las lecturas genéticas del grupo" no se superponen con las lecturas genéticas del grupo'.

12. Método de acuerdo con una cualquiera de las reivindicaciones 1 a 10, en donde la pluralidad de grupos comprende un grupo' y un grupo" y en donde las lecturas genéticas del grupo" se superponen completamente con las lecturas genéticas del grupo'.

13. Método de acuerdo con una cualquiera de las reivindicaciones 1 a 10, en donde la pluralidad de grupos comprende un grupo' y un grupo" y en donde las lecturas genéticas del grupo" corresponden al complemento inverso de las lecturas genéticas del grupo'.

14. Método de acuerdo con la reivindicación 13, en donde las lecturas genéticas del grupo' corresponden a una primera cadena de un ácido nucleico bicatenario y las lecturas genéticas del grupo" corresponden a la segunda cadena complementaria de dicho ácido nucleico bicatenario.

15. Método de acuerdo con la reivindicación 13 o 14, en donde se crea una secuencia consenso monocatenaria para el grupo' y en donde se crea una secuencia consenso monocatenaria para el grupo".

16. Método de acuerdo con una cualquiera de las reivindicaciones anteriores, que comprende además: crear una secuencia consenso bicatenaria al

(i) establecer una representación del nucleótido más abundante *man_p* o la etiqueta "N" en cada posición específica *p* de una pluralidad de posiciones en la secuencia consenso bicatenaria si la representación o la etiqueta en la posición específica *p* está presente respectivamente en ambas secuencias consenso monocatenarias del grupo' y del grupo"; y
(ii) establecer la etiqueta "N" en cada posición específica *p* de una pluralidad de posiciones en la secuencia consenso bicatenaria si la etiqueta "N" está presente en la posición específica *p* en una de las secuencias consenso monocatenarias del grupo' o del grupo" o si la representación del nucleótido más abundante *man_p* no es idéntica en la posición específica en ambas secuencias consenso monocatenarias del grupo' o del grupo".

17. Método de acuerdo con la reivindicación 16, en donde en la etapa (e) de la reivindicación 1 se comparan las secuencias consenso bicatenarias.

18. Método de acuerdo con la reivindicación 17, en donde las etapas (e), (f) y (g) de la reivindicación 1 se realizan con las secuencias consenso bicatenarias.

19. Método de acuerdo con una cualquiera de las reivindicaciones 14 a 18, en donde cada posición corresponde a un par de bases.

20. Método de acuerdo con cualquiera de las reivindicaciones anteriores, en donde las lecturas genéticas del grupo correspondiente tienen la misma longitud.

21. Método de acuerdo con una cualquiera de las reivindicaciones anteriores, en donde la secuenciación es secuenciación de última generación.

22. Sistema de secuenciación de ácidos nucleicos que comprende:

(a) una unidad de obtención estando configurada para obtener una pluralidad de lecturas genéticas mediante la secuenciación de una muestra de ácido nucleico; y
(b) una unidad de cálculo estando configurada para alinear la pluralidad de lecturas genéticas con al menos una secuencia genética de referencia;
(c) la unidad de cálculo estando configurada para agrupar las lecturas genéticas que comparten una posición genética en una secuencia genética de referencia de la al menos una secuencia genética de referencia en una pluralidad de grupos;
(d) estando la unidad de cálculo configurada para crear una secuencia consenso para cada grupo de la pluralidad de grupos, en donde se crea una secuencia consenso correspondiente determinando un nucleótido más abundante *man_p* en cada posición específica *p* de una pluralidad de posiciones dentro del grupo correspondiente de lecturas genéticas y

(i) establecer una representación del nucleótido más abundante *man_p* en cada posición específica *p* de la secuencia consenso si una relación r entre el número de lecturas genéticas dentro del grupo correspondiente que tiene el nucleótido más abundante *man_p* en la posición específica *p* y el número de lecturas genéticas dentro del un grupo correspondiente es superior o igual a un umbral predeterminado *t;* y
(ii) establecer una etiqueta *N* si la relación está por debajo del umbral predeterminado *t,* en donde dicho umbral

t es al menos aproximadamente el 76 %;

