(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 923 142**

(51) Int. CI.:

G16B 30/10 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **14.06.2019 PCT/EP2019/065777**

(87) Fecha y número de publicación internacional: **19.12.2019 WO19238963**

(96) Fecha de presentación y número de la solicitud europea: **14.06.2019 E 19730781 (2)**

(97) Fecha y número de publicación de la concesión europea: **08.06.2022 EP 3807885**

(54) Título: **Métodos para detectar variantes en datos genómicos de secuenciación de próxima generación**

(30) Prioridad:

14.06.2018 EP 18177876

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
23.09.2022

(73) Titular/es:

SOPHIA GENETICS S.A. (100.0%)
Rue du Centre 172
1025 Saint Sulpice, CH

(72) Inventor/es:

XU, ZHENYU y
SONG, LIN

(74) Agente/Representante:

IZQUIERDO BLANCO, María Alicia

DESCRIPCIÓN

Métodos para detectar variantes en datos genómicos de secuenciación de próxima generación

CAMPO DE LA INVENCIÓN

[0001] Los métodos descritos en el presente documento se refieren al análisis genómico en general, y más específicamente a las aplicaciones de secuenciación de próxima generación.

ANTECEDENTES DE LA INVENCIÓN

Secuenciación de próxima generación

[0002] Las tecnologías de *secuenciación de próxima generación* (NGS) o *secuenciación masivamente paralela (MPS)* de alto rendimiento han reducido significativamente el coste de la secuenciación de ADN en la última década. NGS tiene una amplia aplicación en biología y cambió drásticamente la forma de investigación o metodologías de diagnóstico. Por ejemplo, el perfil de expresión de ARN o la secuenciación de ADN solo se pueden realizar con unos pocos genes con métodos tradicionales, como la PCR cuantitativa o la secuenciación de Sanger. Incluso con micromatrices, el perfilado de la expresión génica o la identificación de la mutación a nivel del genoma completo solo se puede implementar para organismos cuyo tamaño del genoma es relativamente pequeño. Con la tecnología NGS, la creación de perfiles de ARN o la secuenciación del genoma completo se ha convertido en una práctica rutinaria ahora en la investigación biológica. Por otro lado, debido al alto rendimiento de NGS, se han desarrollado métodos multiplexados no solo para secuenciar más regiones sino también para secuenciar más muestras. En comparación con la tecnología de secuenciación tradicional de Sanger, NGS permite la detección de mutaciones en muchas más muestras en diferentes genes en paralelo. Debido a sus superiores sobre el método de secuenciación tradicional, los secuenciadores NGS ahora están reemplazando a Sanger en el diagnóstico de rutina. En particular, las variaciones genómicas de individuos (línea germinal) o de tejidos cancerosos (somáticos) ahora se pueden analizar de forma rutinaria para una serie de aplicaciones médicas que van desde el diagnóstico de enfermedades genéticas hasta el ajuste fino de la farmacogenómica de la medicación en la práctica de la medicina de precisión. NGS consiste en procesar múltiples lecturas de secuencias de ADN fragmentadas, típicamente cortas (menos de 300 pares de bases de nucleótidos). Las lecturas resultantes pueden luego compararse con un genoma de referencia por medio de una serie de métodos bioinformáticos, para identificar pequeñas variantes como polimorfismos de nucleótido único (SNP) correspondientes a una sustitución de un solo nucleótido, así como inserciones y deleciones cortas (INDEL) de nucleótidos en la secuencia de ADN en comparación con su referencia.

Enriquecimiento dirigido

[0003] En algunas patologías, una variante genética específica se ha asociado con la enfermedad, como los genes BRCA1 y BRCA2 en ciertas formas de cánceres hereditarios de mama y ovario o el gen CFTR en la fibrosis quística. En lugar de secuenciar el genoma completo (WGS) a partir de una muestra individual, el análisis genómico puede centrarse en la región del genoma asociada con la enfermedad, *dirigiéndose,* con un conjunto de cebadores o sondas de ADN específicos de la región, y *enriqueciendo o amplificando,* por ejemplo. con PCR (reacción en cadena de la polimerasa), la muestra de ADN biológico específicamente para las subregiones correspondientes al gen a lo largo de la cadena de ADN. Se han desarrollado varios *ensayos de secuenciación de próxima generación* siguiendo esos principios como kits biológicos listos para usar, como por ejemplo los kits de ensayo Multiplicom MASTR™ o Illumina TruSeq® Amplicon para facilitar el diagnóstico basado en ADN con secuenciadores de próxima generación, como como por ejemplo el secuenciador Illumina MiSeq®, en la investigación médica y la práctica clínica.

[0004] El *enriquecimiento objetivo* puede lograrse a partir de una pequeña muestra de ADN por medio de hibridación basada en sonda (en matrices o en solución) o enriquecimiento de exón dirigido basado en PCR altamente multiplexada, de modo que se maximiza tanto la cobertura génica/profundidad de lectura como la especificidad de la amplificación (amplificación de la región derecha, medida mediante una mayor alineación con las regiones diana deseadas). Los ejemplos de sistemas de enriquecimiento de objetivos disponibles en el mercado incluyen Agilent SureSelect™ Target Enrichment System, Roche NimbleGen SeqCap EZ, Illumina Nextera Rapid Capture, Agilent Haloplex™ y Multiplicom MASTR™.

[0005] Con el fin de maximizar el uso del secuenciador NGS de procesamiento paralelo masivo, se multiplexa una cantidad de muestras en el experimento NGS específico; por lo tanto, se puede ingresar simultáneamente un grupo de 48 o más muestras de enriquecimiento objetivo al secuenciador Illumina MiSeq para instancia. Los datos de secuenciación sin procesar del secuenciador NGS pueden luego analizarse para identificar subsecuencias específicas, por ejemplo, mediante la alineación con un genoma de referencia. Como resultado, la amplificación puede producir más de mil lecturas para un amplicón dado en una muestra de paciente.

Automatización del flujo de trabajo de secuenciación de próxima generación

[0006] La secuenciación de próxima generación (NGS) permite, en particular, detectar e informar pequeños cambios en la secuencia de ADN, como polimorfismos de un solo nucleótido (SNP), inserciones o deleciones (INDEL), en comparación

con el genoma de referencia., a través de métodos bioinformáticos como secuenciación de alineación de lectura, llamada de variantes y anotación de variantes. Los flujos de trabajo de NGS se refieren a la configuración y combinación de dichos métodos en una aplicación de análisis genómico de extremo a extremo. En la práctica de la investigación genómica, los flujos de trabajo de NGS a menudo se configuran y optimizan manualmente utilizando, por ejemplo, scripts dedicados en un sistema operativo UNIX, plataformas dedicadas que incluyen una representación gráfica de canalización como el proyecto Galaxy y/o una combinación de los mismos. A medida que se desarrolla la práctica clínica, es posible que los flujos de trabajo de NGS ya no se configuren experimentalmente caso por caso, sino que se integren en las ofertas de SaaS (software como servicio), PaaS (plataforma como servicio) o IaaS (infraestructura como servicio). por terceros proveedores. En ese contexto, una mayor automatización de los flujos de trabajo de NGS es clave para facilitar la integración rutinaria de esos servicios en la práctica clínica.

Optimización del flujo de trabajo de secuenciación de próxima generación

[0007] Si bien los métodos de secuenciación de próxima generación han demostrado ser más eficientes que la secuenciación de Sanger tradicional en la detección de SNP e INDEL, su especificidad (tasa de detección positiva verdadera para una variante genómica dada) y sensibilidad (tasa de detección positiva verdadera para una variante genómica dada) exclusión negativa para una variante genómica determinada) puede mejorarse aún más en la práctica clínica. La especificidad y la sensibilidad del análisis genómico NGS pueden verse afectadas por una serie de factores:

- Sesgos introducidos por la tecnología de secuenciación, por ejemplo debido a:

    o Longitud de las lecturas en relación con la longitud de los fragmentos;
    o Número demasiado pequeño de lecturas (profundidad de lectura);
    o Errores o bases de baja calidad introducidos durante la secuenciación;

    Dificultades inherentes al contar tramos de homopolímero, en particular con pirosecuenciación (como en las plataformas Roche 454) o secuenciación de semiconductores (como en las plataformas Ion Torrent, como describe, por ejemplo, Rothberg en la solicitud de patente de EE. UU. 2009/0026082), lo que genera errores de inserción y eliminación;
- Sesgos introducidos por la tecnología de enriquecimiento de ADN, por ejemplo debido a:

    o La unión no específica de los cebadores o las sondas, por ejemplo debido al almacenamiento del ensayo a baja temperatura durante demasiado tiempo, o debido a una cantidad demasiado pequeña de ADN en la muestra;
    o Introducción de errores de secuencia causados por amplificación y ciclos de PCR imperfectos, por ejemplo, debido a cambios de temperatura;
    o Diseño subóptimo de las sondas o cebadores. Por ejemplo, las mutaciones pueden caer dentro de las regiones de las sondas o cebadores;
    o Limitaciones del método de enriquecimiento. Por ejemplo, la eliminación larga puede abarcar la región amplificada;
    o Contaminación cruzada de conjuntos de datos, pérdida de lectura y disminución de la calidad de lectura debido al etiquetado de fragmentos con códigos de barras, adaptadores y varias etiquetas de secuencia predefinidas;
    o Lecturas quiméricas en lecturas de dos extremos de inserción larga.

- Sesgos introducidos por la propia muestra, por ejemplo debido a:

    o Características somáticas, en particular en el diagnóstico de cáncer basado en la secuenciación de la muestra del tumor;
    o El tipo de muestra biológica, por ejemplo, sangre, orina, saliva y los problemas de preparación de muestras asociados, por ejemplo, que causan la degradación del ADN, la contaminación con ADN extraño o una entrada de ADN demasiado baja.

- Sesgos introducidos por la estructura de datos genómicos de ciertas regiones específicamente, por ejemplo debido a:

    o Alta proporción de contenido de GC en la región de interés;
    o Presencia de homopolímeros y/o heteropolímeros, es decir, repeticiones de secuencias genómicas parciales de uno o más nucleótidos en ciertas regiones, lo que provoca ambigüedades en la alineación inicial y posiblemente errores de secuenciación inherentes, en particular con las tecnologías de secuenciador Roche 454 e Ion Torrent;
    o Presencia de regiones homólogas y de baja complejidad;
    o Presencia de pseudogenes no funcionales que pueden confundirse con genes funcionales, en particular en regiones genómicas de alta repetición del genoma humano cuando los fragmentos de ADN no son lo suficientemente largos en comparación con la longitud de lectura.

[0008] Esto limita el despliegue eficiente de NGS en aplicaciones de análisis genómico de rutina, ya que un flujo de trabajo de análisis de datos genómicos diferente debe organizarse y configurarse manualmente con diferentes conjuntos de

parámetros por parte de personal altamente especializado para cada aplicación para cumplir con las expectativas clínicas en términos de especificidad y sensibilidad. La automatización de los flujos de trabajo de procesamiento de datos genómicos es particularmente desafiante, ya que los flujos de trabajo deben tener en cuenta las estructuras de sesgos de datos específicos introducidos por los procesos biológicos NGS aguas arriba, por un lado, y los datos genómicos inherentes a la solicitud actual por otro lado. En la implementación inicial de las pruebas genómicas, una cantidad limitada de pruebas y configuraciones fueron procesadas por plataformas dedicadas, que podían ser configuradas, configuradas y mantenidas manualmente por personal especializado altamente calificado. Este enfoque es costoso y no escala bien, ya que se deben realizar más y más pruebas en la operación diaria mediante una única plataforma de análisis genómico multipropósito.

