(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 942 359**

(51) Int. CI.:

G16B 20/20 (2009.01)
G16B 40/00 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **15.05.2019 PCT/CN2019/086972**

(87) Fecha y número de publicación internacional: **19.11.2020 WO20227952**

(96) Fecha de presentación y número de la solicitud europea: **15.05.2019 E 19928972 (9)**

(97) Fecha y número de publicación de la concesión europea: **18.01.2023 EP 3971902**

(54) Título: **Método y aparato de detección de mutaciones de bases basados en datos de secuenciación, y medio de almacenamiento**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
31.05.2023

(73) Titular/es:

BGI GENOMICS CO., LIMITED (100.0%)
Floors 7-14, Building No.7 BGI Park, No.21
Hongan 3rd Street Yantian District, Shenzhen
Guangdong 518083, CN

(72) Inventor/es:

LIU, SIYANG;
HUANG, SHUJIA y
JIN, XIN

(74) Agente/Representante:

URÍZAR VILLATE, Ignacio

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

DESCRIPCIÓN

Método y aparato de detección de mutaciones de bases basados en datos de secuenciación, y medio de almacenamiento

Campo

Las realizaciones de la presente divulgación se refieren a la tecnología de detección de mutaciones genéticas, por ejemplo, se refieren a un método y dispositivo de detección de mutaciones de bases basados en datos de secuenciación, y a un medio de almacenamiento.

Antecedentes

Con el avance y la comercialización de la tecnología de secuenciación de última generación desde 2008, los datos de secuenciación han seguido creciendo, y también se han desarrollado activamente algoritmos para varias aplicaciones posteriores basados los datos de secuenciación. Un tipo de algoritmos que tiene perspectivas de tener amplias aplicaciones se denomina método de detección de mutaciones de bases basado en datos de secuenciación. La variación del genoma incluye la variación de un solo nucleótido, pequeñas inserciones y deleciones, y variaciones estructurales. Las diferentes mutaciones genómicas tienen sus respectivos algoritmos de detección diferentes. La presente divulgación se dirige a la detección de mutaciones de bases. Para la detección de mutaciones de bases en la tecnología de secuenciación de última generación, ha habido algoritmos de código abierto relativamente estables, tales como el kit de herramientas de análisis del genoma (GATK, *Genome Analysis ToolKit*).

Sin embargo, los algoritmos de detección de mutaciones de bases de la técnica relacionada todavía tienen los siguientes principales defectos: en primer lugar, no pueden detectar sitios de mutación trialélica ni tetralélica; y, en segundo lugar, no son adecuados para situaciones de baja profundidad de secuenciación y datos de muestra grandes. Por ejemplo, debido a problemas de memoria y velocidad, los algoritmos de detección de mutaciones de bases no pueden ni siquiera analizar datos de forma estable de más de 50.000 muestras. Las pruebas prenatales no invasivas son un tipo de secuenciación del genoma completo y de profundidad ultrabaja, con una profundidad de secuenciación promedio tan baja como de x0,06 y una tasa de cobertura del genoma de aproximadamente el 6 %. Aunque la profundidad de secuenciación es muy baja, una gran cantidad de datos de secuenciación prenatal no invasiva, tal como la secuenciación de datos de cien mil, un millón o incluso diez millones de casos, constituyendo así un excelente conjunto de datos sin procesar para la detección de mutaciones. Además, se han acumulado en todo el mundo datos de secuenciación prenatal no invasiva de aproximadamente 7,2 millones de personas. El espectro de mutación específico de la población obtenido basándose en los datos acumulados de secuenciación del genoma de la población desempeña un importante papel que guía la investigación de enfermedades y el desarrollo de fármacos. Bing Shan Li *et al.* enseñan un método para detectar mutaciones de bases, "A Likelihood-Based Framework for Variant Calling and De Novo Mutation Detection in Families" (PLOS GENETICS, vol. 8, n.º 10, página e1002944, 2012-10-4). Las iteraciones de maximización de las expectativas (EM, *Expectation Maximization*) se dan a conocer en "A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data" (BIOINFORMATICS, vol. 27, n.º 21, páginas 2987-2993, 2011-11-1). Rosset *et al.* desvela la metodología de máxima verosimilitud (véase *Genetics*, vol. 180, n.º 3, 30 de noviembre de 2008).

En dicho escenario de muestras a gran escala, pero con baja profundidad de secuenciación, los algoritmos de detección de mutaciones de bases conocidos en la técnica relacionada ya no funcionan. Por lo tanto, parece muy necesario desarrollar un nuevo algoritmo para la detección de mutaciones de bases que se adapte al escenario de aplicación anterior.

Sumario

La presente divulgación desvela un método de detección de mutaciones de bases implementado por ordenador, un aparato de detección de mutaciones de bases y un medio de almacenamiento basado en datos de secuenciación. A través del método, se puede realizar la detección de mutaciones de bases para los datos de secuenciación con una profundidad de secuenciación baja y una gran cantidad de datos.

Las realizaciones de la presente divulgación desvelan un método de detección de mutaciones de bases implementado por ordenador basado en datos de secuenciación, que incluye: determinar una frecuencia inicial de datos de secuenciación para una pluralidad de muestras que se han de detectar, correspondiente a una base específica en un locus de interés; calcular, basándose en la frecuencia inicial, un valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; actualizar, utilizando cada valor esperado, la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; calcular además, utilizando la frecuencia inicial actualizada, el valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés, actualizar adicionalmente, utilizando cada nuevo valor esperado, la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han

de detectar, correspondiente a la base específica en el locus de interés, y repetir la iteración anterior hasta que converja el valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; y determinar, basándose en cada valor esperado convergente, un tipo de mutación de bases y una confianza de mutación en el locus de interés de cada una de la pluralidad de muestras que se han de detectar. La base específica incluye una base de adenina A, una base de timina T, una base de citosina C o una base de guanina G. Dicho cálculo, basándose en la frecuencia inicial, el valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés incluye: calcular el valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés de acuerdo con la siguiente ecuación:

$$p(\mathrm{b}_{i,j}\left|p_j,\mathrm{d}_i\right.)=\frac{p(\mathrm{b}_{i,j}\left|p_j\right.)p(\mathrm{d}_i\left|\mathrm{b}_{i,j}\right.)}{\sum_{\mathrm{b}_{i,j=0}}^{3}p(\mathrm{b}_{i,j}\left|p_j\right.)p(\mathrm{d}_i\left|\mathrm{b}_{i,j}\right.)}$$

, donde $\mathrm{b}_{i,j}$ representa que una muestra i que se ha de detectar es una base j específica en un locus de interés; p_j representa una frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica j en el locus de interés; d_i representa un conjunto de bases de secuencias de bases en el locus de interés cubierto por la muestra i que se ha de detectar; $p(\mathrm{b}_{i,j}|p_j,\mathrm{d}_i)$ representa un valor esperado de la muestra i que se ha de detectar correspondiente a la base específica j en el locus de interés; $p(\mathrm{b}_{i,j}|p_j)$ representa una probabilidad a priori de que la muestra i que se ha de detectar corresponda a la base específica j en el locus de interés a una p_j dada, y $p(\mathrm{d}_i|\mathrm{b}_{i,j})$ representa una puntuación de calidad de las bases de las secuencias de bases en el locus de interés cubierto por la muestra i que se ha de detectar. Dicha actualización, utilizando cada valor esperado, de la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés incluye: actualizar, utilizando cada valor esperado, la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés de acuerdo con la siguiente ecuación:

$$\hat{p}_j=\frac{\sum_{i=1}^{N}p(\mathrm{b}_{i,j}\left|p_j,\mathrm{d}_i\right.)}{n}$$

, donde $\hat{p}_j$ representa una frecuencia inicial actualizada de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; $p(\mathrm{b}_{i,j}|p_j,\mathrm{d}_i)$ representa un valor esperado de la muestra i que se ha de detectar correspondiente a la base específica j en el locus de interés; N representa el número para la pluralidad de muestras que se han de detectar; y n representa el número total de bases de cuatro tipos de los datos de secuenciación para la pluralidad de muestras que se han de detectar. Dicha determinación, basándose en cada valor esperado convergente, del tipo de mutación de bases y de la confianza de mutación en el locus de interés para cada una de la pluralidad de muestras que se han de detectar incluye: calcular, basándose en cada valor esperado convergente, una estimación de máxima verosimilitud para la pluralidad de muestras que se han de detectar pertenecientes a cada uno de los cuatro tipos de mutación de bases específicas en el locus de interés; calcular un cociente de las estimaciones de máxima verosimilitud de dos tipos de mutación de bases específicas adyacentes; procesar el cociente según una regla preestablecida para obtener una probabilidad correspondiente al cociente; y cuando la probabilidad sea inferior a un umbral establecido, determinar el tipo de mutación de base específica en el locus de interés de cada muestra que se ha de detectar como el tipo de mutación de base específica correspondiente a un denominador actual. Los cuatro tipos de mutación de bases específicas comprenden la mutación monoalélica, mutación bialélica, mutación trialélica y mutación tetralélica.