(e) estando la unidad de cálculo configurada para comparar las secuencias consenso de la pluralidad de grupos con la secuencia genética de referencia en cada posición específica *p* de una pluralidad de posiciones de las secuencias consenso y en donde una diferencia en una posición específica entre las secuencias consenso y la secuencia genética de referencia indica una variación en la posición específica;
(f) estando la unidad de cálculo configurada para determinar el número de secuencias consenso que comprenden la variación en cada posición específica *p* de una pluralidad de posiciones y determinar el número de secuencias consenso que comprenden la etiqueta *N* en cada posición específica *p* de una pluralidad de posiciones; y
(g) estando la unidad de cálculo configurada para identificar la variación en cada posición específica *p* de una pluralidad de posiciones como una verdadera variación si una relación *r** entre el número de secuencias consenso que comprenden la etiqueta *N* en la posición específica *p* y el número de secuencias consenso que comprenden la variación en la posición específica *p* está por debajo de un umbral *t*,* en donde dicho umbral *t** es igual o por debajo de 1,8.

23. El sistema de acuerdo con la reivindicación 22, en donde en la etapa (c) cada lectura genética en un grupo correspondiente de la pluralidad de grupos comprende al menos una secuencia de ácido nucleico particular, tal como al menos una secuencia de código de barras.

24. El sistema de acuerdo con una cualquiera de las reivindicaciones 22 a 23, en donde cada secuencia de ácido nucleico particular, tal como cada secuencia de código de barras, corresponde a una molécula respectiva.

25. El sistema de acuerdo con una cualquiera de las reivindicaciones 22 a 24, en donde las lecturas genéticas de la etapa (c) se agrupan en función de su posición genética y su secuencia de código de barras.

26. El sistema de acuerdo con una cualquiera de las reivindicaciones 22 a 23, en donde en la etapa (d) un grupo correspondiente de la pluralidad de grupos comparte al menos una secuencia de ácido nucleico particular, tal como al menos una secuencia de código de barras.

27. El sistema de acuerdo con una cualquiera de las reivindicaciones 22 a 26, en donde la unidad de cálculo estando configurada para crear las secuencias consenso y está configurada para establecer una representación respectiva o una etiqueta N respectiva para todas las posiciones respectivas dentro del un grupo correspondiente, en donde la unidad de cálculo estando configurada para comparar las secuencias consenso con la secuencia genética de referencia en todas las posiciones, en donde la unidad de cálculo estando configurada para determinar el número de secuencias consenso que comprenden la variación y para determinar el número de secuencias consenso que comprenden la etiqueta *norte* para todas las posiciones respectivas de las secuencias consenso, y/o en donde la unidad de cálculo estando configurada para identificar la variación en todas las posiciones y está configurada para establecer una representación respectiva o una etiqueta N respectiva para todas las posiciones respectivas de las secuencias consenso.

28. El sistema de acuerdo con una cualquiera de las reivindicaciones 22 a 27, en donde el número de posiciones en las lecturas genéticas es 72.

29. El sistema de acuerdo con una cualquiera de las reivindicaciones 22 a 28, en donde un grupo correspondiente comprende al menos 3 lecturas genéticas.

30. El sistema de acuerdo con una cualquiera de las reivindicaciones 22 a 29, en donde la pluralidad de grupos es al menos dos grupos.

31. El sistema de acuerdo con una cualquiera de las reivindicaciones 22 a 30, en donde la pluralidad de grupos comprende un grupo' y un grupo'' y en donde las lecturas genéticas del grupo'' se superponen al menos parcialmente con las lecturas genéticas del grupo'.

32. El sistema de acuerdo con una cualquiera de las reivindicaciones 22 a 30, en donde la pluralidad de grupos comprende un grupo' y un grupo'' y en donde las lecturas genéticas del grupo'' no se superponen con las lecturas genéticas del grupo'.

33. El sistema de acuerdo con una cualquiera de las reivindicaciones 22 a 30, en donde la pluralidad de grupos comprende un grupo' y un grupo'' y en donde las lecturas genéticas del grupo'' se superponen completamente con las lecturas genéticas del grupo'.

34. El sistema de acuerdo con una cualquiera de las reivindicaciones 22 a 30, en donde la pluralidad de grupos comprende un grupo' y un grupo'' y en donde las lecturas genéticas del grupo'' corresponden al complemento inverso de las lecturas genéticas del grupo'.