[0009] En términos de automatización del análisis NGS, se debe prestar especial atención a las dificultades inherentes a la caracterización de las variantes indel en las regiones de homopolímero y/o heteropolímero del genoma humano de referencia, en particular cuando el laboratorio emplea pirosecuenciación (como en Roche 454) o secuenciación de semiconductores (como en las plataformas Ion Torrent). La caracterización errónea de algunas variantes de homopolímeros o heteropolímeros puede resultar en la detección de falsos positivos de ciertos rasgos y enfermedades en una diversidad de aplicaciones de diagnóstico, por ejemplo, en base a algunas variaciones genéticas en genes relacionados con el cáncer, como lo destacan, por ejemplo, Singh et al. en "Clinical validation of a Next-Generation Sequencing Screen for Mutational Hotspots in 46 Cancer-Related Genes", The Journal of Molecular Diagnostics, Vol.15, No.5, septiembre de 2013. Para superar esta limitación, Singh et al. propuso excluir variantes que tienen una frecuencia de población de más del 20 % de las muestras secuenciadas como probablemente sesgadas por el proceso de secuenciación. En la práctica actual de NGS, la plataforma del analizador genómico de los solicitantes, Sophia DDM®, puede configurarse para ignorar los INDEL ubicados en regiones de homopolímero de más de 10 pb, como se describe, por ejemplo, en la propuesta para la evaluación de la prueba genética para el linfedema primario con un panel de 15 genes por London South West RGC St George's (https://ukgtn.nhs.uk/uploads/tx ukgtn/Primary Lymphoedema 15 GP GD Sept 2017.pdf - Fecha de aprobación enero 2018) o "Performance characteristics - BRCA MASTR Dx with drMID Dx for Illumina NGS Systems", rev. Julio de 2017 - https://www.agilent.com/cs/library/datasheets/public/Performance %20characteristics %20BRCA %20MASTR % 20Dx %205991-8424ENE.pdf

[0010] Otras soluciones bioinformáticas como el módulo Sequence Pilot SeqNext de JSI Medical Systems GmbH, Kippenheim, Alemania puede configurarse para llamar a variantes en homopolímeros de al menos 6 pb solo si la frecuencia de la variante supera el 20 % de las lecturas, como describen Nunziato et al. en "Fast Detection of a BRCA2 Large Genomic Duplication by Next Generation Sequencing as a Single Procedure: A Case Report", Intl J Mol Sci v.18(11), noviembre de 2017. Por lo tanto, la práctica actual en las pruebas de panel de genes NGS consiste en ignorar o parametrizar cuidadosamente la detección de variantes en regiones de homopolímero o heteropolímero del genoma humano cuando se sabe que la plataforma NGS las sesga. Sin embargo, estos enfoques pueden dar lugar a caracterizaciones negativas falsas.

[0011] En "Eficiencia y confiabilidad mejoradas del análisis de datos de secuenciación de amplicón de NGS para procedimientos de diagnóstico genético usando el software AGSA", Biomed research international, Vol. 2016, Art ID 5623089, Poulet et al. identifica las limitaciones de diferentes flujos de trabajo de análisis de software como CORAL, HECTOR, AmpliconNoise para la detección de variantes del gen BRCA asociadas con el riesgo familiar de cáncer de mama y de ovario y propone un método de mejora (implementado por los autores como el software AGSA) basado en el análisis del SFF archivo, recopilando el valor del diagrama de flujo para cada lectura de una muestra de interés y derivando una imagen de histograma que puede ser inspeccionada por el usuario final. Para una inserción o eliminación heterocigótica, la distribución de los valores de lectura se divide en dos poblaciones, lo que muestra que algunas lecturas (de un alelo) tienen *n* bases idénticas mientras que otras (del otro alelo) tienen *n* + 1 (inserción) o *n* - 1 (supresión) bases idénticas. Por el contrario, en el caso de artefactos de sesgo de secuenciación, se observa una sola población entre *n* y *n* + 1 (o *n* - 1) picos en el histograma. En el caso de la variación homocigótica, una sola población se centra en *n* + 1 (o *n* - 1), lo que demuestra que todas las lecturas tienen el mismo número de bases en el homopolímero y que este número es diferente del tipo salvaje.

[0012] Poulet et al. también sugiere que la distribución mono o bimodal de los valores del histograma también puede evaluarse estadísticamente, pero no revela métodos para lograrlo. Además, su enfoque requiere el análisis de los archivos de flujo directamente desde el secuenciador, lo que complica el diseño y la implementación de una plataforma de flujo de trabajo de analizador genómico automatizado en un entorno informático en red, en particular cuando la solución de análisis genómico se implementa independientemente del equipo de laboratorio, como es por ejemplo la plataforma Sophia DDM (Data Driven Medicine) SAAS (Software As A Service).

[0013] Similar a la propuesta de Poulet et al., también se describió en la solicitud de patente de EE. UU. 2014/0052381 de Utirametur et al. un método estadístico para mejor detectar la inserción y eliminación en una región homopolimérica y detectar la correspondiente heterocigosidad. Observaron que en un flujo de trabajo del analizador genómico NGS, la alineación de lectura no es necesariamente correcta, pero puede ser posible determinar la heterocigosidad a partir de la distribución de residuos de llamada de base en función de los valores medidos y predichos por el modelo en las regiones de homopolímero mediante el uso de un enfoque de detección de picos bayesiano y modelo de mejor ajuste, ya que las regiones homocigotas tienden a tener una distribución unimodal mientras que las regiones heterocigotas tienden a tener una distribución unimodal. A partir del modelo de mejor ajuste, también es posible derivar el valor de longitud del

homopolímero para ambos alelos en el caso de homocigosis (distribución unimodal), o dos valores de longitud de homopolímero diferentes, uno para cada alelo, en el caso de heterocigosis (distribución bimodal). Si bien este método puede facilitar la identificación de la longitud de las regiones cortas del homopolímero, ya que las densidades del espacio de flujo asociadas exhiben claramente un valor máximo, observamos que es significativamente más difícil clasificar los homopolímeros y los heteropolímeros más largos.

[0014] Por lo tanto, existe la necesidad de una mejor solución para automatizar los flujos de trabajo de llamada de variantes de procesamiento de datos genómicos para aplicaciones médicas basadas en datos, de modo que la misma plataforma de procesamiento de datos genómicos pueda operar en una diversidad de datos genómicos que pueden generarse a partir de diferentes configuraciones de laboratorio de secuenciación de próxima generación mientras se optimiza la especificidad y la sensibilidad de los resultados de búsqueda de variantes para mejorar la investigación y la práctica clínica en comparación con los métodos de estado de la técnica en contextos de datos genómicos que implican la caracterización desafiante de variantes de patrones repetidos de homopolímero y/o heteropolímero.

BREVE RESUMEN

[0015] Se propone un método para detectar e informar, con un procesador, una variante como la repetición de al menos dos patrones de nucleótidos en la secuencia genómica de una muestra de un paciente, comprendiendo el método:

(a) identificar un patrón repetido de referencia $P_{ref} = N^*I$ como la repetición de $I(I>=2)$ patrones genómicos N en una región genómica de una secuencia de referencia del genoma humano;
(b) obtener, con un secuenciador de próxima generación, n conjuntos de pacientes de lecturas de datos de secuenciación de próxima generación que cubren la región genómica del patrón repetido de referencia $S= \{S_1, S_2,.., S_i,.., S_n\}$ de un conjunto de n muestras de pacientes genómicos enriquecidos, estando asociado cada conjunto S_i con una muestra de paciente, siendo el número n de muestras de pacientes genómicos enriquecidos al menos 4;
(c) para cada muestra de paciente i en el conjunto S de muestras de pacientes, medir la distribución P_i de la longitud del patrón repetido en el conjunto de lecturas de secuenciación de próxima generación S_i;
(d) para un posible par de muestras de pacientes i y j, $j>i$, estimar un modelo de mejor ajuste $[V_{ij}^1 | V_{ij}^2]$ de las dos variantes alélicas para la muestra i en relación con la muestra j, con un nivel de confianza L_{ij};
(e) para cada triplete posible de muestras de pacientes i, $j>i$, $k>j$, comparar sus respectivos modelos de mejor ajuste $[V_{ij}^1 | V_{ij}^2]$ $[V_{jk}^1 | V_{jk}^2]$, $[V_{ik}^1 | V_{ik}^2]$, agrupar los modelos de mejor ajuste coincidentes en grupos de modelos variantes de mejor ajuste con un mayor nivel de confianza, e iterar la comparación hasta que se formen grupos estables de modelos variantes de mejor ajuste;
(f) identificar el grupo más probable que lleva la variante de tipo salvaje;
(g) para cada muestra en el grupo que lleva la variante de tipo salvaje, reportar la variante de la muestra como el patrón de repetición de referencia de tipo salvaje $P_{ref} = N^*I$.
(h) para cada muestra del grupo que lleva la variante de tipo salvaje, eliminar el sesgo del modelo de variante de mejor ajuste del grupo que comprende esta muestra en función del modelo de variante de mejor ajuste para el grupo de tipo salvaje identificado, y reportar la muestra variante como la variante no sesgada.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

[0016]

La FIG. 1 representa un sistema de secuenciación de próxima generación de la técnica anterior.

La FIG. 2 muestra el diagrama de flujo de un flujo de trabajo de análisis genómico de secuenciación de próxima generación.

La FIG. 3 ilustra la distribución de probabilidad de la longitud relativa esperada de una variante del patrón de repetición en relación con la referencia del genoma humano (centrado en 0) sin error de experimento, respectivamente para muestras sin mutación (tabla superior) y para muestras con una deleción heterocigota en el alelo 1 (tabla inferior).

La FIG. 4 ilustra la distribución de probabilidad de la longitud relativa medida de una variante del patrón de repetición en relación con la referencia del genoma humano (centrado en 0) como se puede medir sujeto a error experimental, respectivamente para muestras que no tienen mutación (tabla superior) y para muestras con una deleción heterocigótica en el alelo 1 (tabla inferior).

La FIG. 5a) muestra la representación gráfica de todas las posibles distribuciones de probabilidad esperadas de la longitud relativa de los patrones repetidos para varios escenarios de eliminación e inserción que pueden derivarse de la distribución de probabilidad de la longitud relativa en los datos de control sin mutación pero sujetos a errores experimentales sesgos, y

La FIG. 5b) muestra la representación gráfica de una distribución de probabilidad medida para los datos del

paciente, para que coincida con el escenario de distribución de probabilidad esperado más cercano.

La FIG. 6 muestra un ejemplo de distribuciones de probabilidad medidas de la longitud de los patrones repetidos para un par de muestras 1 y 2, así como las distribuciones de probabilidad esperadas de la longitud de los patrones repetidos para cada muestra en relación con la otra supuesta como referencia.

La FIG. 7 muestra otro ejemplo de distribuciones de probabilidad medidas de la longitud de patrones repetidos para un par de muestras 1 y 2, así como las distribuciones de probabilidad esperadas de la longitud de patrones repetidos para cada muestra en relación con la otra supuesta como referencia.

La FIG. 8 muestra otro ejemplo de distribuciones de probabilidad medidas de la longitud de los patrones repetidos para un par de muestras 1 y 2, así como las distribuciones de probabilidad esperadas de la longitud de los patrones repetidos para cada muestra en relación con la otra supuesta como referencia.

La FIG. 9 ilustra un posible diagrama de flujo general de un método refinado de llamada de variantes de acuerdo con la presente descripción, que permite identificar con precisión variantes de patrones repetidos con mayor sensibilidad y especificidad.