Las realizaciones de la presente divulgación desvelan además un aparato de detección de mutaciones de bases basado en datos de secuenciación, que incluye: un módulo de determinación de la frecuencia inicial configurado para determinar una frecuencia inicial de una frecuencia inicial de datos de secuenciación para una pluralidad de muestras que se han de detectar, correspondiente a una base específica en un locus de interés; un módulo de cálculo de valores esperados configurado para calcular, basándose en la frecuencia inicial, un valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; un módulo de actualización configurado para actualizar, utilizando cada valor esperado, la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; un módulo de iteración configurado para: calcular además, utilizando la frecuencia inicial actualizada, el valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés, actualizar adicionalmente, utilizando cada nuevo valor esperado, la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés, y repetir la iteración anterior hasta que converja el valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; y un módulo de determinación del tipo de mutación configurado para determinar, basándose en cada valor esperado convergente, un tipo de mutación de bases y una confianza de mutación en el locus de interés de cada una de la pluralidad de muestras que se han de detectar. La base específica incluye base de adenina A, base de timina T, base de citosina C o base de guanina G. El módulo de cálculo del valor esperado está además configurado para calcular el valor esperado para cada una de la pluralidad de muestras que se han de

detectar, correspondiente a la base específica en el locus de interés, de acuerdo con la siguiente ecuación:

$$p(\mathrm{b}_{i,j}\left|p_j,\mathrm{d}_i\right.)=\frac{p(\mathrm{b}_{i,j}\left|p_j\right.)p(\mathrm{d}_i\left|\mathrm{b}_{i,j}\right.)}{\sum_{\mathrm{b}_{i,j=0}}^{3}p(\mathrm{b}_{i,j}\left|p_j\right.)p(\mathrm{d}_i\left|\mathrm{b}_{i,j}\right.)}$$

, donde $\mathrm{b}_{i,j}$ representa que una muestra i que se ha de detectar es una base j específica en un locus de interés; p_j representa una frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica j en el locus de interés; d_i representa un conjunto de bases de secuencias de bases en el locus de interés cubierto por la muestra i que se ha de detectar; $p(b_{i,j}|p_j,d_i)$ representa un valor esperado de la muestra i que se ha de detectar correspondiente a la base específica j en el locus de interés; $p(\mathrm{b}_{i,j}|p_j)$ representa una probabilidad a priori de que la muestra i que se ha de detectar corresponda a la base específica j en el locus de interés a una p_j dada; and $p(d_i|\mathrm{b}_{i,j})$ representa una puntuación de calidad de las bases de las secuencias de bases en el locus de interés cubierto por la muestra i que se ha de detectar. El módulo de actualización está además configurado para actualizar, utilizando cada valor esperado, la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés de acuerdo con la siguiente ecuación:

$$\hat{p}_j=\frac{\sum_{i=1}^{N}p(\mathrm{b}_{i,j}\left|p_j,\mathrm{d}_i\right.)}{n}$$

, donde $\hat{p}_j$ representa una frecuencia inicial actualizada de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; $p(\mathrm{b}_{i,j}|p_j,\mathrm{d}_i)$ representa un valor esperado de la muestra i que se ha de detectar correspondiente a la base específica j en el locus de interés; N representa el número para la pluralidad de muestras que se han de detectar; y n representa el número total de bases de cuatro tipos de los datos de secuenciación para la pluralidad de muestras que se han de detectar. El módulo de determinación del tipo de mutación se configura además para: calcular, basándose en cada valor esperado convergente, una estimación de máxima verosimilitud para la pluralidad de muestras que se han de detectar pertenecientes a cada uno de los cuatro tipos de mutación de bases específicas en el locus de interés; calcular un cociente de las estimaciones de máxima verosimilitud de dos tipos de mutación de bases específicas adyacentes; procesar el cociente según una regla preestablecida para obtener una probabilidad correspondiente al cociente; y cuando la probabilidad sea inferior a un umbral establecido, determinar el tipo de mutación de base específica en el locus de interés de cada muestra que se ha de detectar como el tipo de mutación de base específica correspondiente a un denominador actual. Los cuatro tipos de mutación de bases específicas comprenden la mutación monoalélica, mutación bialélica, mutación trialélica y mutación tetralélica.

Las realizaciones de la presente divulgación desvelan además un medio de almacenamiento que incluye una instrucción ejecutable por ordenador. La instrucción ejecutable por ordenador, cuando es ejecutada por un procesador de un ordenador, implementa el método de detección de mutaciones de bases mencionado anteriormente basado en datos de secuenciación.

El método de detección de mutaciones de bases mencionado anteriormente basado en datos de secuenciación desvelado en las realizaciones de la presente divulgación puede realizar la detección de las mutaciones de las bases para una gran cantidad de datos de secuenciación con una profundidad de secuenciación baja.

Breve descripción de los dibujos

La FIG. 1 es un diagrama de flujo esquemático de un método de detección de mutaciones de bases basado en datos de secuenciación de acuerdo con una realización de la presente divulgación;
la FIG. 2 es un diagrama estructural esquemático de un aparato de detección de mutaciones de bases basado en datos de secuenciación de acuerdo con una realización de la presente divulgación;
la FIG. 3 es un diagrama esquemático que ilustra recuperaciones de diferentes frecuencias de mutación con diferentes datos de muestra utilizando el método de detección de mutaciones de bases mencionado anteriormente de acuerdo con una realización de la presente divulgación;
la FIG. 4 es un diagrama esquemático que ilustra la recuperación de diferentes tipos de mutación bajo diferentes frecuencias de mutación utilizando el método de detección de mutaciones de bases mencionado anteriormente de acuerdo con una realización de la presente divulgación, donde el número de muestras es 140.000;
la FIG. 5 es un diagrama esquemático que ilustra las diferencias entre los valores verdaderos y estimados de la frecuencia de mutación de un sitio de mutación a diferentes frecuencias de mutación utilizando el método de detección de mutaciones de bases mencionado anteriormente de acuerdo con una realización de la presente divulgación; y
la FIG. 6 es un diagrama esquemático que ilustra las diferencias entre los valores verdaderos y estimados de la frecuencia de mutación de un sitio de mutación a diferentes frecuencias de mutación utilizando el método de detección de mutaciones de bases mencionado anteriormente de acuerdo con otra realización de la presente divulgación.

Descripción de las realizaciones

A continuación, se presenta una descripción general de los temas descritos en el presente documento. Esta descripción general no pretende limitar el alcance de la protección de las reivindicaciones.

A menos que se definan de otra manera, todos los términos técnicos y científicos utilizados en el presente documento tienen el significado común que pueden entender los expertos en la materia. La terminología utilizada en el presente documento pretende describir realizaciones específicas y no se utiliza para limitar la presente divulgación. El término "y/o" como se utiliza en el presente documento incluye todas y cada una de las combinaciones de uno o más elementos enumerados.

El método de detección de mutaciones de bases basado en los datos de secuenciación desvelado en las realizaciones de la presente divulgación puede aplicarse a pruebas prenatales u otros datos de secuenciación de población a gran escala, cuya profundidad de secuenciación suele ser baja. El método de detección de mutaciones de bases basado en los datos de secuenciación proporcionados en la presente divulgación también se puede utilizar en situaciones de mayor profundidad de secuenciación. Aunque la profundidad de secuenciación de cada muestra sea muy baja, cada sitio está cubierto por muchas secuencias de bases que se multiplican (en lo sucesivo en el presente documento denominadas lecturas) cuando las muestras son numerosas. Cada base derivada de una lectura independiente y que cubre el sitio actual corresponderá a una puntuación de calidad de la base generada por el secuenciador. La puntuación de calidad de la base se registra como $p(d_i|b_{i,j})$, que representa específicamente una puntuación de calidad de una base j, que se obtiene de una lectura independiente d_i de una muestra i y cubre el sitio actual. La puntuación de calidad de la base refleja una probabilidad de error de detección de la base correspondiente. Por ejemplo, la base cubierta por una primera lectura en un primer locus L1 = {A,T} es una base A, y una puntuación de calidad de la base de A es 30, lo que significa que la probabilidad de error de detección de la base A en el primer locus de la primera lectura es $10^{-(30/10)} = 10^{-3}$, y la probabilidad correcta correspondiente es $1\text{-}10^{-3}$, donde A representa una base de adenina, T representa una base de timina, C representa una base de citosina y G representa una base de guanina. Si la muestra es una muestra prenatal y el sitio actual está cubierto por más de una lectura, solo se extrae una secuencia de lectura de cada muestra, para evitar el problema de que estas lecturas no sean independientes debido a la mezcla de ADN fetal y materno.

La FIG. 1 es un diagrama de flujo esquemático de un método de detección de mutaciones de bases basado en datos de secuenciación de acuerdo con una realización de la presente divulgación. Como se muestra en la FIG. 1, el método incluye las siguientes etapas.

En la etapa 110, se determina una frecuencia inicial de una base específica en un locus de interés en los datos de secuenciación para una pluralidad de muestras que se han de detectar.

El locus de interés se refiere a una posición en la que se ha de detectar si se produce una mutación de una base. La base específica incluye una base de adenina (A), una base de timina (T), una base de citosina (C) o una base de guanina (G).

En una realización, dicha determinación de la frecuencia inicial de una base específica en un locus de interés en los datos de secuenciación para la pluralidad de muestras que se han de detectar: calcular estadísticamente el número de la base específica en los datos de secuenciación para la pluralidad de muestras que se han de detectar y el número total de bases de cuatro tipos en los datos de secuenciación para la pluralidad de muestras que se han de detectar; y determinar el cociente de dividir el número de la base específica entre el número total de las bases de cuatro tipos como la frecuencia inicial de la base específica en el locus de interés basándose en los datos de secuenciación para la pluralidad de muestras que se han de detectar.

El proceso mencionado anteriormente para determinar la frecuencia inicial se expresa mediante una ecuación:

$$p_j = \frac{\sum_{i=1}^{N} b_{ij}}{n} \qquad (1)$$

donde b_{ij} representa el número de bases específicas j portadas en los datos de secuenciación de una muestra i que se ha de detectar; N representa el número total de muestras que se han de detectar; $\sum_{i=1}^{N} b_{ij}$ representa el número total de base j específica portada en los datos de secuenciación para la pluralidad de muestras que se han de detectar; n representa el número total de bases de los cuatro tipos en los datos de secuenciación para la pluralidad de muestras que se han de detectar; p_j representa la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica j en el locus de interés, j = {0,1,2,3}; p_0 representa la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base A en el locus de interés; p_1 representa la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base C en el locus de interés; p_2 representa la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han

de detectar correspondiente a la base G en el locus de interés; y p_3 representa la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar correspondiente a la base T en el locus de interés.