35. El sistema de acuerdo con la reivindicación 34, en donde las lecturas genéticas del grupo' corresponden a una primera cadena de un ácido nucleico bicatenario y las lecturas genéticas del grupo" corresponden a la segunda cadena complementaria de dicho ácido nucleico bicatenario.

36. El sistema de acuerdo con la reivindicación 34 o 35, en donde se crea una secuencia consenso monocatenaria para el grupo' y en donde se crea una secuencia consenso monocatenaria para el grupo".

37. El sistema de acuerdo con una cualquiera de las reivindicaciones 22 a 36, en donde la unidad de cálculo estando configurada para crear una secuencia consenso bicatenaria al

(i) establecer una representación del nucleótido más abundante *man_p* o la etiqueta "N" en cada posición específica p de una pluralidad de posiciones en la secuencia consenso bicatenaria si la representación o la etiqueta en la posición específica *p* está presente respectivamente en ambas secuencias consenso monocatenarias del grupo' y del grupo"; y
(ii) establecer la etiqueta "N" en cada posición específica *p* de una pluralidad de posiciones en la secuencia consenso bicatenaria si la etiqueta "N" está presente en la posición específica *p* en una de las secuencias consenso monocatenarias del grupo' o del grupo" o si la representación del nucleótido más abundante *man_p* no es idéntica en la posición específica en ambas secuencias consenso monocatenarias del grupo' o del grupo".

38. El sistema de acuerdo con la reivindicación 37, en donde la unidad de cálculo estando configurada para comparar secuencias consenso bicatenarias.

39. El sistema de acuerdo con la reivindicación 37 o 38, en donde la unidad de cálculo estando configurada para comparar las secuencias consenso bicatenarias, para determinar el número de secuencias consenso bicatenarias y para identificar la variación en las secuencias consenso bicatenarias.

40. El sistema de acuerdo con una cualquiera de las reivindicaciones 22 a 39, en donde cada posición corresponde a un par de bases.

41. El sistema de acuerdo con una cualquiera de las reivindicaciones 22 a 40, en donde las lecturas genéticas del grupo correspondiente tienen la misma longitud.

42. El sistema de acuerdo con una cualquiera de las reivindicaciones 22 a 41, en donde la secuenciación es secuenciación de última generación.

43. Producto de programa informático que comprende uno o más medios legibles por ordenador que tienen instrucciones ejecutables por ordenador para realizar las etapas del método de una cualquiera de las reivindicaciones 1 a 21.