La FIG. 10 muestra una tabla de ejemplo de emparejamiento de 8 muestras respectivamente entre sí asumidas como referencia, donde las muestras S1, S2 y S3 se emparejan entre sí como si no tuvieran ninguna mutación (0/0).

La FIG. 11 muestra otra tabla de ejemplo de emparejamiento de 8 muestras respectivamente entre sí asumidas como referencia, donde las muestras S2, S7 y S8 solo marchan parcialmente entre sí, S2 frente a S7 y S7 frente a S8 sin ninguna mutación (0/0), pero S2 vs S8 portador de una mutación heterocigota (-1/0).

La FIG. 12 muestra otra tabla de ejemplo de coincidencia de 8 muestras respectivamente entre sí asumidas como referencia, donde las muestras S1, S2 y S6 no coinciden entre sí y no se puede encontrar una coincidencia relevante entre S2 y S6 (NA/NA).

La FIG. 13 muestra otra tabla de ejemplo de emparejamiento de 8 muestras respectivamente entre sí asumidas como referencia, después de iterar el método de emparejamiento propuesto.

La FIG. 14 ilustra un posible agrupamiento de muestras en una tabla de ejemplo de emparejamiento de 8 muestras respectivamente entre sí asumidas como referencia.

La FIG. 15 ilustra el diagrama de flujo lógico del método propuesto para identificar el grupo de muestras correspondiente a la referencia del genoma humano de tipo salvaje en el grupo de muestras.

La FIG. 16 muestra la cobertura de datos de NGS para un patrón de homopolímero repetido ejemplar en el gen CHEK2 de un grupo experimental que comprende una mezcla de muestras de pacientes mutados y muestras de tipo salvaje, y la FIG. 17 muestra la distribución medida correspondiente de las longitudes del patrón.

La FIG. 18 muestra la cobertura de datos de NGS para un patrón de homopolímero repetido ejemplar en el gen RAD54L de un grupo de experimentos que comprende una mezcla de muestras de pacientes mutados y muestras de tipo salvaje, y la FIG. 19 muestra la distribución medida correspondiente de las longitudes del patrón.

La FIG. 20 muestra un diagrama de la cobertura de datos de NGS para un patrón de homopolímero repetido ejemplar en el gen ATM de un grupo de experimentos que comprende una mezcla de muestras de pacientes mutados y muestras de tipo salvaje, y la FIG. 21 muestra la distribución medida correspondiente de las longitudes del patrón.

La FIG. 22 muestra un diagrama de la cobertura de datos de NGS para un patrón de heteropolímero repetido ejemplar en el gen ATM de un grupo experimental que comprende una mezcla de muestras de pacientes con diferentes mutaciones, y la FIG. 23 muestra la distribución medida correspondiente de las longitudes del patrón.

DESCRIPCIÓN DETALLADA

Sistema de análisis de secuenciación de próxima generación

[0017] La FIG. 1 muestra un sistema de análisis genómico ejemplar que comprende un ensayo de enriquecimiento de ADN 100, un secuenciador 110 de próxima generación y un analizador de datos genómicos 120.

[0018] En un laboratorio de NGS, una reserva de muestras de ADN se procesa mediante el ensayo de enriquecimiento de ADN 100 para generar una biblioteca de amplicones agrupados (para el enriquecimiento basado en amplicones) o fragmentos (para el enriquecimiento basado en sondas) como entrada de fragmentos de ADN al secuenciador de próxima

generación 110, cada conjunto de amplicones/fragmentos correspondiente a una muestra diferente. El número de amplicones/fragmentos depende de la aplicación. En algunos experimentos de análisis genómico, el enriquecimiento de objetivos puede requerir 150 cebadores para enriquecer 75 regiones diferentes que se seleccionarán fuera del genoma de la muestra, lo que da como resultado un conjunto de 75 amplicones para cada muestra. El número de muestras también puede adaptarse a la capacidad de procesamiento paralelo de secuenciador de secuenciación de próxima generación 110, por ejemplo, 48 muestras en forma de una biblioteca de amplicones combinados pueden secuenciarse en paralelo mediante un secuenciador Illumina MiSeq. Se pueden utilizar otras tecnologías de secuenciador NGS, como por ejemplo los secuenciadores Roche 454™ GS Junior o GS FLX, Illumina MiSeq® o Life Technologies Ion PGM™.

[0019] El secuenciador 110 de próxima generación analiza las muestras de entrada y genera lecturas de secuencia en un formato de archivo legible por computadora que representa datos de secuenciación NGS sin procesar. Dependiendo de la tecnología NGS, el secuenciador NGS 110 puede generar uno o más archivos. En algunas formas de realización, por ejemplo con secuenciadores Illumina, el formato de archivo FASTQ se puede usar con dos archivos diferentes para lecturas directas e inversas o como un archivo único unido. Este archivo de texto generalmente comienza con un encabezado de secuencia marcado con un carácter de inicio '@', seguido de una línea de información de secuencia representada como una cadena de caracteres de nucleótidos 'A', 'T', 'C', 'G', luego por un encabezado de calidad marcado con un carácter de inicio '+', seguido de una línea de métricas de calidad, una puntuación de calidad que coincide con cada lectura de nucleótidos. El formato de las métricas de calidad para cada nucleótido en la cadena de información de la secuencia puede depender del secuenciador. Algunos secuenciadores heredados generan los datos de secuenciación sin procesar en el formato de archivo binario SFF (formato de diagrama de flujo estándar), que comprende un encabezado informativo y los datos leídos. También son posibles otras formas de realización, por ejemplo, algunos secuenciadores Roche heredados generan múltiples archivos FASTQ para el análisis de un solo paciente, mientras que otros secuenciadores, por ejemplo, los secuenciadores Ion Torrent PGM, han migrado al formato de archivo BAM comprimido sin mapear, como se puede reconocer en la extensión de archivo.*basecaller.bam.* Como saben los expertos en la técnica de los sistemas de comunicación, el laboratorio opera una infraestructura informática para almacenar el archivo de datos de secuenciación NGS sin procesar resultante en un biobanco de laboratorio. La infraestructura informática del laboratorio se conecta, con credenciales de autenticación, a través de una red de comunicación, al analizador de datos genómicos 120 y transmite una solicitud de análisis genómico que comprende el archivo de secuenciación NGS sin procesar al analizador de datos genómicos 120.

[0020] El sistema informático del analizador de datos genómicos 120 (también "sistema" en este documento) 120 está programado o configurado de otro modo para implementar diferentes métodos de análisis de datos genómicos, como recibir y/o combinar datos de secuenciación y/o anotar datos de secuenciación.

[0021] El analizador de datos genómicos 120 puede ser un sistema informático o parte de un sistema informático que incluye una unidad central de procesamiento (CPU, "procesador" o "procesador informático" en este documento), memoria como RAM y unidades de almacenamiento como un disco duro e interfaces de comunicación para comunicarse con otros sistemas informáticos a través de una red de comunicación, por ejemplo, Internet o una red local. Los ejemplos de sistemas informáticos, entornos y/o configuraciones de analizadores de datos genómicos incluyen, entre otros, sistemas informáticos personales, sistemas informáticos de servidor, clientes ligeros, clientes pesados, dispositivos de mano o portátiles, sistemas multiprocesadores, sistemas basados en microprocesadores, conjuntos de cajas superiores, electrónica de consumo programable, PC de red, sistemas de minicomputadoras, sistemas de computadora central y similares. En algunas formas de realización, el sistema informático puede comprender uno o más servidores informáticos, que son operativos con muchos otros sistemas informáticos de propósito general o especial y pueden permitir la computación distribuida, como la computación en la nube, por ejemplo en una granja de datos genómicos. En algunas formas de realización, el analizador de datos genómicos 120 puede integrarse en un sistema paralelo masivo. En algunas formas de realización, el analizador de datos genómicos 120 puede integrarse directamente en un sistema de secuenciación de próxima generación.

[0022] El sistema informático del analizador de datos genómicos 120 puede adaptarse en el contexto general de las instrucciones ejecutables del sistema informático, tales como módulos de programa, que se ejecutan mediante un sistema informático. En general, los módulos de programa pueden incluir rutinas, programas, objetos, componentes, lógica, estructuras de datos, etc., que realizan tareas particulares o implementan tipos de datos abstractos particulares. Como es bien sabido por los expertos en la técnica de la programación informática, los módulos de programa pueden utilizar funciones nativas del sistema operativo y/o del sistema de archivos, aplicaciones independientes; plugins de navegador o aplicación, applets, etc.; bibliotecas comerciales o de código abierto y/o herramientas de biblioteca que pueden programarse en Python, Biopython, C/C++ u otros lenguajes de programación; scripts personalizados, como scripts Perl o Bioperl.

[0023] Las instrucciones pueden ejecutarse en entornos informáticos en la nube distribuidos donde las tareas se realizan mediante dispositivos de procesamiento remoto que están vinculados a través de una red de comunicaciones. En un entorno de computación en la nube distribuido, los módulos de programa pueden ubicarse en medios de almacenamiento de sistemas informáticos tanto locales como remotos, incluidos los dispositivos de almacenamiento de memoria.

[0024] Como se ilustra en la FIG. 1, el analizador de datos genómicos 120 puede comprender un módulo de alineación de secuencias 121, que compara los datos de secuenciación NGS sin procesar con un genoma de referencia. El módulo

de alineación de secuencias 121 puede configurarse para ejecutar diferentes algoritmos de alineación. Se pueden usar algoritmos estándar de alineación de datos sin procesar como Bowtie2 o BWA que se han optimizado para un procesamiento rápido de numerosas lecturas de secuenciación de datos genómicos, pero también son posibles otras formas de realización. Los resultados de la alineación se pueden representar como uno o varios archivos en formato BAM o SAM, como saben los expertos en bioinformática, pero también se pueden usar otros formatos, por ejemplo, formatos comprimidos o formatos optimizados para el cifrado de conservación del orden, dependiendo de los requisitos del analizador de datos genómicos 120 para la optimización del almacenamiento y/o la aplicación de la privacidad de los datos genómicos.

[0025] Los datos de alineación resultantes pueden ser filtrados y analizados adicionalmente por un módulo de llamada de variantes 122 para recuperar información de variantes tales como polimorfismos SNP e INDEL. El módulo de llamadas de variantes 122 puede configurarse para ejecutar diferentes algoritmos de llamadas de variantes. La información de la variante detectada resultante puede luego ser emitida por el módulo analizador de datos genómicos 120 como un informe de variante genómica para su posterior procesamiento por parte del usuario final, por ejemplo con una herramienta de visualización, y/o por un módulo adicional de procesamiento de anotaciones de variantes (no representado).

[0026] El analizador de datos genómicos 120 puede adaptarse para detectar automáticamente, con un procesador, un conjunto de características que determinan de manera única los datos de secuenciación de entrada y los datos genéticos correspondientes, el contexto de enriquecimiento de ADN, como el tipo de muestra o las características del proceso de laboratorio, la tecnología de enriquecimiento de ADN, como el kit de objetivo de enriquecimiento dirigido o las características del ensayo de sonda de captura, y/o la tecnología de secuenciación NGS. Como será evidente para los expertos en la técnica de la secuenciación de próxima generación, estas características experimentales pueden causar sesgos específicos en la alineación de la secuencia y/o en los resultados de llamada de variantes.