En la etapa 120, se calcula un valor esperado para cada muestra que se ha de detectar, correspondiente a la base específica en el locus de interés, basado en la frecuencia inicial.

En una realización, el valor esperado del caso en que cada muestra que se ha de detectar tenga la base específica en el locus de interés se calcula de acuerdo con la siguiente ecuación (2):

$$p(\mathrm{b}_{i,j} \left| p_j, \mathrm{d}_i \right.) = \frac{p(\mathrm{b}_{i,j} \left| p_j \right.) p(\mathrm{d}_i \left| \mathrm{b}_{i,j} \right.)}{\sum_{\mathrm{b}_{i,j}=0}^{3} p(\mathrm{b}_{i,j} \left| p_j \right.) p(\mathrm{d}_i \left| \mathrm{b}_{i,j} \right.)} \qquad (2)$$

donde $\mathrm{b}_{i,j}$ representa que una muestra i que se ha de detectar es una base j específica en un locus de interés; p_j representa una frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica j en el locus de interés; d_i representa un conjunto de bases de secuencias de bases en el locus de interés cubierto por la muestra i que se ha de detectar; $p(\mathrm{b}_{i,j}$ representa un valor esperado de la muestra i que se ha de detectar, correspondiente a la base específica j en el locus de interés; $p(\mathrm{b}_{i,j}|p_j)$ representa una probabilidad a priori de que la muestra i que se ha de detectar corresponda a la base específica j en el locus de interés a una p_j dada; y $p(\mathrm{d}_i\mathrm{b}_{i,j})$ representa una puntuación de calidad de las bases de secuencias de bases en el locus de interés cubierto por la muestra i que se ha de detectar.

En la etapa 130, utilizando cada valor esperado, se actualiza la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés.

En una realización, la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés, se actualiza utilizando cada valor esperado de acuerdo con la siguiente ecuación (3):

$$\hat{p}_j = \frac{\sum_{i=1}^{N} p(\mathrm{b}_{i,j} \left| p_j, \mathrm{d}_i \right.)}{n} \qquad (3)$$

donde $\hat{p}_j$ representa una frecuencia inicial actualizada de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; $p(\mathrm{b}_{i,j}$ representa un valor esperado de la muestra i que se ha de detectar, correspondiente a la base específica j en el locus de interés; N representa el número para la pluralidad de muestras que se han de detectar; y n representa el número total de bases de cuatro tipos de los datos de secuenciación para la pluralidad de muestras que se han de detectar. Las bases de cuatro tipos se refieren a las bases A, T, C y G.

En la etapa 140, el valor esperado de cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés se calcula además utilizando la frecuencia inicial actualizada; la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés se actualiza adicionalmente utilizando cada nuevo valor esperado; y se repite la iteración anterior hasta que converja el valor esperado de cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés.

En una realización, el resultado $\hat{p}_j$ obtenido de la ecuación anterior (3) se utiliza para reemplazar p_j en la ecuación (2), de modo que el nuevo valor esperado de cada muestra que se ha de detectar, correspondiente a la base específica en el locus de interés se obtiene de la ecuación (2). El nuevo valor esperado se utiliza para actualizar adicionalmente la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés, en función de la ecuación anterior (3), hasta que converja el valor esperado obtenido de la ecuación (2). Las condiciones para la convergencia del valor esperado incluyen: que la diferencia entre los valores esperados obtenidos en sucesivas iteraciones sea inferior a una diezmilésima para tres veces consecutivas. Por ejemplo, el valor esperado obtenido de la 10.ª iteración es q10, el valor esperado obtenido de la 11.ª iteración es q11 y el valor esperado obtenido en la 12.ª iteración es q12, y el valor esperado obtenido en 13.ª iteración es q13; cuando (q11-q10) < 0,0001, (q12-q11) < 0,0001 y (q13-q12) < 0,0001, se puede determinar que el valor esperado cumple las condiciones de convergencia, y el valor esperado q13 es el

valor esperado convergente.

La frecuencia de bases óptima no se puede obtener hasta que el valor esperado converja, es decir, la frecuencia de bases óptima es un valor de la ecuación anterior (3) correspondiente a la convergencia del valor esperado, es decir, la frecuencia del locus de interés correspondiente a la base específica j.

En la etapa 150, se determina un tipo de mutación de bases y una confianza de mutación en el locus de interés para cada muestra que se ha de detectar basándose en cada valor esperado convergente.

En una realización, dicha determinación, basándose en cada valor esperado convergente, del tipo de mutación de bases y de la confianza de mutación en el locus de interés para cada una de la pluralidad de muestras que se han de detectar incluye: calcular, basándose en cada valor esperado convergente, una estimación de máxima verosimilitud para la pluralidad de muestras que se han de detectar pertenecientes a cada uno de los cuatro tipos de mutación de bases específicas en el locus de interés; calcular un cociente de las estimaciones de máxima verosimilitud de dos tipos de mutación de bases específicas adyacentes; procesar el cociente según una regla preestablecida para obtener una probabilidad correspondiente al cociente; y determinar el tipo de mutación de base específica en el locus de interés de cada muestra que se ha de detectar como el tipo de mutación de base específica correspondiente a un denominador actual, cuando la probabilidad es inferior a un umbral establecido. Los cuatro tipos de mutación de bases específicas comprenden la mutación monoalélica, mutación bialélica, mutación trialélica y mutación tetralélica.

Dicho cálculo, basándose en cada valor esperado convergente, la estimación de máxima verosimilitud para la pluralidad de muestras que se han de detectar pertenecientes a cada uno de los cuatro tipos de mutación de bases específicas en el locus de interés incluye: calcular la estimación de máxima verosimilitud para la pluralidad de muestras que se han de detectar pertenecientes a cada uno de los cuatro tipos específicos de mutación de bases en el locus de interés de acuerdo con la siguiente ecuación (4):

$$p(\mathrm{D}\left|p_j\right.)=\prod_{i=1}^{N}\sum_{j=0}^{3}\left[\,p(\mathrm{b}_{i,j}\left|p_j\right.)\,\mathrm{p}(\mathrm{d}_i\left|b_{i,j}\right.)\right] \qquad (4)$$

donde D representa datos de observación que comprenden conjuntos de bases de secuencias de bases en el locus de interés cubierto por toda la pluralidad que se ha de detectar; p_j representa una frecuencia de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a una base específica j en el locus de interés, que se obtiene a partir de cada valor esperado convergente; $p(D|p_j)$ representa una estimación de máxima verosimilitud para la pluralidad de muestras que se han de detectar, correspondiente a un tipo de mutación de bases correspondiente a j en el locus de interés; cuando j = 0, el tipo de mutación de base específica correspondiente a j es la mutación monoalélica; cuando j = 1, el tipo de mutación de base específica correspondiente a j es la mutación bialélica; cuando j = 2, el tipo de mutación de base específica correspondiente a j es la mutación trialélica; y cuando j = 3, el tipo de mutación de base específica correspondiente a j es la mutación tetralélica.

En una realización, dicho cálculo del cociente de las estimaciones de máxima verosimilitud de los dos tipos de mutaciones de bases específicas adyacentes incluye: suponer que una estimación de máxima verosimilitud del locus de interés correspondiente a la mutación tetralélica es f_4, y una estimación de máxima verosimilitud del locus de interés correspondiente a la mutación trialélica es f_3; y determinarlo $\frac{f_3}{f_4}$ como el cociente de las estimaciones de máxima verosimilitud de los dos tipos de mutaciones de bases específicas adyacentes.

Suponiendo que una estimación de máxima verosimilitud del locus de interés correspondiente a la mutación bialélica es f_2, y un valor mínimo de las estimaciones de máxima verosimilitud de cuatro combinaciones de mutaciones en las mutaciones trialélicas es f_3 m en; las cuatro combinaciones de mutaciones de la mutación trialélica son {A,T,C}, {A,T,G}, {T,C,G} y {A,C,G}, respectivamente, y cada combinación de mutaciones corresponde a una estimación de máxima verosimilitud descrita en la ecuación anterior (4). Por tanto, $\frac{f_2}{f_{3\,\min}}$ se determina como el cociente de las estimaciones de máxima verosimilitud de los dos tipos de mutaciones de bases específicas adyacentes.

Suponiendo que una estimación de máxima verosimilitud del locus de interés correspondiente a una mutación monoalélica es f_1, y un valor mínimo de las estimaciones de máxima verosimilitud de 16 combinaciones de mutaciones en las mutaciones bialélicas es f_2 m en, entonces, $\frac{f_1}{f_{2\,\min}}$ se determina como el cociente de las estimaciones de máxima verosimilitud de los dos tipos de mutaciones de bases específicas adyacentes.

Correspondientemente, dicho procesamiento del cociente de acuerdo con la regla preestablecida para obtener la

probabilidad correspondiente al cociente incluye: calcular un logaritmo natural del cociente para obtener un primer resultado; multiplicar el primer resultado obtenido por -2, obteniéndose un segundo resultado; y obtener una probabilidad correspondiente al segundo resultado consultando la tabla de distribución de chi-cuadrado.

Se supone que la frecuencia de mutación de una base cualquiera es 0 en el caso de la mutación trialélica, la frecuencia de mutación de dos bases cualesquiera en el caso de una mutación bialélica se supone que es 0, y la frecuencia de mutación de tres bases cualesquiera en el caso de una mutación monoalélica se supone que es 0. En este sentido, el numerador de cada una de las fracciones anteriores $\frac{f_3}{f_4}$, $\frac{f_2}{f_{3\,\min}}$ y $\frac{f_1}{f_{2\,\min}}$) es inferior al denominador por un parámetro de frecuencia Pj. Por lo tanto, el valor estadistico, que se obtiene calculando el logaritmo natural del cociente anterior y multiplicando el logaritmo natural por -2, obedece a una distribución chi-cuadrado con un grado de libertad. Por tanto, la probabilidad correspondiente a cada cociente se puede obtener utilizando la tabla de distribución chi-cuadrado.