FIG 1

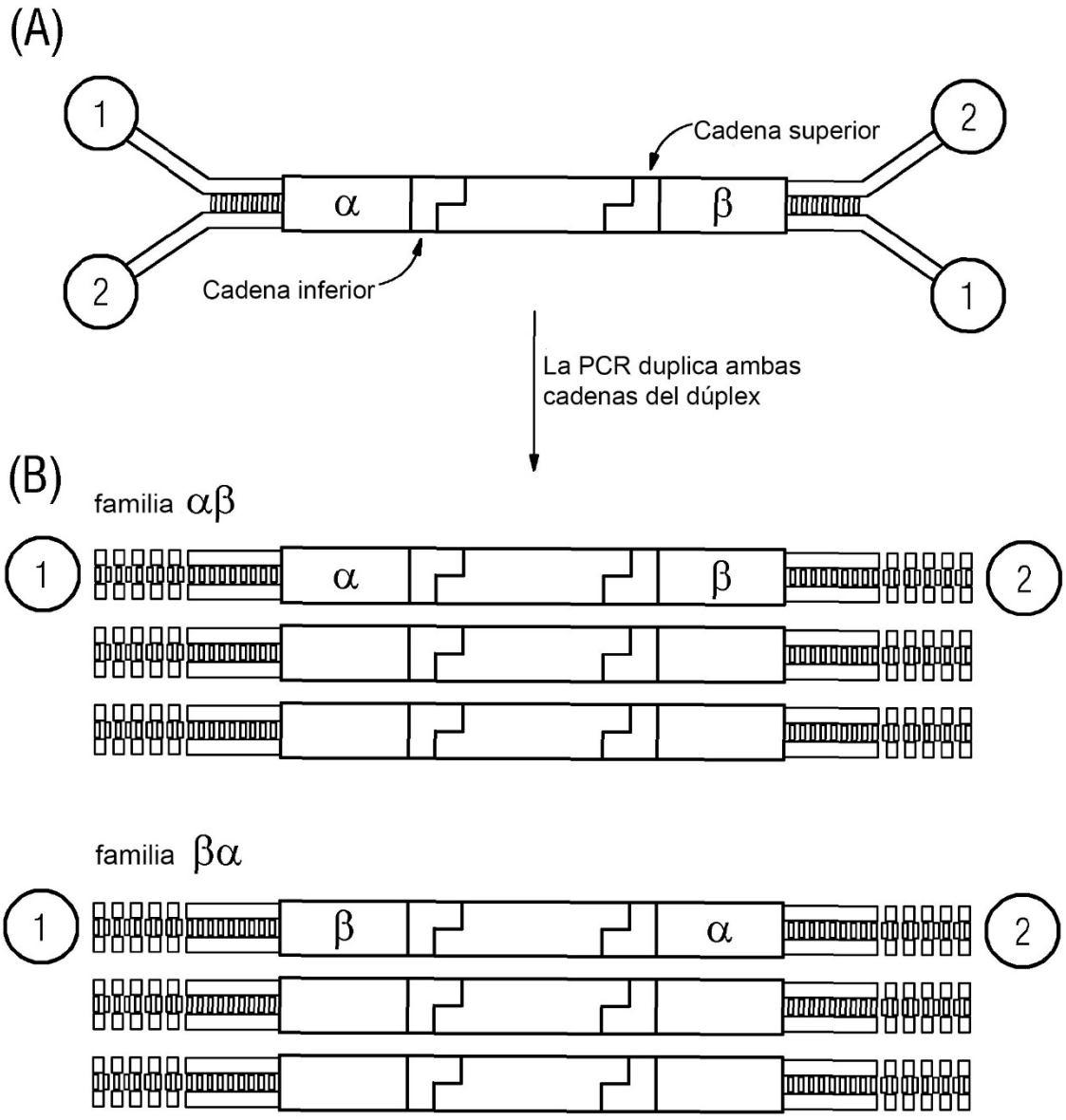
(A)
1
2
Cadena superior
α
β
Cadena inferior
2
1
La PCR duplica ambas cadenas del dúplex
(B)
familia αβ
1
α
β
2
familia βα
1
β
α
2

FIG 2

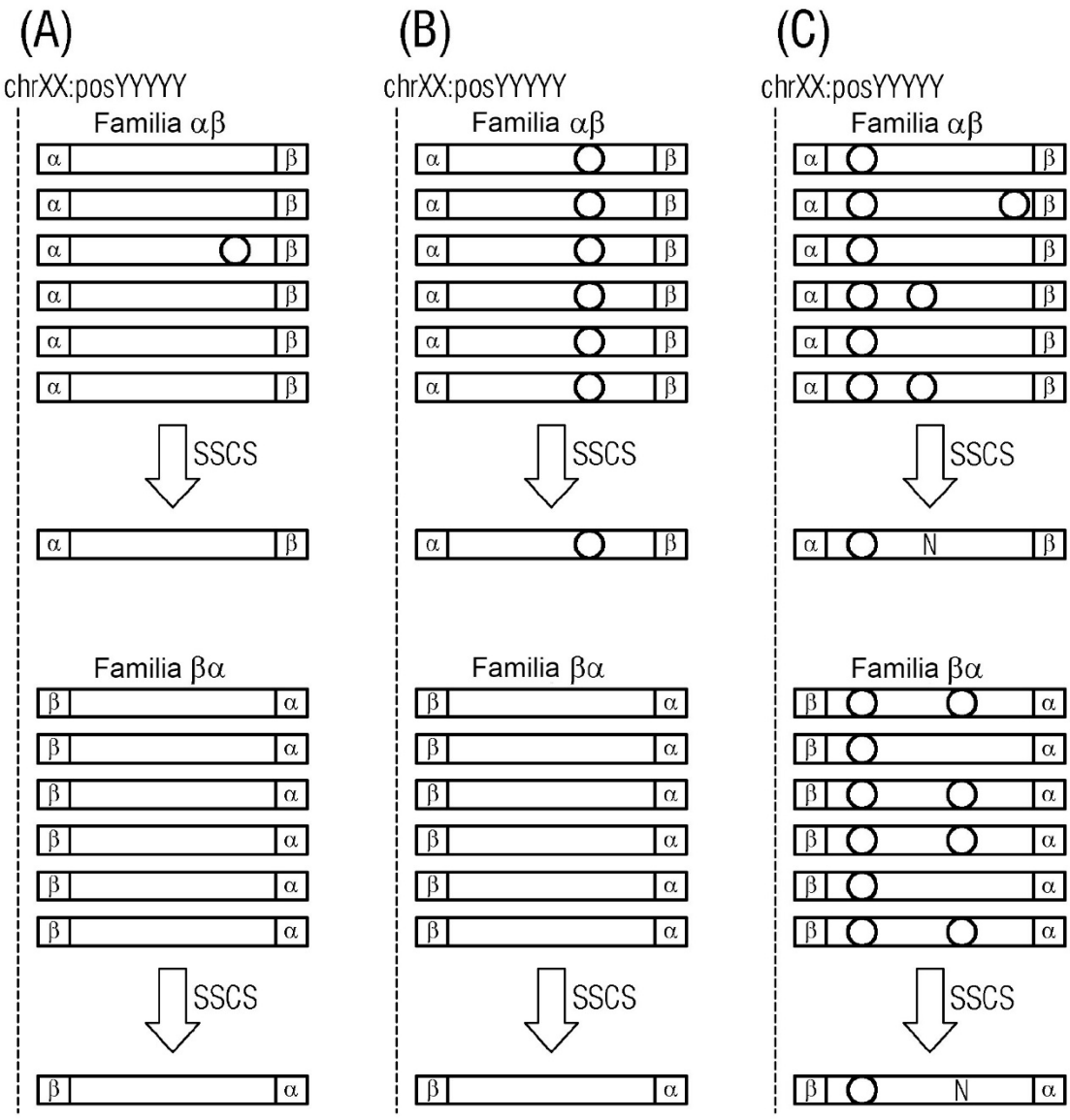
(A)
chrXX:posYYYYY
Familia αβ
α
β
SSCS
Familia βα
(B)
chrXX:posYYYYY
Familia αβ
SSCS
Familia βα
(C)
chrXX:posYYYYY
Familia αβ
SSCS
N
Familia βα
SSCS
N