[0027] El sistema analizador de datos genómicos 120 propuesto puede así servir a las solicitudes de análisis genómico de secuenciación de próxima generación de diferentes laboratorios que están operando de forma independiente diferentes tecnologías de secuenciador y diferentes tecnologías de enriquecimiento de ADN en diferentes muestras para diferentes genes. El sistema analizador de datos genómicos 120 propuesto puede detectar automáticamente un conjunto de características a partir de los datos de entrada y las solicitudes recibidas del laboratorio y puede adaptar la configuración del módulo de alineación de secuencias 121 y el módulo de llamada de variantes 122 en consecuencia, sin requerir una configuración manual costosa y que requiere mucho tiempo para minimizar los sesgos de datos posiblemente inducidos por cada flujo de trabajo biológico diferente. Como será evidente para los expertos en la materia, puede haber docenas o incluso cientos de configuraciones de laboratorio clínico diferentes para múltiples laboratorios de abastecimiento que operen con el mismo analizador genómico 120, y es probable que este número y diversidad aumente aún más con el despliegue de más tecnologías y ensayos a medida que se desarrolla la práctica clínica de medicina personalizada basada en NGS.

[0028] Dependiendo de las características detectadas del experimento genómico, el analizador de datos genómicos 120 puede configurar el módulo de alineación de secuencias 121 para operar pasos de procesamiento de datos adicionales y/o usar diferentes conjuntos de parámetros de configuración de modo que sean minimizados los sesgos de datos causados por las características del experimento genómico.

[0029] Dependiendo de las características de entrada detectadas, el analizador de datos genómicos puede configurar además el módulo de llamada de variantes 122 para operar pasos de procesamiento de datos adicionales y/o usar diferentes conjuntos de parámetros de configuración de modo que se minimicen los sesgos de datos causados por las características del experimento genómico.

[0030] Dependiendo de los resultados del alineamiento de secuencias inicial por el módulo de alineamiento de secuencias 121, el analizador de datos genómicos 120 puede adaptarse adicionalmente para identificar los sesgos de alineamiento de datos de secuenciación de próxima generación que se hacen evidentes al analizar los datos de alineamiento. En consecuencia, el analizador de datos genómicos puede configurar el módulo de alineación de secuencias 121 para ejecutar un paso adicional de realineación de los datos de secuenciación NGS sin procesar. Esta realineación puede estar restringida por parámetros adicionales que pueden determinarse a partir de los resultados de la alineación inicial. En una posible forma de realización, la realineación se aplica específicamente en una subregión de la secuencia genómica. Los datos de realineación resultantes pueden ser filtrados y analizados adicionalmente por el módulo de llamada de variantes 122 para generar un informe de variantes genómicas más relevante con mayor sensibilidad y especificidad para la detección de variantes.

[0031] Dependiendo de los resultados de la llamada de variantes por parte del módulo de llamada de variantes 122, el analizador de datos genómicos 120 puede adaptarse adicionalmente para identificar los sesgos de llamada de variantes que se hacen evidentes cuando se llaman variantes en los datos de alineación. En consecuencia, el analizador de datos genómicos puede configurar el módulo de llamada de variantes 122 para ejecutar un paso adicional de recuperación de variantes en todos o parte de los datos de alineación. Este paso refinado de llamadas de variantes puede estar limitado por parámetros adicionales que pueden determinarse a partir de los resultados anteriores de alineación y/o realineación y/o llamadas de variantes. En una posible forma de realización, las variantes se solicitan específicamente en un subconjunto de los datos genómicos alineados. Los datos de llamada de variante refinados resultantes pueden

combinarse adicionalmente con los resultados de llamada de variante estándar mediante el módulo de llamada de variante 122 para generar un informe de variante genómica más relevante con mayor sensibilidad y especificidad para la detección de variantes. En una posible forma de realización, algunos resultados de llamadas de variantes pueden excluirse del informe de variantes genómicas identificados posiblemente sesgados por el módulo de llamada de variantes 122, de modo que el analizador de datos genómicos 120 genera un informe de variantes genómicas más relevante con mayor sensibilidad y especificidad para detección de variantes.

[0032] La FIG. 2 muestra en consecuencia un posible flujo de trabajo de análisis genómico para el analizador de datos genómicos 120, que comprende:

- recibir 200 una solicitud de análisis de secuenciación de próxima generación;
- identificar 211 un primer conjunto de características asociadas con la solicitud de análisis de secuenciación de próxima generación, comprendiendo el primer conjunto de características al menos un identificador de tecnología de enriquecimiento objetivo, un identificador de tecnología de secuenciación y un identificador de contexto genómico;
- configurar 231 un módulo de alineación de datos 121 para alinear los datos de secuenciación de entrada de acuerdo con al menos una característica del primer conjunto de características;
- alinear 232, con el módulo de alineación de datos configurado 121, los datos de secuenciación de entrada a una secuencia genómica, y reportar los datos de alineación en un archivo de datos de alineación sin procesar;
- identificar 212 un segundo conjunto de características asociadas con los datos de alineación del archivo de datos de alineación sin procesar, comprendiendo el segundo conjunto de características al menos un identificador de patrón de alineación de datos;
- configurar 233 el módulo de alineación de datos 121 para refinar al menos un subconjunto de los datos de secuenciación de entrada de acuerdo con al menos una característica del primer conjunto de características y al menos una característica del segundo conjunto de características;
- refinar 234, con el módulo de alineación de datos configurado 121, el subconjunto de datos de secuenciación de entrada para producir un archivo de datos de alineación refinado;
- identificar 213 un tercer conjunto de características asociadas con los datos de realineación del archivo de datos de alineación refinados, comprendiendo el tercer conjunto de características al menos un contexto genómico identificador;
- configurar 241 un módulo de llamada de variantes 122 para detectar variantes asociadas con los datos de alineación refinados de acuerdo con al menos una característica del primer conjunto de características, al menos una característica del segundo conjunto de características y al menos una característica del tercero conjunto de características;
- detectar 242 un primer conjunto de variantes genómicas, con el módulo de llamada de variante configurado 122, en los datos de alineación refinados;
- identificar 214 un cuarto conjunto de características asociadas con las variantes genómicas detectadas, comprendiendo el cuarto conjunto de características al menos una variante que llama identificador de refinamiento;
- configurar 243 el módulo de llamada de variantes 122 para detectar variantes asociadas con los datos de alineación refinados de acuerdo con al menos una característica del primer conjunto de características, al menos una característica del segundo conjunto de características, al menos una característica del tercer conjunto de características y al menos una característica del cuarto conjunto de características;
- detectar 244 variantes genómicas refinadas, con el módulo de llamada de variante configurado 122, en los datos de alineación refinados y las variantes genómicas detectadas, para producir un conjunto refinado de variantes genómicas;
- informando 250 el conjunto refinado de variantes genómicas.

[0033] El analizador de datos genómicos multipropósito genérico 120 facilita el análisis y el informe de múltiples variantes genómicas diferentes a partir de datos de secuenciación de próxima generación sin procesar recibidos de una diversidad de configuraciones clínicas operadas por múltiples laboratorios de abastecimiento sin requerir una configuración manual dedicada o metadatos exhaustivos documentación para adaptarse a cada combinación de configuración biológica y contexto de diagnóstico para cada análisis clínico.

Método refinado de llamada de variantes - flujo de trabajo ejemplar

[0034] Una forma de realización ejemplar del método refinado de llamada de variantes 244 propuesto para una identificación de variantes de patrones repetidos más precisa (homopolímeros y/o heteropolímeros) se describirá ahora con más detalle. El flujo de trabajo de análisis de datos genómicos totalmente automatizado de la FIG. 2 funciona con datos genómicos obtenidos de al menos un laboratorio de secuenciación de próxima generación.

[0035] Como ejemplo, el laboratorio puede operar un secuenciador Ion Torrent y xGen Lockdown Probes de Integrated DNA Technologies para identificar variantes genómicas en hasta 27 genes posiblemente relevantes para cánceres hereditarios de mama, ovario y gastrointestinales. Como será evidente para los expertos en la técnica de la genómica humana, algunas de esas variantes genómicas corresponden a regiones de homopolímero bastante largas en los cromosomas de referencia humanos de tipo salvaje, como por ejemplo (lista no exhaustiva) la repetición de 13 nucleótidos A en la región genómica CHEK2 en la posición 29130814 del cromosoma 22; la repetición de 14 nucleótidos T en la región genómica RAD54L en la posición 46739975 en el cromosoma 1; o la repetición de 19 nucleótidos T en la región genómica

ATM en la posición 108195977 en el cromosoma 11.

[0036] Como otro ejemplo, el laboratorio puede operar un secuenciador Illumina MiSeq® con "el ensayo Multiplicom CFTR MASTR™ Dx como la tecnología de enriquecimiento objetivo para identificar variantes genómicas en el gen CFTR cubierto por este kit específico basado en amplicón. El número de repeticiones en dinucleótidos TG (heteropolímeros) es típicamente 11 en el gen CFTR del cromosoma 7 humano, pero puede variar de 9 (2 deleciones) a 14 (3 inserciones) se repite. Las últimas variantes influyen en el empalme del exón 9 en el gen CFTR y se han asociado con la enfermedad de fibrosis quística cuando el gen CFTR también lleva una variante abreviada de solo homopolímeros 5T en el tracto de politimidina, que es típicamente de 7 Repeticiones de nucleótidos T, pero pueden variar de 5 a 9 repeticiones de nucleótidos T. Además, 12 o 13 repeticiones de heteropolímero TG también están asociadas con alguna patología de fibrosis quística menos común, mientras que 11 dinucleótidos TG las repeticiones están menos asociadas con la enfermedad (Hefferon et al, "Una repetición de dinucleótido variable en el gen CFTR contribuye a la diversidad de fenotipos al formar estructuras secundarias de ARN que alteran el empalme", Proc Natl Acad Sci EE. UU. 101: 3504-3509, 2004 - http: //www.pnas.OR9/content/101/10/3504.long).

[0037] En función de al menos una de las primeras características, a saber, el identificador de tecnología de enriquecimiento objetivo, el identificador de tecnología de secuenciación y/o el identificador de contexto genómico, el analizador de datos genómicos 120 configura 231 el módulo de alineación de datos 121 para ejecutar 232 una primera alineación de datos sin procesar. El módulo de alineación de datos 121 también puede ejecutar 232 pasos de preprocesamiento, como eliminar los adaptadores específicos del ensayo de las lecturas.

[0038] El módulo de alineación de datos 121 alinea 232 con una secuencia genómica de referencia, con un algoritmo de alineación de datos sin procesar como conocen los expertos en bioinformática, los datos de secuenciación sin procesar preprocesados para producir un archivo de alineación de datos. Se pueden usar algoritmos estándar como Bowtie2 o BWA que se han optimizado para el procesamiento rápido de numerosas lecturas de secuenciación de datos genómicos, pero también son posibles otras formas de realización. El archivo de alineación de datos resultante puede representarse como uno o varios archivos en formato BAM o SAM, pero también son posibles otras formas de realización, en particular, el módulo de alineación de datos 121 también puede ejecutar 232 pasos de procesamiento posterior, como comprimir y/o cifrar la alineación. datos, por ejemplo, con un cifrado que conserva el orden, un cifrado homomórfico, un cifrado simétrico y/o un esquema de cifrado asimétrico y/o una combinación de los mismos, dependiendo de los requisitos de analizador de datos genómicos 120 para la optimización del almacenamiento y/o la aplicación de la privacidad de los datos genómicos a lo largo del procesamiento del flujo de trabajo del análisis genómico.