El valor estadístico obtenido al calcular el logaritmo natural del cociente anterior y multiplicar el logaritmo natural por -2 es el siguiente:

$$LRT_{4vs3} = -2\ln(\frac{f_3}{f_4})$$

$$LRT_{3vs2} = -2\ln(\frac{f_2}{f_{3\,\min}})$$

$$LRT_{2vs1} = -2\ln(\frac{f_1}{f_{2\,\min}})$$

Cuando la probabilidad es inferior al umbral establecido, por ejemplo, cuando la probabilidad correspondiente a LRT_{4vS3} es inferior a 10^{-6}, se determina que la suposición no es cierta, es decir, se determina que el locus de interés no pertenece a la mutación trialélica, sino que pertenece a la mutación tetralélica. Cuando la probabilidad correspondiente a LRT_{3vs2} es inferior a 10^{-6}, se determina que el locus de interés no pertenece a la mutación bialélica, sino que pertenece a la mutación trialélica. Cuando la probabilidad correspondiente a LRT_{2vs1} es inferior a 10^{-6}, se determina que el locus de interés no pertenece a la mutación monoalélica, sino que pertenece a la mutación bialélica.

En una realización, dicha determinación, basándose en cada valor esperado convergente, la confianza de la mutación en el locus de interés para cada muestra que se ha de detectar incluye: realizar una transformación regular en la escala de Phred sobre la probabilidad correspondiente a $\frac{f_1}{f_{2\,\min}}$ para obtener una puntuación de calidad de Phred; y determinar la puntuación de calidad de Phred como la confianza de la mutación en el locus de interés para cada muestra que se ha de detectar.

La confianza de la mutación se puede utilizar para determinar además si la mutación de base en el locus de interés es verdadera.

Las realizaciones de la presente divulgación desvelan un método de detección de mutaciones de bases basado en datos de secuenciación, que es completamente diferente del método de detección de mutaciones conocido en la técnica relacionada, y que tiene ventajas significativas en la complejidad temporal. Al utilizar las características de datos de una gran cantidad de datos de muestra y la baja profundidad de secuenciación de muestras individuales, se utiliza directamente el método de función de probabilidad para observar los datos en función de los tipos de mutación alélica, en lugar de usar la función de probabilidad de observar los datos basados en el genotipo. De esta manera, todo el método de detección es más rápido y eficiente, y permite el análisis de más de 100.000 o incluso un millón de muestras (por ejemplo, un millón de casos de datos de pruebas prenatales). Además, en las realizaciones de la presente divulgación, en lugar de presuponer que el locus de interés es una mutación bialélica, por primera vez, se utiliza una pluralidad de pruebas de verosimilitud para probar las estimaciones de máxima verosimilitud de diferentes combinaciones de bases. Por lo tanto, el método de detección de mutaciones de bases de las realizaciones de la presente divulgación tiene la capacidad de detectar mutaciones bialélicas, así como mutaciones monoalélicas y multialélicas. Cuando el método de detección de mutaciones de bases de las realizaciones de la presente divulgación se usa para analizar los datos de pruebas prenatales de cien mil o un millón de casos, se puede obtener información de la frecuencia y posición de las mutaciones de la población con una alta precisión, y dicha información tiene valor tanto para la investigación científica como para la industria.

La FIG. 2 es un diagrama estructural esquemático de un aparato de detección de mutaciones de bases basado en datos de secuenciación de acuerdo con una realización de la presente divulgación. Como se muestra en la FIG. 2, el dispositivo incluye un módulo de determinación de la frecuencia inicial 210, un módulo de cálculo de valor esperado 220, un módulo de actualización 230, un módulo de iteración 240 y un módulo de determinación del tipo de mutación 250.

En esta realización, el módulo de determinación de la frecuencia inicial 210 está configurado para determinar una

frecuencia inicial de una frecuencia inicial de datos de secuenciación para una pluralidad de muestras que se han de detectar, correspondiente a una base específica en un locus de interés; el módulo de cálculo del valor esperado 220 está configurado para calcular, basándose en la frecuencia inicial, un valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; el módulo de actualización 230 está configurado para actualizar, utilizando cada valor esperado, la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; el módulo de iteración 240 está configurado para calcular además, utilizando la frecuencia inicial actualizada, el valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés, actualizar adicionalmente, utilizando cada nuevo valor esperado, la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés, y repetir la iteración anterior hasta que converja el valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; el módulo de determinación del tipo de mutación 250 está configurado para determinar, basándose en cada valor esperado convergente, un tipo de mutación de bases y una confianza de mutación en el locus de interés de cada una de la pluralidad de muestras que se han de detectar. La base específica incluye la base de adenina A, base de timina T, base de citosina C o base de guanina G.

El aparato de detección de mutaciones de bases basado en datos de secuenciación desvelado en la realización de la presente divulgación puede lograr la detección de mutaciones de bases para una gran cantidad de datos de secuenciación con poca profundidad de secuenciación, consistiendo en determinar una frecuencia inicial de datos de secuenciación para una pluralidad de muestras que se han de detectar, correspondiente a una base específica en un locus de interés; calcular, basándose en la frecuencia inicial, un valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; actualizar, utilizando cada valor esperado, la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; calcular además, utilizando la frecuencia inicial actualizada, el valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés, actualizar adicionalmente, utilizando cada nuevo valor esperado, la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés, y repetir la iteración anterior hasta que converja el valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; determinar, basándose en cada valor esperado convergente, un tipo de mutación de bases y una confianza de mutación en el locus de interés de cada una de la pluralidad de muestras que se han de detectar.

Cabe señalar que, el principio de funcionamiento específico de cada componente de la realización del dispositivo puede referirse a la parte correspondiente de la realización del método, que no se repetirá en el presente documento.

En una realización, para verificar el efecto del método de detección de mutaciones de bases mencionado anteriormente, ahora se proporciona un ejemplo de análisis mediante el uso del método de detección de mutaciones de bases mencionado anteriormente. En este ejemplo, el método de detección de mutaciones de bases mencionado anteriormente se utiliza para analizar datos de simulación a gran escala, para probar la recuperación de mutaciones, la tasa de falsos positivos y la diferencia entre los valores estimados y verdaderos de la frecuencia de posición de las mutaciones del método de detección de mutaciones de bases mencionado anteriormente.

Ejemplo

Resumen de los datos de simulación

Los datos de simulación simulan un total de 100 locus monopolímicos, 50.000 locus dialélicos, 50.000 locus trialélicos y 50.000 locus tetralélicos. Para cada uno de los tres grupos anteriores de 50.000 locus, la frecuencia mínima de mutación alélica se establece en un intervalo de 1/10.000 de los locus, por ejemplo, tal como se muestra en la tabla 1:

[Tabla 1] Distribución de la frecuencia de las mutaciones de los datos de simulación

Locus alélico	Frecuencia de la mutación (A)	Frecuencia de la mutación (C)	Frecuencia de la mutación (G)	Frecuencia de la mutación (T)	Tipo de alelo
1	1	0	0	0	1
2	1	0	0	0	1
3	1	0	0	0	1
…					
98	1	0	0	0	1

99	1	0	0	0	1
100	1	0	0	0	1
…					

(continuación)

Locus alélico	Frecuencia de la mutación (A)	Frecuencia de la mutación (C)	Frecuencia de la mutación (G)	Frecuencia de la mutación (T)	Tipo de alelo
101	1,00E-05	0,99999	0	0	2
102	2,00E-05	0,99998	0	0	2
103	3,00E-05	0,99997	0	0	2
…					
50098	0,49998	0,50002	0	0	2
50099	0,49999	0,50001	0	0	2
50100	0,5	0,5	0	0	2
…					
50101	1,00E-05	0,499995	0,499995	0	3
50102	2,00E-05	0,49999	0,49999	0	3
50103	3,00E-05	0,499985	0,499985	0	3
…					
100098	0,49998	0,25001	0,25001	0	3
100099	0,49999	0,250005	0,250005	0	3
100100	0,5	0,25	0,25	0	3
…					
100101	1,00E-05	0,33333	0,33333	0,33333	4
100102	2,00E-05	0,333326667	0,333326667	0,333326667	4
100103	3,00E-05	0,333323333	0,333323333	0,333323333	4
…					
150098	0,49998	0,166673333	0,166673333	0,166673333	4
150099	0,49999	0,16667	0,16667	0,16667	4
150100	0,5	0,166666667	0,166666667	0,166666667	4

Tras establecer la distribución de la frecuencia de las mutaciones de bases como se ilustra en la Tabla 1, la profundidad de secuenciación de la muestra se establece en x0,06 y la tasa de error de secuenciación se establece en aproximadamente 0,01. El objetivo principal de los datos de simulación es observar la recuperación de mutaciones, la tasa de falsos positivos y la diferencia entre los valores estimados y verdaderos de la frecuencia de los locus de las mutaciones del método de detección de mutaciones de bases mencionado anteriormente en función de los datos de secuenciación proporcionados en las realizaciones de la presente divulgación en diferentes condiciones (diferentes tamaños de muestra, diferentes frecuencias de mutación y diferentes tipos de mutación). La FIG. 3 es un diagrama esquemático de recuperaciones de diferentes frecuencias de mutación utilizando el método de detección de mutaciones de bases mencionado anteriormente con diferentes datos de muestra de acuerdo con una realización de la presente divulgación, donde el eje horizontal representa la frecuencia de mutación y el eje vertical representa la recuperación. La FIG. 3 ilustra las recuperaciones de tamaños de muestra de 44.000 personas, 140.000 personas y 1 millón de personas bajo diferentes frecuencias de mutación. A un cierto tamaño de muestra, la frecuencia mínima de mutación que el método de detección de mutaciones de bases basado en datos de secuenciación mencionado anteriormente puede detectar (cuando la recuperación es superior a 0) y la frecuencia mínima de mutación cuando la recuperación alcanza el 100 % suele recibir la atención de los investigadores. En la FIG. 3, se puede observar que, con el aumento del tamaño de la muestra de 44.000 a 140.000 o a 1.000.000, las frecuencias de mutación mínimas detectadas cuando la recuperación es superior a 0 son 0,002, 0,001 y 0,00002, respectivamente, y las frecuencias de mutación mínimas cuando la recuperación alcanza el 100 % (frecuencias de mutación de detección completa) son 0,015, 0,005 y 0,002, respectivamente, revelando los buenos rendimientos del método de detección de mutaciones de bases mencionado anteriormente basado en datos de secuenciación. En particular, ninguna de las mutaciones de los 100 locus monopolímicos son falsos positivos, es decir, la tasa de falsos positivos es cero. Los resultados mostrados en la FIG. 3 indican que el método de detección de mutaciones de

bases basado en datos de secuenciación proporcionado en las realizaciones de la presente divulgación puede detectar con sensibilidad y precisión mutaciones monoalélicas a partir de una gran cantidad de datos de secuenciación con una profundidad tan baja como x0,06.