FIG 3

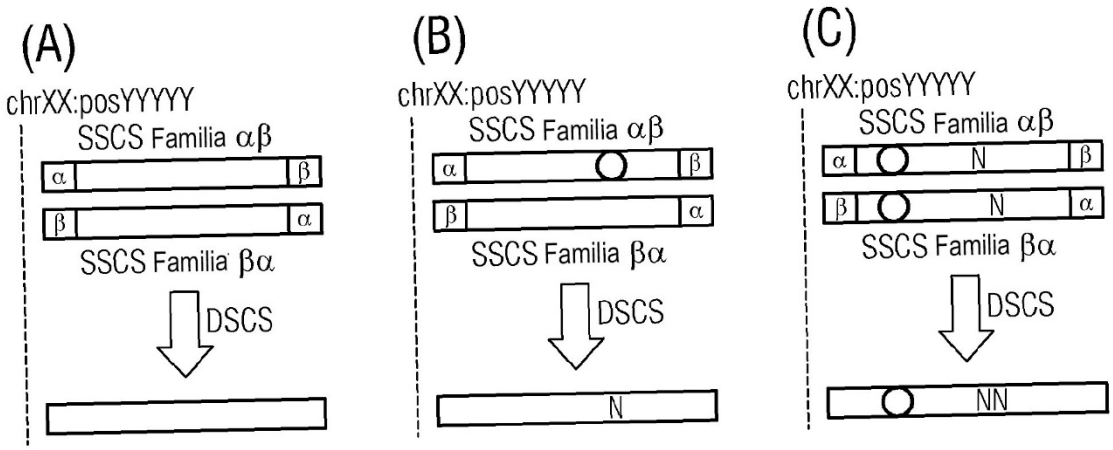
(A)
chrXX:posYYYYY
SSCS Familia αβ
α
β
β
α
SSCS Familia βα
DSCS
(B)
chrXX:posYYYYY
SSCS Familia αβ
α
β
β
α
SSCS Familia βα
DSCS
N
(C)
chrXX:posYYYYY
SSCS Familia αβ
α
N
β
β
N
α
SSCS Familia βα
DSCS
NN

FIG 4 Ejemplo 1: para una posición dada en la cual se detecta un nucleótido diferente del nucleótido anotado en el genoma de referencia, si en esa posición hay más del doble de "N" presentes (en comparación con la variante), no se llama mutación.