[0039] El analizador de datos genómicos 120 puede derivar automáticamente 212 un segundo conjunto de características de los resultados de la alineación de datos 232, como un patrón de alineación de datos específico que requiere el refinamiento de la alineación y/o los algoritmos de llamada de variantes. El analizador de datos genómicos puede, por ejemplo, detectar la presencia de desajustes de alineación, especialmente al principio y/o al final de las lecturas ("recorte suave"), como puede deberse a un cebado incorrecto del cebador. Este sesgo frecuente en las tecnologías basadas en amplicones puede causar:

- Falsos positivos, cuando un artefacto de cebado incorrecto está presente en suficientes lecturas para estar desalineado con el genoma de referencia, lo que provocará una variante incorrecta que llama a la interpretación 242 como un SNP en la muestra de ADN;
- Falsos negativos, cuando el módulo de alineación 121 no puede discriminar entre artefactos de cebado incorrecto en ciertas lecturas y los datos de amplicón correctos en otras lecturas, lo que hace que las regiones correspondientes sean recortadas suavemente por el módulo de alineación de datos 121, lo que a su vez hará que la variante que llama 242 pierda posibles variantes de relevancia patológica en los datos de amplicón correctos.

[0040] Los patrones de recorte suave corresponden a datos de secuenciación en los límites 5' o 3' de las lecturas que no pudieron alinearse correctamente con los algoritmos de alineación sin procesar 232 del módulo de alineación de datos 121. Las alineaciones recortadas suaves están marcadas específicamente en la cadena CIGAR del archivo de datos de alineación, por lo que los patrones correspondientes se pueden identificar fácilmente después de la alineación de datos 232. Como saben los expertos en la técnica de la secuenciación de próxima generación, la información de recorte suave se puede volver a mapear en el flujo de trabajo de análisis genómico con algoritmos específicos para detectar además variantes estructurales de potencial relevancia clínica.

[0041] El analizador de datos genómicos 120 puede así identificar automáticamente 212 las lecturas con regiones de recorte suave, a partir de los resultados de la alineación de datos 232, y configurar 233 el módulo de alineación de datos 121 para operar una realineación de datos adicional 234 en esas lecturas específicamente tomando en cuenta la información de anclajes de cebadores correspondiente a la tecnología de enriquecimiento de ADN específica, en el algoritmo de alineación. Como será evidente para los expertos en la técnica de la bioinformática, se puede usar un algoritmo más robusto que Bowtie2 o BWA específicamente en esas regiones, incluso si es menos eficiente desde el punto de vista computacional. De hecho, solo un subconjunto de todos los datos NGS necesita realinearse de esta manera y el flujo de trabajo propuesto está completamente automatizado, por lo que el rendimiento de eficiencia computacional general del analizador de datos genómicos 120 no se ve afectado significativamente, mientras que esta realineación de datos la automatización del refinamiento permite aumentar la especificidad y la sensibilidad del analizador 120 de datos

genómicos para que sea comparable a la obtenida con configuraciones manuales de prueba y error de la práctica de investigación de la técnica anterior. Se han descrito ejemplos de dichos algoritmos, por ejemplo, por Suzuki et al. en "ClipCrop: una herramienta para detectar variaciones estructurales con resolución de base única utilizando información de recorte suave", BMC Bioinformatics 2011 12 (Suppl 14): S7 y por Schröder et al en "Socrates: identificación de reordenamientos genómicos en genomas tumorales por realineación de lecturas recortadas suaves", Bioinformatics (2014), pero también son posibles otras formas de realización. En particular, el algoritmo de realineación más eficiente puede ser configurado automáticamente 233 por el analizador de datos genómicos propuesto 120 como una función tanto del contexto genómico como de los patrones de recorte suave de datos de alineación sin procesar.

[0042] Dependiendo del identificador de contexto genómico, el analizador de datos genómicos 120 también puede identificar a partir de los datos de alineación la presencia de algunas regiones que son particularmente difíciles de alinear, como regiones de homopolímero, regiones de heteropolímero o, más generalmente, cualesquiera regiones con patrones de repetición específicos. La alineación adecuada de las lecturas de secuenciación de próxima generación correspondientes es particularmente desafiante ya que esas múltiples repeticiones causan ambigüedades de alineación. El analizador de datos genómicos 120 puede así identificar automáticamente 212 a partir de los resultados de la alineación de datos sin procesar 232 un contexto genómico específico que requiere refinamiento en las lecturas que se superponen a esas regiones ambiguas. En consecuencia, el analizador de datos genómicos 120 puede configurar 233 el módulo de alineación de datos 121 para operar una nueva realineación de datos 234 en esas lecturas para identificar otras posibles soluciones de alineación, como por ejemplo teniendo en cuenta la tasa de error de PCR y comparando las lecturas entre sí.

[0043] El analizador de datos genómicos 120 puede entonces usar el identificador de tecnología de enriquecimiento objetivo para configurar 241 el módulo de llamada de variantes 122 para ejecutar diferentes algoritmos de llamada de variantes de acuerdo con el identificador de contexto genómico identificado inicialmente (por ejemplo, CHEK2, ATM, BRCA2, CFTR...) y el refinamiento del contexto genómico específico identificado a partir de los resultados de alineación sin procesar (*p. ej.,* la presencia de ciertos patrones repetidos de homopolímeros en los datos). El módulo de llamada de variantes 122 llama a 242 variantes sobre los datos de alineación refinados para producir un primer archivo VCF. En algunos casos, las variantes resultantes pueden no ser lo suficientemente precisas para que las informe el analizador de datos genómicos 120.

[0044] Ahora se describirá con más detalle una posible forma de realización de un método refinado de llamada de variantes 244 adecuado para identificar mejor los tractos de patrones repetidos. Como saben los expertos en la técnica de la bioinformática, el analizador de datos genómicos 120 puede recibir del laboratorio datos de secuenciación de NGS sin procesar correspondientes a una diversidad de pacientes cuyas muestras se han agrupado en la misma ejecución de secuenciación. Debido a la menor precisión del experimento de secuenciación de ADN en presencia de repeticiones de nucleótidos de homopolímero o heteropolímero, se pueden detectar inserciones y eliminaciones falsas incluso cuando en realidad no hubo mutación en las muestras. En la práctica de laboratorio de última generación, pueden faltar hasta un 30 % de los homopolímeros, por lo que se analizan como deleciones debido a errores de secuenciación en esas regiones, lo que introduce un sesgo estadístico significativo y por lo tanto disminuye la precisión del análisis genómico. Algunos laboratorios también pueden secuenciar los datos de control de una muestra de tipo salvaje además de los datos del paciente, todos generados con el mismo proceso experimental (captura de ADN 100 y secuenciador NG 110), de modo que la referencia de datos de control pueda usarse para neutralizar los datos y facilitar la búsqueda de variantes incluso en contextos genómicos repetitivos que son particularmente difíciles de caracterizar, como se describe, por ejemplo, en la solicitud de patente europea en trámite WO/2018/104466. Sin embargo, dependiendo de la configuración del laboratorio, es posible que los datos de control no estén disponibles. En tal configuración, la llamada de variante precisa se ve desafiada por la dificultad de estimar correctamente la longitud de los patrones de secuencia repetitivos, como por ejemplo, el tramo de homopolímero A en el gen CHEK2, el tramo de homopolímero T en el gen RAD54L, el tramo de homopolímero T en el gen ATM, o el tramo de heteropolímero TG en el gen CFTR sin ninguna pista sobre los errores de secuenciación subyacentes que potencialmente sesgan las mediciones. De hecho, la longitud medida del patrón de repetición sigue una distribución de probabilidad discreta de la longitud de los patrones de repetición ("longitud de distribución") que puede depender tanto del sesgo del experimento como de la variante genómica real. Para una mayor sensibilidad y especificidad en el flujo de trabajo del análisis genómico, es deseable disminuir la contribución del sesgo del experimento en los datos de medición tanto como sea posible. Esto se puede lograr mediante el análisis cruzado de las medidas de la muestra bajo el supuesto de que una de ellas corresponde a la referencia del genoma humano (es decir, que no tiene ninguna variante). La medida de la muestra de referencia supuesta se puede usar como datos de referencia para predecir modelos variantes, a los que el método refinado de llamada de variantes 244 puede hacer coincidir las medidas de cada otra muestra, permitiendo así una identificación variante más precisa para esas muestras.

[0045] Como se describe, por ejemplo, en la solicitud de patente europea en trámite EP 16202691.8, la distribución de la longitud puede medirse como la distribución de probabilidad discreta de la longitud absoluta de un patrón repetido en un conjunto de lecturas de datos de secuencia genómica con cobertura suficiente. En otra posible forma de realización, la longitud de distribución puede medirse como la distribución de probabilidad normalizada discreta de la longitud relativa (que indica deleciones o inserciones) de un patrón repetido en un conjunto de lecturas de datos de secuencia genómica con cobertura suficiente, en relación con el patrón de referencia teórico del genoma humano, (también correspondiente a muestras de tipo salvaje que se encuentran más comúnmente sin mutación). A lo largo de esta divulgación, se considerará la longitud relativa para facilitar la comprensión de los métodos propuestos, sin embargo, como será evidente para los

expertos en la materia, los métodos propuestos también se aplican a las medidas de longitud absoluta como un simple cambio de las medidas de longitud relativa. Dependiendo del contexto genómico, el patrón de repetición puede ser un homopolímero, como la repetición de un solo nucleótido. En la aplicación de análisis genómico CHEK2, el homopolímero puede ser el patrón poli-A, la longitud absoluta de este patrón generalmente se puede medir en el intervalo de 11 a 15 repeticiones, o la longitud relativa en el intervalo de -2 (2 deleciones) a +2 (2 inserciones) con 0 que representa el patrón de repetición de tipo salvaje de nucleótidos 13A sin mutación. En la aplicación de análisis genómico RAD54L, el homopolímero puede ser el patrón poli-T, la longitud absoluta de este patrón generalmente se puede medir en el intervalo de 9 a 13, o la longitud relativa en el intervalo de -2 (2 deleciones) a +2 (2 inserciones) con 0 que representa el patrón de repetición de tipo salvaje de 11 nucleótidos T sin mutación. En la aplicación de análisis genómico ATM, el homopolímero puede ser el patrón poli-T, la longitud absoluta de este patrón puede medirse normalmente en el intervalo de 17 a 21, o la longitud relativa en el intervalo de -2 (2 deleciones) a +2 (2 inserciones) con 0 que representa el patrón de repetición de tipo salvaje de 19 nucleótidos T sin mutación.

[0046] Dependiendo del contexto genómico, el patrón de repetición también puede ser un heteropolímero, como la repetición de un par o triplete o más nucleótidos. En el ejemplo de CFTR, el patrón de repetición puede ser la repetición del heteropolímero TG, el rango de longitud absoluta para este patrón puede ser 11. En la aplicación de análisis genómico CFTR, el heteropolímero puede ser el patrón de dinucleótido TG, la longitud absoluta de este patrón suele ser medido en el intervalo de 9 a 14, o la longitud relativa en el intervalo de -2 (2 deleciones) a +3 (3 inserciones) representando 0 el patrón de repetición de tipo salvaje de 11 dinucleótidos TG sin mutación.

[0047] En las aplicaciones ejemplares de CHEK2, RAD54L o ATM, y análisis genómico CFTR respectivamente, el método refinado de llamada de variantes 244 puede evaluar la longitud del tracto de homopolímero poli-A, el tracto de homopolímero poli-T o el tracto de heteropolímero poli-TG respectivamente, para caracterizar mejor la variante genómica correspondiente. Para estimar con precisión la longitud de las repeticiones del patrón, el método de llamada de variantes 244 debería minimizar los sesgos causados por los errores del experimento. Esto se puede lograr estimando las longitudes de distribución esperadas para varios escenarios de inserción y eliminación en cada posible patrón de secuencia repetitiva para diferentes hipótesis en los datos de entrada y seleccionando la mejor coincidencia ("modelo de variante de mejor ajuste").