La FIG. 4 es un diagrama esquemático de recuperaciones de mutaciones de diferentes tipos utilizando el método de detección de mutaciones de bases mencionado anteriormente con diferentes frecuencias de mutación cuando el tamaño de la muestra es de 140.000, donde el eje horizontal representa la frecuencia de mutación y el eje vertical representa la recuperación. En la FIG. 4, se puede observar que, para la mutación bialélica, la frecuencia de mutación mínima que se puede detectar (cuando la recuperación es superior a 0) y la frecuencia de mutación mínima cuando la recuperación alcanza el 100 % son de 0,002 y 0,005, respectivamente. Aunque los recuerdos de las mutaciones trialélicas y tetralélicas relativamente difíciles son ligeramente más bajos que los de las mutaciones bialélicas, la diferencia en efecto es insignificante. Para las mutaciones trialélicas, la frecuencia de mutación mínima que se puede detectar (cuando la recuperación es superior a 0) y la frecuencia de mutación más baja cuando la recuperación alcanza el 100 % son 0,002 y 0,008, respectivamente. Para las mutaciones tetralélicas, la frecuencia de mutación mínima que se puede detectar (cuando la recuperación es superior a 0) y la frecuencia de mutación más baja cuando la recuperación alcanza el 100 % son 0,001 y 0,008, respectivamente. De modo similar, no se detecta ninguno de los sitios sin mutación, lo que refleja la alta precisión del método. Los resultados mostrados en la FIG. 4 demuestran que el método de detección de mutaciones de bases basado en datos de secuenciación proporcionado por las realizaciones de la presente divulgación puede detectar las mutaciones trialélicas o tetralélicas con sensibilidad y precisión.

La FIG. 5 y la FIG. 6 son diagramas esquemáticos que ilustran las diferencias entre los valores estimados y verdaderos de la frecuencia de mutación de la posición de mutación utilizando el método de detección de mutaciones de bases mencionado anteriormente con diferentes frecuencias de mutación, donde el eje horizontal representa la frecuencia de mutación, el eje vertical de la FIG. 5 representa el error cuadrático medio y el eje vertical de la FIG. 6 representa el coeficiente de variación del error cuadrático medio. La FIG. 5 refleja una diferencia promedio entre una frecuencia de cada banda de frecuencia y una frecuencia de mutación verdadera, que refleja intuitivamente la diferencia aproximada entre la frecuencia detectada y la frecuencia real, pero que no permite una comparación transversal de las diferentes frecuencias debido a las diferentes medidas de las diferentes bandas de frecuencia. En la FIG. 6, se unifican las medidas de las distintas bandas de frecuencia, de manera que la frecuencia detectada por el método y la frecuencia real puedan compararse en diferentes bandas de frecuencia para evaluar la diferencia entre ellas. En la FIG. 5, se puede observar que, con el aumento de la frecuencia de mutación simulada, la frecuencia detectada es ligeramente diferente de la frecuencia real. En la FIG. 6, se puede observar que una diferencia de medición de la frecuencia de baja frecuencia (<0,003) y la banda de frecuencia real es de aproximadamente 1 vez a 1,5 veces, y la mutación con una frecuencia superior a 0,003 es inferior a la diferencia de medición. A través de este análisis, se puede ver que el método anterior puede detectar con precisión la verdadera frecuencia de mutación para la posición de la mutación detectada.

Una realización de la presente divulgación proporciona además un medio de almacenamiento que contiene una instrucción ejecutable por ordenador, y la instrucción ejecutable por ordenador, cuando es ejecutada por un procesador de un ordenador, implementa el método de detección de mutaciones de bases mencionado anteriormente basado en datos de secuenciación.

Cabe señalar que los términos de relación como "primero" y "segundo" de la presente divulgación solo se utilizan para distinguir una entidad u operación de otra entidad u operación, y no necesariamente requieren o implican la presencia de dicha relación real o secuencia de estas entidades u operaciones. Por otra parte, el término "incluir", "comprender", "contener", o cualquier otra variante del mismo significa una inclusión no excluyente. Por tanto, un proceso, método, artículo o dispositivo que se describe que incluye una serie de elementos, no solo incluye los elementos mencionados, sino que también incluye otros elementos que no se enumeran explícitamente, o incluye además elementos inherentes al proceso, método, artículo o dispositivo. Sin más restricciones, un elemento definido por la expresión "que incluye un...", "que comprende un..." no excluye la presencia de otros elementos idénticos en el proceso, método, artículo o dispositivo que incluye el elemento.

La descripción anterior de las realizaciones desveladas permite a los expertos en la materia implementar o utilizar la presente divulgación.

REIVINDICACIONES

1. Un método de detección de mutaciones de bases implementado por ordenador basado en datos de secuenciación, comprendiendo el método de detección de mutaciones de bases:

determinar una frecuencia inicial de datos de secuenciación para una pluralidad de muestras que se han de detectar, correspondiente a una base específica en un locus de interés;
calcular, basándose en la frecuencia inicial, un valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés;
actualizar, utilizando cada valor esperado, la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés;
calcular además, utilizando la frecuencia inicial actualizada, el valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés, actualizar adicionalmente, utilizando cada nuevo valor esperado, la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés, y repetir la iteración anterior hasta que converja el valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; y
determinar, basándose en cada valor esperado convergente, un tipo de mutación de bases y una confianza de mutación en el locus de interés de cada una de la pluralidad de muestras que se han de detectar,
en donde la base específica comprende una base de adenina A, una base de timina T, una base de citosina C o una base de guanina G,
en donde dicho cálculo, basándose en la frecuencia inicial, el valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés comprende:

calcular el valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés de acuerdo con la siguiente ecuación:

$$p(\mathrm{b}_{i,j}\left|p_j,\mathrm{d}_i\right.)=\frac{p(\mathrm{b}_{i,j}\left|p_j\right.)p(\mathrm{d}_i\left|\mathrm{b}_{i,j}\right.)}{\sum_{\mathrm{b}_{i,j=0}}^{3}p(\mathrm{b}_{i,j}\left|p_j\right.)p(\mathrm{d}_i\left|\mathrm{b}_{i,j}\right.)}$$

donde $b_{i,j}$ representa que una muestra i que se ha de detectar es una base j específica en un locus de interés; p_j representa una frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica j en el locus de interés; d, representa un conjunto de bases de secuencias de bases en el locus de interés cubierto por la muestra i que se ha de detectar; $p(\mathrm{b}_{i,j}|p_j,\mathrm{d}_i)$ representa un valor esperado de la muestra i que se ha de detectar correspondiente a la base específica j en el locus de interés; $p(\mathrm{b}_{i,j}|p_j)$ representa una probabilidad a priori de que la muestra i que se ha de detectar corresponda a la base específica j en el locus de interés a una p_j dada; and $p(\mathrm{d}_i|\mathrm{b}_{i,j})$ representa una puntuación de calidad de las bases de las secuencias de bases en el locus de interés cubierto por la muestra i que se ha de detectar,

en donde dicha actualización, utilizando cada valor esperado, de la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés, comprende:

actualizar, utilizando cada valor esperado, la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés de acuerdo con la siguiente ecuación:

$$\hat{p}_j=\frac{\sum_{i=1}^{N}p(\mathrm{b}_{i,j}\left|p_j,\mathrm{d}_i\right.)}{n}$$

donde $\hat{p}_j$ representa una frecuencia inicial actualizada de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; $p(\mathrm{b}_{i,j}|p_j,\mathrm{d}_i)$ representa un valor esperado de la muestra i que se ha de detectar correspondiente a la base específica j en el locus de interés; N representa el número para la pluralidad de muestras que se han de detectar; y n representa el número total de bases de cuatro tipos de los datos de secuenciación para la pluralidad de muestras que se han de detectar, y

en donde dicha determinación, basándose en cada valor esperado convergente, del tipo de mutación de bases y de la confianza de mutación en el locus de interés para cada una de la pluralidad de muestras que se han de

detectar comprende:

calcular, basándose en cada valor esperado convergente, una estimación de máxima verosimilitud para la pluralidad de muestras que se han de detectar pertenecientes a cada uno de los cuatro tipos de mutación de bases específicas en el locus de interés;
calcular un cociente de las estimaciones de máxima verosimilitud de dos tipos de mutación de bases específicas adyacentes; procesar el cociente según una regla preestablecida para obtener una probabilidad correspondiente al cociente; y
cuando la probabilidad es inferior a un umbral establecido, determinar el tipo de mutación de base específica en el locus de interés de cada muestra que se ha de detectar como el tipo de mutación de base específica correspondiente a un denominador actual,
en donde los cuatro tipos de mutación de bases específicas comprenden la mutación monoalélica, mutación bialélica, mutación trialélica y mutación tetralélica.