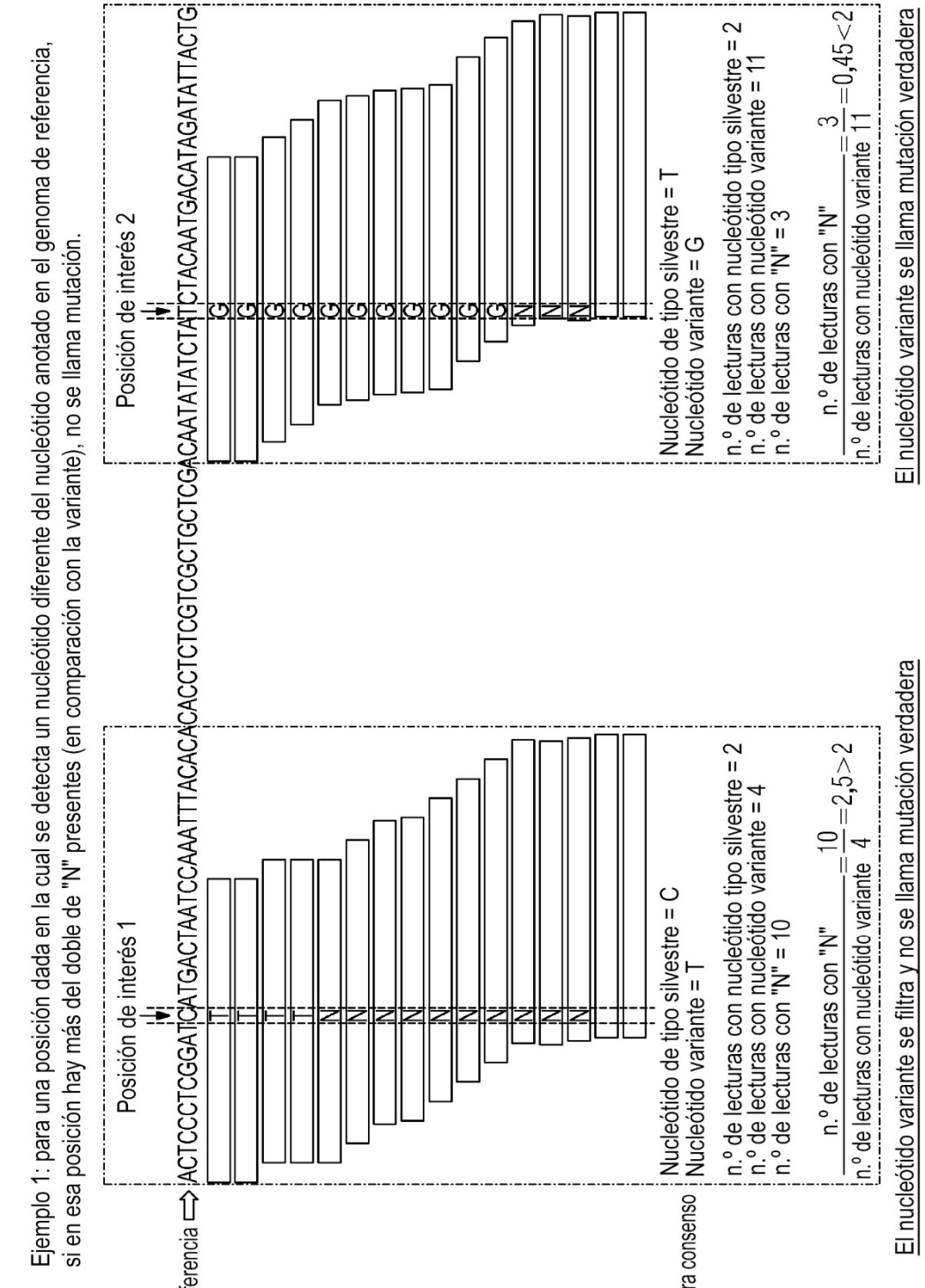

FIG 5 Ejemplo 2: para una posición dada en la cual se detecta un nucleótido diferente del nucleótido anotado en el genoma de referencia, si en esa posición hay más de cuatro veces de "N" presentes (en comparación con la variante), no se llama mutación.