[0048] Como será evidente para los expertos en la técnica de la bioinformática, este método mejorará significativamente la precisión de la estimación de la longitud del patrón repetitivo siempre que las lecturas de secuenciación de próxima generación tengan una cobertura estadística suficientemente grande. También será evidente para los expertos en la técnica de la biología que no existe un método simple para que un operador humano identifique la referencia de tipo salvaje y/o la longitud real de los tramos de patrones repetidos en las muestras de ADN del paciente, por lo que existe una ventaja significativa en el empleo de un método de automatización de procesamiento de señales, con uno o más procesadores de computadora, para facilitar el análisis real de los datos de la muestra de ADN del paciente secuenciados con un secuenciador de próxima generación.

[0049] Bajo el supuesto de que los sesgos se aplican por igual a todas las muestras en el grupo de muestras, en particular también se aplicarán a las muestras de tipo salvaje. Por lo tanto, proponemos modelos de variantes de mejor ajuste en datos de muestra en relación con otra muestra en el grupo, que se supone que corresponde a la muestra de tipo salvaje. En el método de la solicitud de patente europea co-pendiente EP 16202691.8, la muestra de tipo salvaje es la muestra de datos de control. Si no hay ninguna muestra de datos de control presente en el conjunto, se pueden realizar y verificar diferentes hipótesis mediante el análisis cruzado de cada muestra con respecto a las otras muestras, como se describirá ahora con mayor detalle.

Mejor ajuste de muestra a muestra de modelos variantes

[0050] Una distribución teórica de la longitud de un tracto de patrón repetido en muestras de pacientes en relación con un tracto de patrón repetido en la referencia del genoma humano se ilustra en la FIG. 3 respectivamente para una muestra de datos de control sin mutación y sin sesgo de error del experimento, y un escenario de una sola mutación de eliminación de motivo básico, por ejemplo, la eliminación de un nucleótido en un solo alelo en la región genómica del homopolímero CHEK2 (longitud del patrón de homopolímero de referencia REF=13) o ATM (referencia longitud del patrón del homopolímero REF=19), o la eliminación de un dinucleótido (dos nucleótidos) en un solo alelo en la región genómica del heteropolímero CFTR (longitud del patrón del heteropolímero de referencia REF=11). Esta deleción heterocigota se representa como una diferencia de longitud de patrón de repetición de -1 en un alelo en relación con la referencia, y una diferencia de longitud de patrón de repetición de 0 en el otro alelo. Esto corresponde al caso teórico ideal en el que cada alelo contribuye con el 50 % de la medición de la longitud de distribución, por lo que se espera medir respectivamente una longitud de (REF-1) patrones repetidos y una longitud de repeticiones REF a partir de los datos de realineación, cada uno con una probabilidad igual de 0,5 en la distribución de probabilidad discreta normalizada de la longitud del tracto de patrón repetido.

[0051] Una distribución medida ejemplar de la longitud del mismo tramo de patrón repetido en muestras de pacientes en relación con un tramo de patrón repetido en la referencia del genoma humano se ilustra en la FIG. 4 respectivamente para un conjunto de datos experimentales sin mutación pero sujeto a un experimento sesgo de error, que causa (tabla superior de la FIG. 4) mediciones erróneas de longitudes más cortas de 2 eliminaciones (10 % de los datos de cobertura de control)

o 1 eliminación (20 % de los datos de cobertura de control) y mediciones erróneas de longitudes más largas de 1 inserción (10 % de los datos de cobertura de control), de modo que solo el 60 % de los datos de cobertura corresponden a la longitud real de la región repetitiva, por ejemplo, una longitud real de nucleótidos REF=13 A para el patrón estándar homocigoto 13-A sin mutación en el gen CHEK2 en la referencia del genoma humano o una longitud real de nucleótidos REF=19 T para el patrón homocigótico estándar 19-T sin mutación en el gen ATM en la referencia del genoma humano. La distribución de probabilidad discreta medida de la longitud del patrón para los datos del paciente estará sesgada en consecuencia y, por lo tanto, este sesgo inducido por el experimento puede tenerse en cuenta al estimar la distribución de probabilidad discreta esperada de la longitud del patrón para cada posible escenario de variante de mutación para mejorar la sensibilidad y especificidad. Como ejemplo, la tabla inferior de la FIG. 4 muestra la distribución de probabilidad discreta esperada resultante de la estimación de la longitud del patrón para el escenario de una eliminación de un solo motivo básico en relación con la referencia del genoma humano, por ejemplo, la eliminación de un nucleótido en un solo alelo en el tracto de homopolímero poli-A del gen CHEK2 o el tracto de homopolímero poli-T del gen ATM o RAD54L (deleción heterocigota - longitud del patrón = REF-1 en un alelo, y longitud del patrón = REF en el otro alelo). En el alelo 1 mutado, toda la distribución de probabilidad de la longitud del patrón puede estar desplazada hacia la izquierda, debido a la eliminación real de un motivo básico de nucleótido. Dado que ambos alelos contribuyen por igual a los datos de medición generales, sus contribuciones se pueden resumir y promediar simplemente para proporcionar la distribución de probabilidad esperada de la longitud del patrón para este escenario de mutación de eliminación única, teniendo en cuenta el sesgo de error del experimento de los datos de tipo salvaje: en esta ilustración de ejemplo, podemos esperar que el 5 % de los datos de un paciente con una mutación de deleción única se miden como portadores de 3 deleciones, el 15 % como portadores de 2 deleciones, el 40 % como portadores de 1 deleción (resultado correcto), el 35 % como portadores sin mutación, y el 5 % como portador de una sola inserción del motivo básico, por ejemplo, el nucleótido T en el tramo de homopolímero poli-T en el RAD54L o los genes ATM, o el dinucleótido TG en el tramo de heteropolímero poli-TG en el CFTR gen.

[0052] La FIG. 5a) muestra distribuciones de probabilidad discretas esperadas ejemplares de las longitudes del patrón de repetición en relación con la longitud del patrón de repetición de referencia del genoma para una diversidad de modelos de escenarios de variantes genómicas que van desde una deleción doble homocigótica (arriba a la izquierda - longitud de -2 en comparación con la referencia centrada en 0) a una doble inserción homocigota (abajo a la derecha - longitud de +2 en comparación con la referencia centrada en 0) cuando el sesgo del experimento provoca la siguiente medición de distribución de probabilidad errónea incluso en los datos de tipo salvaje sin mutación (escenario 0/0): 40 % de medición sin mutación, 30 % de medición de una sola eliminación (longitud de -1 en comparación con la referencia centrada en 0), 30 % de medición de una sola inserción (longitud de +1 en comparación con la referencia centrada en 0).

[0053] Como será evidente para los expertos en la técnica de la estadística, el módulo de llamada de variantes 122 puede aplicar diferentes métodos, por ejemplo, una medida de distancia estadística, para comparar la distribución de probabilidad discreta normalizada medida de la longitud en los datos del paciente, como se ilustra, por ejemplo, en la FIG. 5b), con la distribución de probabilidad discreta normalizada esperada de la longitud para cada escenario, como se ilustra, por ejemplo, en la FIG. 5a). El módulo de llamada variante 122 puede entonces seleccionar el escenario de comparación más cercano como el escenario variante que da como resultado la distancia mínima estimada (modelo de mejor ajuste).

[0054] Se pueden usar las longitudes relativas (número de inserciones o eliminaciones en relación con la longitud *I* de las repeticiones de tipo salvaje en la referencia del genoma humano) o las longitudes absolutas para representar las distribuciones, una de las cuales es un simple cambio de las coordenadas de referencia en comparación con el otro. Un escenario variante de repeticiones R_1 de un patrón de nucleótidos en el primer alelo y repeticiones R_2 en el segundo alelo, por lo tanto, puede señalarse como un escenario variante $[R_1|\ R_2]$ en coordenadas absolutas, o como $[V_1|\ V_2] = [R_1 - I\ |\ R_2 - I]$ en coordenadas relativas, donde *I* es el número de repeticiones en la referencia del genoma humano.

[0055] Así, en una posible forma de realización, la distancia estadística entre la distribución de probabilidad discreta medida de la longitud en los datos del paciente y la distribución de probabilidad discreta esperada de la longitud para un escenario variante $[R_1|R_2]$ de las repeticiones R_1 de un patrón de nucleótidos en el primer alelo y las repeticiones R_2 en el segundo alelo se pueden calcular como la distancia euclidiana entre los vectores que representan sus respectivas distribuciones de probabilidad discretas normalizadas. Alternativamente, en una posible forma de realización, la distancia estadística entre la distribución de probabilidad discreta medida de la longitud en los datos del paciente en relación con la referencia del genoma humano y la distribución de probabilidad discreta esperada de la longitud para un escenario variante $[V_1\ |\ V_2]=[R_1 - I\ |\ R_2 - I]$ de las repeticiones R_1 de un patrón de nucleótidos en el primer alelo y las repeticiones R_2 en el segundo alelo se pueden calcular como la distancia euclidiana entre los vectores que representan las respectivas distribuciones de probabilidad.

[0056] Como será evidente para los expertos en la técnica del análisis estadístico, se pueden aplicar varios métodos para determinar el modelo que mejor se ajusta. En una posible forma de realización, puede usarse la minimización del error cuadrático medio entre las distribuciones estadísticas medidas y esperadas. De manera más general, es posible minimizar una distancia n-norma, como la distancia euclidiana o la distancia 1-norma. Como será evidente para los expertos en bioinformática, se pueden usar otros métodos de ajuste estadístico para determinar el modelo de mejor ajuste, como por ejemplo algunos de los métodos mencionados en la solicitud de patente de EE. UU. 2014/0052381 de Utirametur et al.

[0057] En el ejemplo ilustrativo de la FIG. 5 (usando coordenadas de longitud relativa), el escenario de variante [0/1] se

seleccionará en consecuencia como el modelo que mejor se ajusta para la variante genómica de patrón repetido en relación con la referencia del genoma humano (inserción única heterocigótica), por ejemplo 13-A en un alelo y 14-A en el otro alelo, en el ejemplo CHEK2 [13A/14A]).

[0058] Los métodos propuestos anteriormente permiten estimar de forma fiable las respectivas distribuciones de probabilidad no sesgadas de las longitudes del patrón repetido (por ejemplo, la repetición del motivo básico del homopolímero C o el heteropolímero TG) en los datos de muestra del paciente cuando la cobertura de datos de control experimental está disponible y representa el sesgo de medición experimental de tipo salvaje, es decir, cuando se incluye una muestra de datos de control en el grupo de laboratorio. Sin embargo, no todas las configuraciones de análisis genómico proporcionan tales datos de control de tipo salvaje. Por lo tanto, es necesario considerar más a fondo diferentes hipótesis sobre las muestras reales que pueden representar los datos de tipo salvaje ("referencia") en el grupo de pacientes. Esto se entenderá mejor con las ilustraciones de la FIG. 6, FIG. 7 y FIG. 8 que representan ejemplos de tales hipótesis en el caso más simple de comparar 2 muestras, respectivamente del paciente 1 y el paciente 2, entre sí en un intento de determinar si uno de ellos corresponde a la muestra de tipo salvaje, posiblemente con sesgo de medición al identificar erróneamente una mutación en el patrón repetido (la variante de muestra de tipo salvaje debe llamarse [0|0] después de la variante refinada llamando 244, independientemente de la observación sesgada inicial).