2. El método de acuerdo con la reivindicación 1, en donde dicha determinación de la frecuencia inicial de los datos de secuenciación para una pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés comprende:

calcular estadísticamente el número de la base específica en los datos de secuenciación para la pluralidad de muestras que se han de detectar y el número total de bases de cuatro tipos en los datos de secuenciación para la pluralidad de muestras que se han de detectar, en donde las bases de cuatro tipos comprenden las bases de adenina A, las bases de timina T, las bases de citosina C y las bases de guanina G; y
determinar un cociente de la división del número de la base específica entre el número total de las bases de cuatro tipos como la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés.

3. El método de acuerdo con la reivindicación 1, en donde dicho cálculo, basándose en cada valor esperado convergente, de la estimación de máxima verosimilitud para la pluralidad de muestras que se han de detectar pertenecientes a cada uno de los cuatro tipos de mutación de bases específicas en el locus de interés comprende:

calcular la estimación de máxima verosimilitud para la pluralidad de muestras que se han de detectar pertenecientes a cada uno de los cuatro tipos específicos de mutación de bases en el locus de interés de acuerdo con la siguiente ecuación:

$$p(\mathrm{D}\left|p_j\right.)=\prod_{i=1}^{N}\sum_{j=0}^{3}\left[p(\mathrm{b}_{i,j}\left|p_j\right.)\,\mathrm{p}(\mathrm{d}_i\left|b_{i,j}\right.)\right]$$

donde D representa datos de observación que comprenden conjuntos de bases de secuencias de bases en el locus de interés cubierto por toda la pluralidad de muestras que se ha de detectar; p_j representa una frecuencia de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a una base específica j en el locus de interés, que se obtiene a partir de cada valor esperado convergente; *p(D|pj)* representa una estimación de máxima verosimilitud para la pluralidad de muestras que se han de detectar, correspondiente a un tipo de mutación de bases correspondiente a j en el locus de interés; $p(\mathrm{b}_{i,j}|p_j)$ representa una probabilidad a priori de que la muestra i que se ha de detectar corresponda a la base específica j en el locus de interés a una p_j dada; y $p(\mathrm{d}_i|\mathrm{b}_{i,j})$ representa una puntuación de calidad de las bases de una secuencia de bases en el locus de interés cubierto por la muestra i; cuando j = 0, el tipo de mutación de base específica correspondiente a j es la mutación monoalélica; cuando j = 1, el tipo de mutación de base específica correspondiente a j es la mutación bialélica; cuando j = 2, el tipo de mutación de base específica correspondiente a j es la mutación trialélica; y cuando j = 3, el tipo de mutación de base específica correspondiente a j es la mutación tetralélica.

4. El método de acuerdo con la reivindicación 3, en donde dicho cálculo del cociente de las estimaciones de máxima verosimilitud de los dos tipos de mutaciones de bases específicas adyacentes comprende:
determinarlo $\frac{f_3}{f_4}$ como el cociente de las estimaciones de máxima verosimilitud de los dos tipos de mutaciones de bases específicas adyacentes, siempre que una estimación de máxima verosimilitud del locus de interés correspondiente a la mutación tetralélica sea f_4, y una estimación de máxima verosimilitud del locus de interés correspondiente a la mutación trialélica sea f_3.

5. El método de acuerdo con la reivindicación 3, en donde dicho cálculo del cociente de las estimaciones de máxima verosimilitud de los dos tipos de mutaciones de bases específicas adyacentes comprende:
determinarlo $\frac{f_2}{f_3\,\min}$ como el cociente de las estimaciones de máxima verosimilitud de los dos tipos de mutaciones de

bases específicas adyacentes, siempre que una estimación de máxima verosimilitud del locus de interés correspondiente a la mutación bialélica sea f_2, y un valor mínimo de las estimaciones de máxima verosimilitud de cuatro combinaciones de mutaciones en las mutaciones trialélicas sea $f_3 min$.

6. El método de acuerdo con la reivindicación 3, en donde dicho cálculo del cociente de las estimaciones de máxima verosimilitud de los dos tipos de mutaciones de bases específicas adyacentes comprende:

determinarlo $\frac{f_1}{f_2 \min}$ como el cociente de las estimaciones de máxima verosimilitud de los dos tipos de mutaciones de bases específicas adyacentes, siempre que una estimación de máxima verosimilitud del locus de interés correspondiente a la mutación monoalélica sea f_1, y un valor mínimo de las estimaciones de máxima verosimilitud de 16 combinaciones de mutaciones en las mutaciones bialélicas sea $f_2 min$.

7. El método de acuerdo con una cualquiera de las reivindicaciones 4 a 6, en donde dicho procesamiento del cociente de acuerdo con la regla preestablecida para obtener la probabilidad correspondiente al cociente comprende:

calcular un logaritmo natural del cociente para obtener un primer resultado;
multiplicar el primer resultado obtenido por -2, obteniéndose un segundo resultado; y
obtener una probabilidad correspondiente al segundo resultado consultando una tabla de distribución de chi-cuadrado.

8. El método de acuerdo con la reivindicación 6, en donde dicha determinación, basándose en cada valor esperado convergente, de la confianza de la mutación en el locus de interés para cada muestra que se ha de detectar comprende:

realizar una transformación regular en la escala de Phred sobre una probabilidad correspondiente a $\frac{f_1}{f_2 \min}$ para obtener una puntuación de calidad de Phred; y
determinar la puntuación de calidad de Phred como la confianza de la mutación en el locus de interés para cada muestra que se ha de detectar.

9. Un aparato de detección de mutaciones de bases basado en datos de secuenciación, comprendiendo el aparato de detección de mutaciones de bases:

un módulo de determinación de la frecuencia inicial configurado para determinar una frecuencia inicial de una frecuencia inicial de datos de secuenciación para una pluralidad de muestras que se han de detectar, correspondiente a una base específica en un locus de interés;
un módulo de cálculo de valores esperados configurado para calcular, basándose en la frecuencia inicial, un valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés;
un módulo de actualización configurado para actualizar, utilizando cada valor esperado, la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés;
un módulo de iteración configurado para: calcular además, utilizando la frecuencia inicial actualizada, el valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés, actualizar adicionalmente, utilizando cada nuevo valor esperado, la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés, y repetir la iteración anterior hasta que converja el valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; y
un módulo de determinación del tipo de mutación configurado para determinar, basándose en cada valor esperado convergente, un tipo de mutación de bases y una confianza de mutación en el locus de interés de cada una de la pluralidad de muestras que se han de detectar,
en donde la base específica comprende una base de adenina A, una base de timina T, una base de citosina C o una base de guanina G,
en donde el módulo de cálculo del valor esperado está configurado además para:

calcular el valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés de acuerdo con la siguiente ecuación:

$$p(\mathrm{b}_{i,j} \left| p_j, \mathrm{d}_i \right.) = \frac{p(\mathrm{b}_{i,j} \left| p_j \right.) p(\mathrm{d}_i \left| \mathrm{b}_{i,j} \right.)}{\sum_{\mathrm{b}_{i,j=0}}^{3} p(\mathrm{b}_{i,j} \left| p_j \right.) p(\mathrm{d}_i \left| \mathrm{b}_{i,j} \right.)}$$

donde $b_{i,j}$ representa que una muestra i que se ha de detectar es una base j específica en un locus de interés; p_j representa una frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica j en el locus de interés; d_i representa un conjunto de bases de secuencias de bases en el locus de interés cubierto por la muestra i que se ha de detectar; $p(b_{i,j}|p_j,d_i)$ representa un valor esperado de la muestra i que se ha de detectar correspondiente a la base específica j en el locus de interés; $p(b_{i,j}|p_j)$ representa una probabilidad a priori de que la muestra i que se ha de detectar corresponda a la base específica j en el locus de interés a una p_j dada; and $p(d_i|b_{i,j})$ representa una puntuación de calidad de las bases de las secuencias de bases en el locus de interés cubierto por la muestra i que se ha de detectar,

en donde el módulo de actualización está configurado además para:

actualizar, utilizando cada valor esperado, la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés de acuerdo con la siguiente ecuación:

$$\hat{p}_j = \frac{\sum_{i=1}^{N} p(b_{i,j} \mid p_j, d_i)}{n}$$

donde $\hat{p}_j$ representa una frecuencia inicial actualizada de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; $p(b_{i,j}|p_j,d_i)$ representa un valor esperado de la muestra i que se ha de detectar correspondiente a la base específica j en el locus de interés; N representa el número para la pluralidad de muestras que se han de detectar; y n representa el número total de bases de cuatro tipos de los datos de secuenciación para la pluralidad de muestras que se han de detectar, y

en donde el módulo de determinación del tipo de mutación se configura además para:

calcular, basándose en cada valor esperado convergente, una estimación de máxima verosimilitud para la pluralidad de muestras que se han de detectar pertenecientes a cada uno de los cuatro tipos de mutación de bases específicas en el locus de interés;
calcular un cociente de las estimaciones de máxima verosimilitud de dos tipos de mutación de bases específicas adyacentes;
procesar el cociente según una regla preestablecida para obtener una probabilidad correspondiente al cociente; y
cuando la probabilidad es inferior a un umbral establecido, determinar el tipo de mutación de base específica en el locus de interés de cada muestra que se ha de detectar como el tipo de mutación de base específica correspondiente a un denominador actual,
en donde los cuatro tipos de mutación de bases específicas comprenden la mutación monoalélica, mutación bialélica, mutación trialélica y mutación tetralélica.

10. Un medio de almacenamiento, que comprende una instrucción ejecutable por ordenador, en donde la instrucción ejecutable por ordenador, cuando es ejecutada por un procesador de un ordenador, implementa el método de detección de mutaciones de bases basado en datos de secuenciación de acuerdo con una cualquiera de las reivindicaciones 1 a 8.

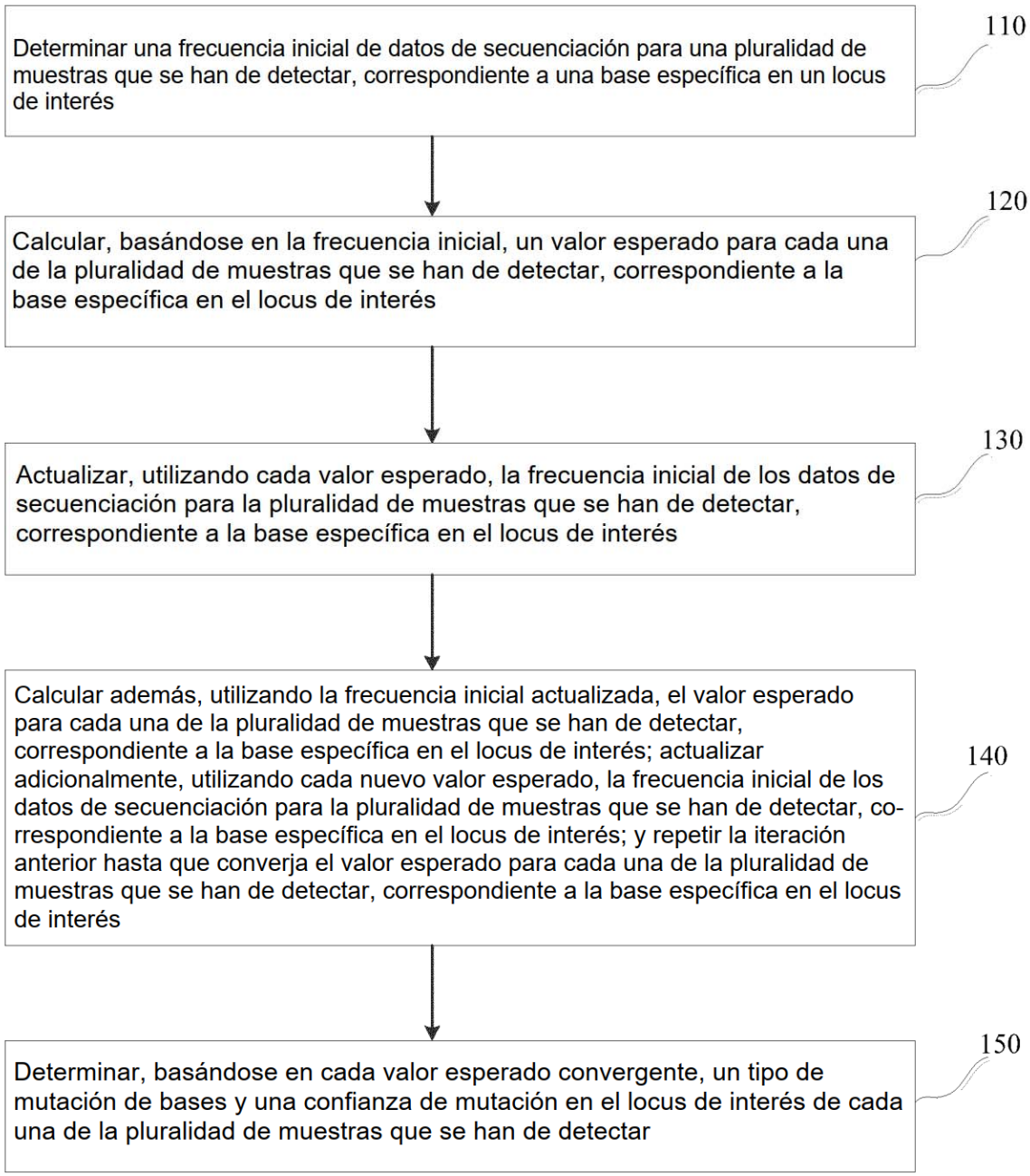

Determinar una frecuencia inicial de datos de secuenciación para una pluralidad de muestras que se han de detectar, correspondiente a una base específica en un locus de interés
110
Calcular, basándose en la frecuencia inicial, un valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés
120
Actualizar, utilizando cada valor esperado, la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés
130
Calcular además, utilizando la frecuencia inicial actualizada, el valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; actualizar adicionalmente, utilizando cada nuevo valor esperado, la frecuencia inicial de los datos de secuenciación para la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés; y repetir la iteración anterior hasta que converja el valor esperado para cada una de la pluralidad de muestras que se han de detectar, correspondiente a la base específica en el locus de interés
140
Determinar, basándose en cada valor esperado convergente, un tipo de mutación de bases y una confianza de mutación en el locus de interés de cada una de la pluralidad de muestras que se han de detectar
150

FIG. 1

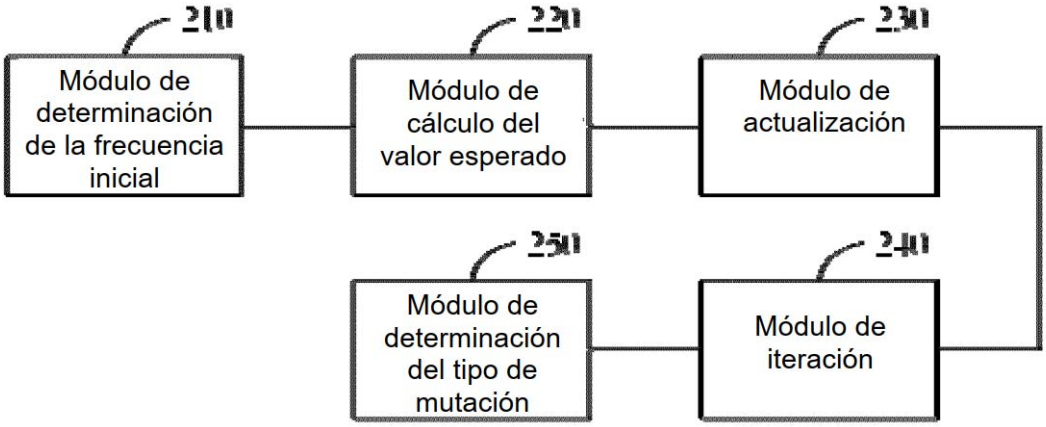

210
Módulo de determinación de la frecuencia inicial
220
Módulo de cálculo del valor esperado
230
Módulo de actualización
250
Módulo de determinación del tipo de mutación
240
Módulo de iteración

FIG. 2

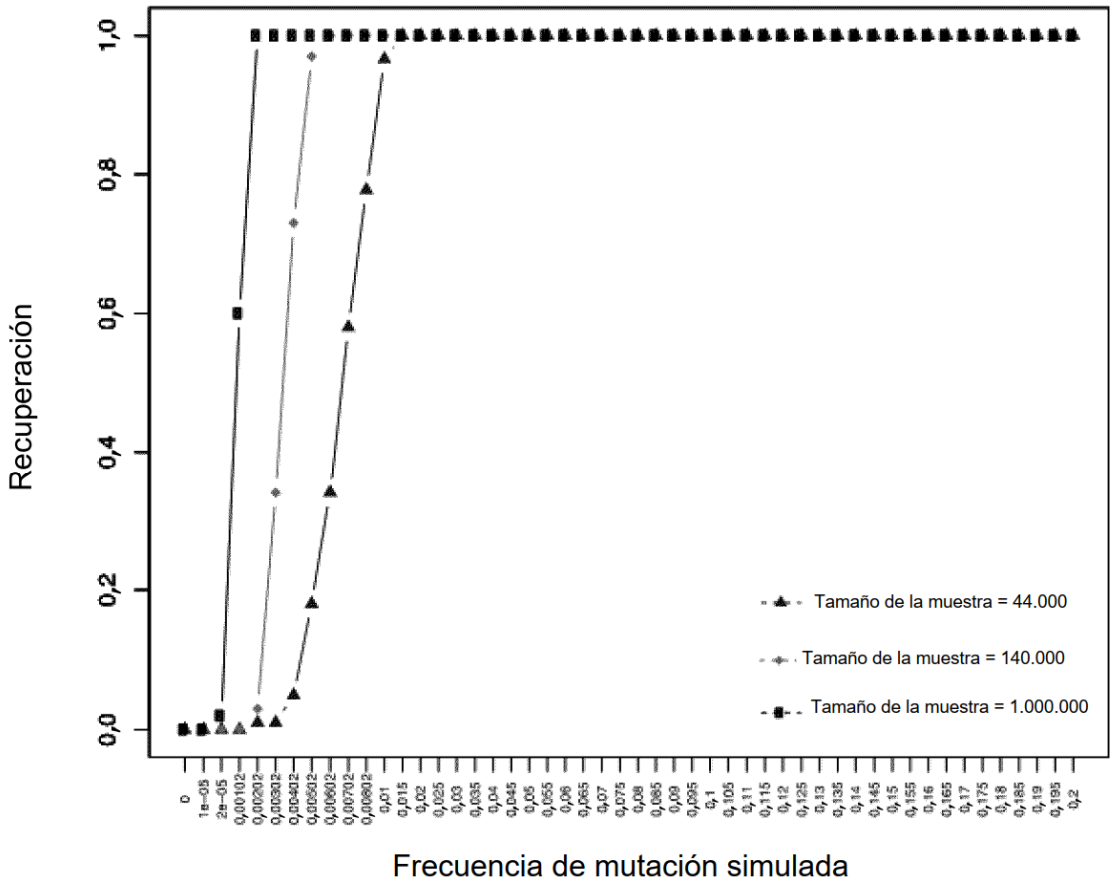

Recuperación
0,0
0,2
0,4
0,6
0,8
1,0
Tamaño de la muestra = 44.000
Tamaño de la muestra = 140.000
Tamaño de la muestra = 1.000.000
Frecuencia de mutación simulada