[0059] En la FIG. 6, la distribución medida de la longitud relativa del patrón repetido de la muestra 1 es unimodal centrada en 0 (sin mutación) mientras que la distribución medida de la longitud relativa del patrón repetido de la muestra 2 es bimodal con dos picos respectivamente en -2 (dos eliminaciones) y +1 (una inserción). En la hipótesis de que la distribución medida para la muestra 1 corresponde a la medida de tipo salvaje con sesgo experimental, se puede encontrar una coincidencia perfecta como la distancia mínima posible $d = 0$ entre la distribución medida para la muestra 2 y la distribución predicha para un modelo variante [-2|1] de 2 deleciones en un alelo y 1 inserción en un alelo para la muestra 2 en relación con la muestra 1. En la hipótesis de que la distribución medida para la muestra 2 corresponde a la medida de tipo salvaje con sesgo experimental, no se puede encontrar coincidencia (d=0,95) entre la distribución medida para la muestra 1 y la distribución predicha para un modelo variante [-112] de 1 deleción en un alelo y 2 deleciones en un alelo para la muestra 1 en relación con la muestra 2. En este ejemplo, la variante que llama el método 244 permite que el analizador de datos genómicos 120 identifique con buena confianza que la longitud del patrón de repetición es la misma que en el genoma de referencia humano para el paciente 1 en ambos alelos, correspondientes al tipo salvaje (*p. ej.,* 13 nucleótidos A en CHEK2, 11 nucleótidos T en RAD54L, 19 T en ATM o 11 TG en CFTR), mientras que para el paciente 2 la longitud del patrón repetido es dos nucleótidos más corta en un alelo (*p. ej.,* 11 nucleótidos A en CHEK2, 9 nucleótidos T en RAD54L, 17 T en ATM o 9 TG en CFTR) y un nucleótido más largo en comparación con el genoma de referencia humano en el otro alelo (*p. ej.,* 14 nucleótidos A en CHEK2, 12 nucleótidos T en RAD54L, 20 T en ATM, o 12 TG en CFTR).

[0060] En la FIG. 7, la distribución medida de la longitud relativa del patrón repetido de la muestra 1 es unimodal centrada en 0 (sin mutación) mientras que la distribución medida de la longitud relativa del patrón repetido de la muestra 2 es unimodal centrada en -1 (una deleción homocigota). En la hipótesis de que la distribución medida para la muestra 1 corresponde a la medida de tipo salvaje sin sesgo experimental (ya que está centrada en 0), se puede encontrar una coincidencia perfecta como la distancia mínima posible d=0 entre la distribución medida para la muestra 2 y la distribución predicha para un modelo variante [-1|-1] de una deleción homocigota de un solo nucleótido para la muestra 2 en relación con la muestra 1. En la hipótesis de que la distribución medida para la muestra 2 corresponde a la medición de tipo salvaje con sesgo experimental de un homocigoto eliminación (ya que se centra en -1), también se puede encontrar una coincidencia perfecta como la distancia mínima posible d=0 entre la distribución medida para la muestra 1 y la distribución predicha para un modelo variante [+1|+1] de una inserción homocigota de un solo nucleótido para la muestra 1 en relación con la muestra 2. En este ejemplo, no es posible discriminar si la muestra 1 o la muestra 2 corresponde al tipo salvaje, un análisis cruzado adicional con otras muestras será necesario para el método de llamada de variantes 244 para permitir que el analizador de datos genómicos 120 identifique variantes reales con buena confianza en la muestra 1 y la muestra 2.

[0061] En la FIG. 8, la distribución medida de la longitud relativa del patrón de repetición de la muestra 1 es bimodal con dos picos respectivamente en -1 (una eliminación) y +2 (dos inserciones) mientras que la distribución medida de la longitud de patrón de repetición relativo de la muestra 2 es unimodal centrada entre -1 (una eliminación) y + 1 (una inserción). En la hipótesis de que la distribución medida para la muestra 1 corresponde a la medida de tipo salvaje con sesgo experimental, la mejor coincidencia se puede encontrar a una distancia d=0,467 entre la distribución medida para la muestra 2 y la distribución predicha para un modelo variante [-110] de una deleción heterocigota de un solo nucleótido para la muestra 2 en relación con la muestra 1. En la hipótesis de que la distribución medida para la muestra 2 corresponde a la medida de tipo salvaje con sesgo experimental, la mejor coincidencia se puede encontrar en la distancia mínima posible d=0,363 entre la distribución medida para la muestra 1 y la distribución predicha para un modelo variante [-1|+2] de una eliminación de un solo nucleótido en un alelo y una inserción de doble nucleótido en el segundo alelo para la muestra 1 en relación con la muestra 2. En este ejemplo, independientemente de qué muestra corresponda al tipo salvaje, observamos que el método de llamada de variantes 244 no permite que el analizador de datos genómicos 120 identifique con suficiente confianza nce la longitud del patrón de repetición para el paciente 1 y el paciente 2. Sin embargo, aún puede ser posible identificarlo mediante el análisis cruzado de más muestras de pacientes, como se describirá ahora con más detalles.

Análisis cruzado de múltiples muestras

[0062] Como se ilustra mediante el flujo de trabajo de la FIG. 9, el enfoque general propuesto en este documento consiste en considerar sucesivamente cada muestra en el conjunto de muestras como posiblemente la muestra de referencia de tipo salvaje (WT) como hipótesis de trabajo. El método puede comenzar calculando el modelo de mejor ajuste con un nivel de confianza para cada par de muestras como se describe en la sección anterior. Como se ilustró anteriormente mediante los ejemplos de la FIG. 6, la FIG. 7 y FIG. 8, el análisis de cada par de muestras puede proporcionar un modelo de mejor ajuste con un cierto nivel de confianza. Una vez que se han identificado inicialmente los modelos de mejor ajuste de muestra a muestra, se puede realizar un análisis cruzado de muestra adicional con cada triplete posible para identificar el mejor modelo de mejor ajuste para ellos, así como un nivel de confianza correspondiente. El nivel de confianza general para cada hipótesis de trabajo se puede calcular iterativamente como una función de las distancias entre las distribuciones medidas de las longitudes de los patrones repetidos y los modelos de variantes de mejor ajuste coincidentes a medida que se refinan en cada iteración de análisis cruzado. Se pueden diseñar diferentes algoritmos para realizar un análisis cruzado de múltiples muestras bajo los supuestos anteriores, de modo que converjan después de un número limitado de iteraciones. A continuación, identificar las muestras correspondientes al tipo salvaje (que debe ser identificado por el método refinado de llamadas de variantes 244 como homocigotas, sin mutación) permite tener en cuenta un posible sesgo experimental en las mediciones del grupo (lo que puede resultar en la detección errónea de mutaciones cuando se aplica la llamada de variante no refinada a los datos medidos en las regiones genómicas de patrones repetidos).

[0063] En un conjunto de muestras, puede haber varias muestras que lleven el patrón de tipo salvaje. Después de realizar un análisis cruzado de las muestras, es ventajoso intentar agruparlas. Luego, se pueden aplicar ciertas heurísticas para identificar qué grupo es más probable que corresponda al tipo salvaje; por ejemplo:

- Es más probable que las muestras con variantes homocigóticas sean muestras de tipo salvaje que aquellas con variantes heterocigóticas, ya que la mayoría de los sesgos experimentales deberían aplicarse indiferentemente a ambos alelos.
- Las muestras con múltiples variantes de deleción pueden ser muestras de tipo salvaje (como puede deberse a que el secuenciador de próxima generación salta nucleótidos), sin embargo, después de la variante refinada llamando a 244 de acuerdo con esta hipótesis, todas las demás muestras aún deberían tener variantes en el intervalo posible conocido (por ejemplo, para CFTR, no más de 2 deleciones del patrón TG en comparación con la referencia del genoma humano). En otras palabras, esta suposición de trabajo solo se puede considerar si las otras muestras tienen suficiente longitud de patrón observada para ajustarse mejor a un modelo de variante plausible.

[0064] Al identificar el grupo de muestras correspondiente al tipo salvaje en el grupo de muestras en función del análisis estadístico de muestras cruzadas, el método refinado de llamada de variantes 244 puede caracterizar con mayor precisión la variante del patrón repetido para cada muestra en el grupo como la variante derivada de la hipótesis de trabajo del sesgo de tipo salvaje con el nivel de confianza más alto. El analizador de datos genómicos propuesto 120, por lo tanto, detectará 244 e informará de 250 con mayor precisión las variantes de patrones repetidos para múltiples muestras de pacientes.

[0065] En una posible forma de realización, el método refinado de llamada de variantes 244 para identificar una variante de patrón repetido como el patrón repetido de al menos dos nucleótidos (caso de homopolímero) o al menos dos grupos de nucleótidos (caso de heteropolímero: dinucleótido, triplete de nucleótidos, etc...) en la secuencia genómica de una muestra de un paciente puede comprender:

(a) identificar un patrón repetido de referencia $P_{ref} = N*l$ como la repetición de $l(l>=2)$ patrones genómicos N en una región genómica de una secuencia de referencia del genoma humano;
(b) obtener, con un secuenciador de próxima generación, n conjuntos de pacientes de lecturas de datos de secuenciación de próxima generación que cubran la región genómica del patrón repetido de referencia $S= \{S_1, S_2, ..., S_i, ..., S_n\}$ de un conjunto de n muestras de pacientes genómicos enriquecidos, estando asociado cada conjunto Si con una muestra de paciente, siendo el número n de muestras de pacientes genómicos enriquecidos al menos 4;
(c) para cada muestra de paciente i en el conjunto S de muestras de pacientes, medir la distribución P_i de la longitud del patrón repetido en el conjunto de lecturas de secuenciación de próxima generación S_i;
(d) para un posible par de muestras de pacientes i y j, $j>i$:

(d1) estimación para la muestra j, bajo el supuesto de que la muestra i lleva el patrón de homopolímero de referencia del genoma humano de tipo salvaje $P_{ref}=N*l$ para cada alelo, un modelo de mejor ajuste $[V_{ji}^1 | V_{ji}^2]$ de las dos variantes alélicas para la muestra j, con un nivel de confianza $L_{j|i}$, así como la distancia más pequeña $D_{j|i}$ entre la distribución medida P_j para la muestra j y la distribución unimodal o bimodal predicha para el modelo de variante de mejor ajuste $[V_{ji}^1 | V_{ji}^2]$;
(d2) estimando para la muestra i, bajo el supuesto de que la muestra j lleva el patrón de homopolímero de referencia del genoma humano de tipo salvaje $P_{ref} =N*l$ para cada alelo, un modelo de mejor ajuste $[V_{ji}^1 | V_{ji}^2]$ de las dos variantes de alelo para la muestra i, con un nivel de confianza $L_{i|j}$, así como la distancia más pequeña $D_{i|j}$ entre la distribución medida P_i para la muestra i y la distribución unimodal o bimodal pronosticada para el modelo de variante de mejor ajuste $[V_{ji}^1 | V_{ji}^2]$;
(d3) si $D_{i|j} \geq D_{j|i}$, seleccionando para el par de muestras (i,j) el modelo de mejor ajuste $[V_{ij}^1 | V_{ij}^2] = [-V_{ji}^1 | -V_{ji}^2]$ como

el modelo variante de mejor ajuste de las dos variantes alélicas y el nivel de confianza $L_{ij} = L_{ji}$ como el valor del nivel de confianza para esta coincidencia de mejor ajuste con la muestra *i* como la muestra de referencia del par *(i,j);*
*(*d4) de lo contrario si $D_{i|j} < D_{j|i}$ seleccionando para el par de muestras *(i,j)* el modelo $[V_{ij}^1|\, V_{ij}^2] = [V_{i|j}^1|\, V_{i|j}^2]$ como el modelo de variante de mejor ajuste de las dos variantes de alelo y el nivel de confianza $L_{ij} = L_{i|j}$ como el valor del nivel de confianza para esta coincidencia de mejor ajuste con la muestra *j* como la muestra de referencia del par *(i,j);*

(e) para cada posible triplete de muestras de pacientes *i, j>i, k>j,* comparar sus respectivos modelos de mejor ajuste $[V_{j|i}^1|\, V_{j|i}^2]$ $[V_{jk}^1|\, V_{jk}^2]$, $[V_{ik}^1|\, V_{ik}^2]$, y:

(el) si los tres modelos de mejor ajuste para el triplete de muestras de pacientes coinciden entre sí, aumentando sus niveles de confianza L_{ij}, L_{jk}, L_{ik};
(e2) de lo contrario, los tres modelos de mejor ajuste no coinciden entre sí, reemplazando el modelo de mejor ajuste con el nivel de confianza más bajo del subconjunto por un modelo de mejor ajuste calculado a partir de las otras dos muestras del subconjunto, y disminuyendo los niveles de confianza L_{ij},L_{jk},L_{ik} de todos los modelos de mejor ajuste para el triplete de muestras de pacientes.