FIG. 3

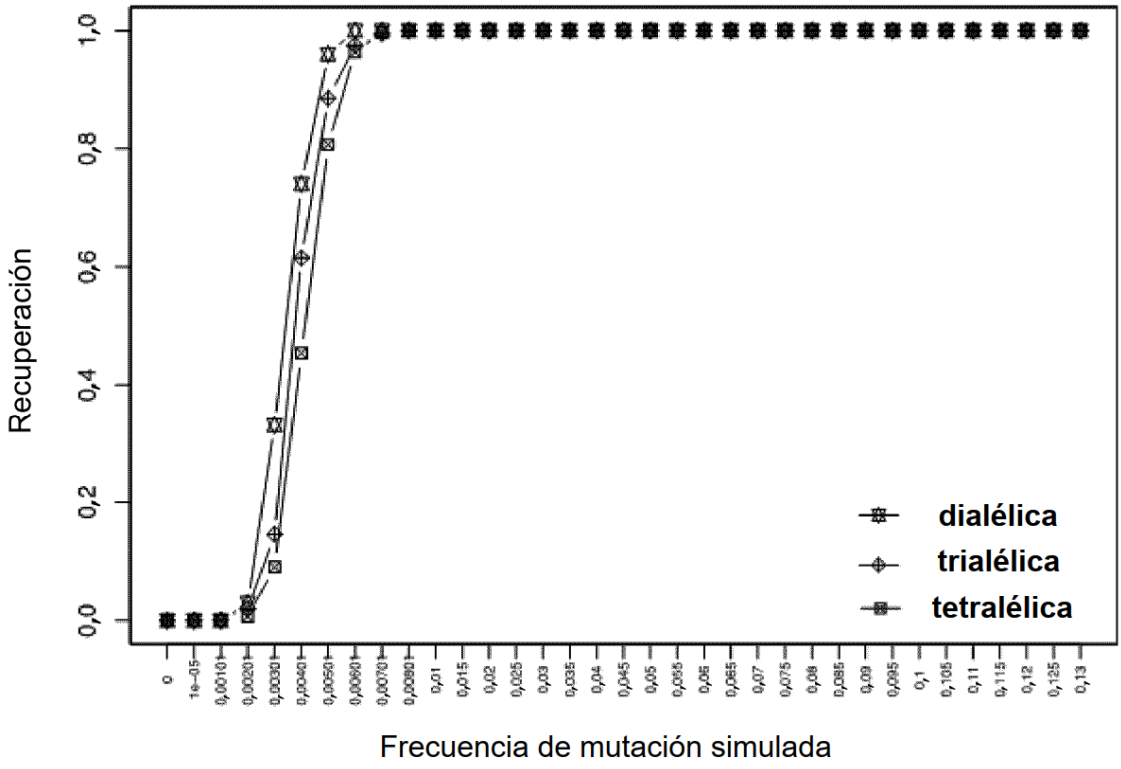

Recuperación
0,0
0,2
0,4
0,6
0,8
1,0
dialélica
trialélica
tetralélica
0
1e−05
0,00101
0,00201
0,00301
0,00401
0,00501
0,00601
0,00701
0,00801
0,01
0,015
0,02
0,025
0,03
0,035
0,04
0,045
0,05
0,055
0,06
0,065
0,07
0,075
0,08
0,085
0,09
0,095
0,1
0,105
0,11
0,115
0,12
0,125
0,13
Frecuencia de mutación simulada

FIG. 4

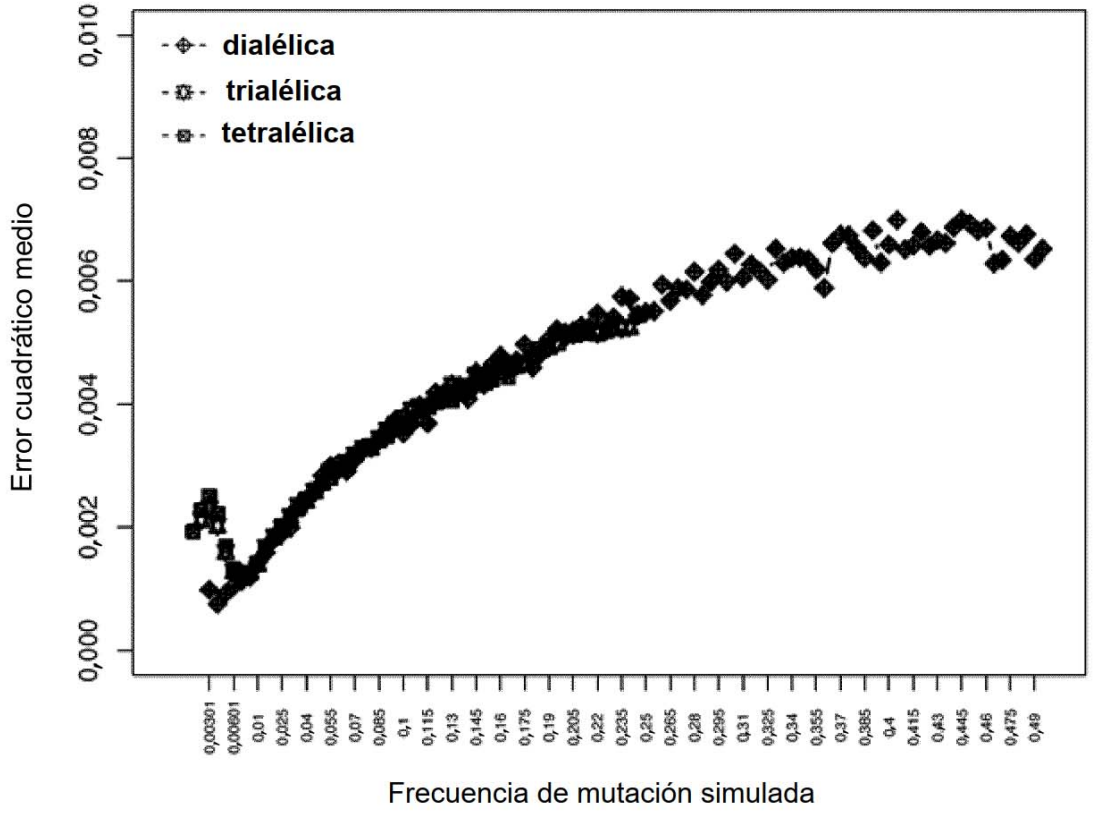

dialélica
trialélica
tetralélica
Error cuadrático medio
0,000
0,002
0,004
0,006
0,008
0,010
0,00301
0,00801
0,01
0,025
0,04
0,055
0,07
0,085
0,1
0,115
0,13
0,145
0,16
0,175
0,19
0,205
0,22
0,235
0,25
0,265
0,28
0,295
0,31
0,325
0,34
0,355
0,37
0,385
0,4
0,415
0,43
0,445
0,46
0,475
0,49
Frecuencia de mutación simulada

FIG. 5

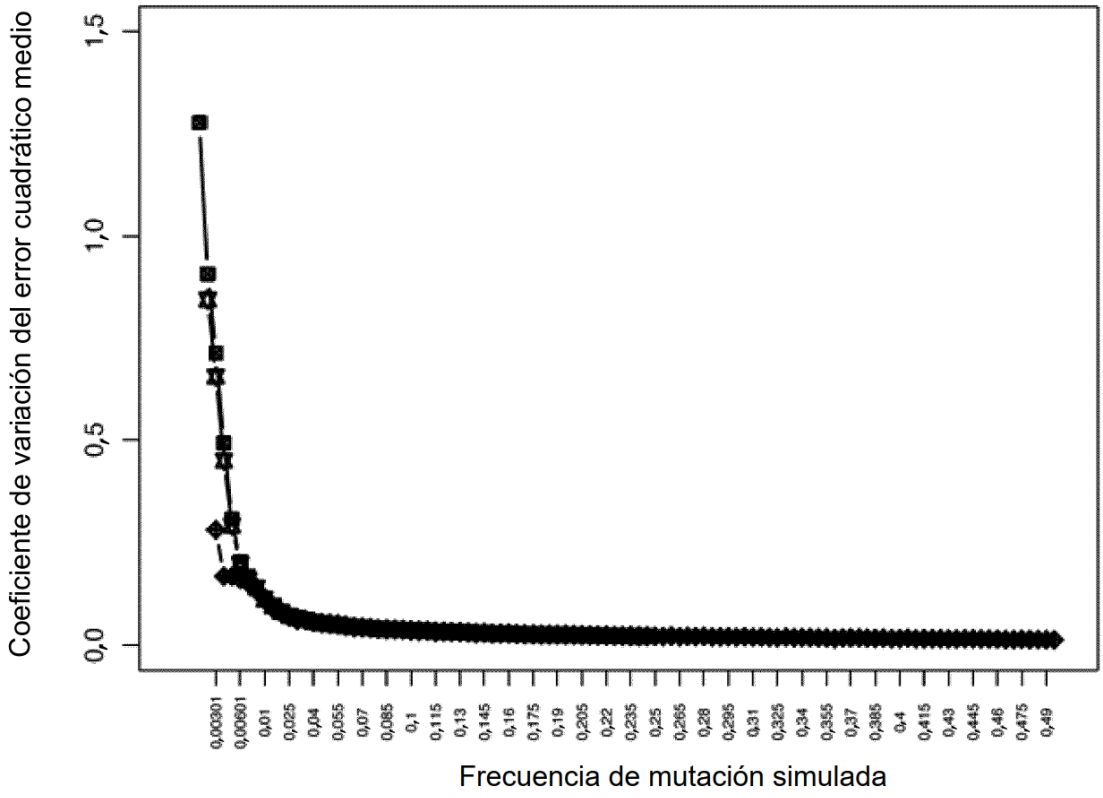

Coeficiente de variación del error cuadrático medio
0,0
0,5
1,0
1,5
0,00901
0,00601
0,01
0,025
0,04
0,055
0,07
0,085
0,1
0,115
0,13
0,145
0,16
0,175
0,19
0,205
0,22
0,235
0,25
0,265
0,28
0,295
0,31
0,325
0,34
0,355
0,37
0,385
0,4
0,415
0,43
0,445
0,46
0,475
0,49
Frecuencia de mutación simulada

FIG. 6