(f) repetir el paso (e) hasta que los resultados ya no varíen;
(g) hacer coincidir grupos de modelos idénticos de variantes de mejor ajuste e identificar el grupo más probable que porta la variante de tipo salvaje;
(h) para cada muestra fuera del grupo que lleva la variante de tipo salvaje, neutralizar el modelo de variante de mejor ajuste del grupo que comprende esta muestra en función del modelo de variante de mejor ajuste para el grupo de tipo salvaje identificado, y reportar la variante no sesgada para la muestra.

[0066] Como será evidente para los expertos en la técnica de las estadísticas, el nivel de confianza puede estimarse inicialmente y refinarse más mediante diferentes métodos matemáticos. En una posible forma de realización, para cualquier par de muestras de pacientes *i* y *j, j>i,* el nivel de confianza L_{ij} de la estimación $[V_{ij}^1|\, V_{ij}^2]$ puede calcularse inicialmente como:

$$L_{ij} = \begin{cases} 1 - D_{i|j}/D_{i|j}', D_{i|j} > D_{j|i} \\ 1 - D_{j|i}/D_{j|i}', \quad D_{i|j} \leq D_{j|i} \end{cases}$$

donde $D_{j|i}$ es la distancia más pequeña y $D_{j|i}'$ es la segunda distancia más pequeña calculada en el paso d1), y $D_{i|j}$ es la distancia más pequeña y $D_{i|j}'$ es la segunda distancia más pequeña calculada en el paso d2). El valor del nivel de confianza L_{ij} es, por lo tanto, un valor flotante en el intervalo [0,1], con el valor 1 cuantificando un nivel de confianza máximo para *i* o *j* como referencia y el valor 0 cuantificando un nivel de confianza mínimo, por ejemplo cuando no es posible diferenciar *i* de *j* como la referencia.

[0067] En una posible forma de realización, los niveles de confianza para cada par en un subconjunto triplete *i, j, k* de modelos coincidentes de mejor ajuste pueden incrementarse respectivamente en el paso e1) como:

$$L_{ij}' = 1-(1- L_{ij})(1- L_{jk}*L_{ik})$$

$$L_{jk}' = 1-(1- L_{jk})(1- L_{ij}*L_{ik})$$

$$L_{ik}' = 1-(1- L_{ik})(1- L_{ij}*L_{jk})$$

[0068] En una posible forma de realización, los niveles de confianza para cada par en un triplete subconjunto *i, j, k* de modelos de mejor ajuste no coincidentes donde el nivel de confianza inicial más bajo es L_{ik} puede disminuirse respectivamente en el paso e2) como:

$$L_{ij}' = L_{ij} - (1 - L_{jk})*L_{ik}$$

$$L_{jk}' = L_{jk} - (1 - L_{ij})*L_{ik}$$

$$L_{ik}' = \max(0, L_{ij}*L_{jk} - L_{ik})$$

y el modelo de mejor ajuste para el par *j, k* con el nivel de confianza más bajo fuera del subconjunto puede ser reemplazado por un modelo de mejor ajuste calculado a partir de la otras dos muestras del subconjunto *j* asumiendo transitividad entre muestras dentro del triplete, como:

$$[V_{ik}^{1}{}'|\ V_{ik}^{2}{}'] = [V_{ij}^{1} + V_{ij}^{1}|\ V_{jk}^{2} + V_{jk}^{2}]$$

[0069] En particular, cuando no se pueden comparar diferentes tipos de mutaciones heterocigóticas, el resultado de $[V_{ij.k}^{1}|\ V_{ij.k}^{2}]$ puede excluirse del análisis cruzado ([NA| NA]):

1. si $V_{ik}^{1} \neq V_{ik}^{2}$ (que indica una mutación heterocigótica en la muestra *i* o en la muestra *k)* y $D_{i|k} > D_{k|i}$ (que indica que la muestra *k* es homocigota) y $V_{kj}^{1} \neq V_{kj}^{2}$ (que indica una mutación heterocigota en la muestra *j* o en la muestra *k)* y $D_{j|k} > D_{k|j}$ (que indica que la muestra *k* es homocigota) y [$V_{ik}^{1} \neq -V_{kj}^{1}$ o $V_{ik}^{2} \neq -V_{kj}^{2}$] (que indica que no es el mismo tipo de mutación heterocigota en la muestra *i* y la muestra *k)*;
2. si $[V_{ik}^{1}|\ V_{ik}^{2}]$ es igual a [NA|NA] o $[V_{kj}^{1}|\ V_{kj}^{2}]$ es igual a [NA|NA].

[0070] En otro caso particular, si $[V_{ij}^{1}|\ V_{ij}^{2}]$ es igual a [NA|NA]: como este resultado puede deberse al ruido de secuenciación en la muestra *i* o en la muestra *j,* el resultado se puede volver a medir con la misma Fórmula anterior, de las otras dos muestras de un triplete.

[0071] De lo contrario, el nivel de confianza correspondiente para $[V_{ij.k}^{1}|\ V_{ij.k}^{2}]$ puede estimarse como:

$$L_{ij.k} = L_{ik} * L_{kj}$$

[0072] Para cada par *i* y *j,* como máximo *n*-2 resultados de valores $[V_{ij.k}^{1}|\ V_{ij.k}^{2}]$ y $L_{ij.k}$ pueden derivarse de tripletes, donde *n* es el número de muestras de pacientes. Además, partiendo de $[V_{ij.0}^{1}|\ V_{ij.0}^{2}] = [V_{ij}^{1}|\ V_{ij}^{2}]$ y $L_{ij.0} = L_{ij,}$ para cada par *i* y *j,* hay como máximo *n*-1 resultados de $[V_{ij.k}^{1}|\ V_{ij.k}^{2}]$ y $L_{ij.k,}$ con *k* =0 a *n*, $k \neq i$ y $k \neq j$.

[0073] Los pasos a) a f) anteriores permiten identificar los modelos variantes de mejor ajuste más probables a partir del análisis cruzado de los datos del grupo de muestras. Una aplicación de ejemplo de los métodos de comparación de muestras cruzadas propuestos anteriormente se describirá ahora como se ilustra en las FIG. 10 a FIG. 13.

[0074] En la FIG. 10, la muestra *i*=1 frente a la muestra j=2 está asociada con el modelo de variante de mejor ajuste sin mutación en relación la una con la otra $[V_{12}^{1}|\ V_{12}^{2}] = [0|0]$, con un nivel de confianza L_{12}=*0,*46; la muestra *i*=1 frente a la muestra *j*=3 está asociada con el modelo de variante de ajuste óptimo de ausencia de mutación entre sí $[V_{13}^{1}|\ V_{13}^{2}] = [0|0]$, con un nivel de confianza L_{13}=*0,*60; la muestra *i*=2 frente a la muestra *j*=3 se asocia con el modelo de variante de mejor ajuste sin mutación entre sí $[V_{23}^{1}|\ V_{23}^{2}] =$ [0|0], con un nivel de confianza L_{23}=*0,*32. Los modelos de mejor ajuste para el triplete de muestras de pacientes (1,2,3) coinciden entre sí, por lo que los niveles de confianza correspondientes $L_{12,}$ $L_{13,}$ L_{23} pueden incrementarse en consecuencia como L_{12}'=0,564, L_{23}'=0,518 y L_{13}'=0,659 respectivamente.

[0075] En la FIG. 11, la muestra *i*=2 frente a la muestra j=7 está asociada con el modelo de variante de mejor ajuste sin mutación entre sí $[V_{27}^{1}|\ V_{27}^{2}] = = [0|0]$, con un nivel de confianza L_{27}=*0,*18; la muestra *i*=2 frente a la muestra *j*=8 está asociada con el modelo de variante de mejor ajuste de una deleción heterocigota en la muestra 2 en relación con la muestra 8 $[V_{28}^{1}|\ V_{28}^{2}] = = [-1|0]$, con un nivel de confianza L_{28}=*0,*10; la muestra *i*=7 frente a la muestra *j*=8 se asocia con el modelo de variante de ajuste óptimo de ausencia de mutación entre sí $[V_{78}^{1}|\ V_{78}^{2}] = [0|0]$, con un nivel de confianza L_{78}=*0,*68. Los modelos de mejor ajuste para el subconjunto de muestras cruzadas 2 vs 7 y 7 vs 8 y 2 vs 8 no coinciden entre sí y el nivel de confianza L_{28}=0,10 es el valor más bajo del triplete, por lo tanto, el modelo de mejor ajuste $[V_{28}^{1}|\ V_{28}^{2}]$ de la muestra 2 frente a la muestra 8 puede reemplazarse por el del subconjunto coincidente $[V_{27}^{1} + V_{78}^{1}|\ V_{27}^{2} + V_{78}^{2}] = [0|0]$, y los niveles de confianza correspondientes $L_{27,}$ $L_{28,}$ L_{78} se reducirán en consecuencia como L_{27}'=0,148, L_{78}'=0,598 y L_{28}'=0,022 respectivamente.

[0076] En la FIG. 12, la muestra *i*=1 frente a la muestra *j*=2 está asociada con el modelo de variante de mejor ajuste sin mutación entre sí $[V_{12}^{1}|\ V_{12}^{2}] = [0|0]$, con un nivel de confianza L_{12}=*0,*46; la muestra *i*=1 frente a la muestra *j*=6 está asociada con el modelo de variante de ajuste óptimo de ausencia de mutación entre sí $[V_{16}^{1}|\ V_{16}^{2}] = [0|0]$, con un nivel de confianza L_{16}=*0,*67; pero la muestra i=2 frente a la muestra j=6 se ha clasificado como poco fiable (NA) en un paso anterior. El modelo de mejor ajuste $[V_{26}^{1}|\ V_{26}^{2}]$ de la muestra 2 frente a la muestra 6 puede ser reemplazado por el del subconjunto coincidente $[V_{21}^{1} + V_{16}^{1}|\ V_{21}^{2} + V_{16}^{2}] = [0|0]$, con un nivel de confianza $L_{26,1}$=0,308. El nivel de confianza L_{12} y L_{16} permanece sin cambios, ya que [NA|NA] entre la muestra 2 y la muestra 6 no aporta ninguna información al triplete.

[0077] La FIG. 13 representa una tabla de ejemplo de los modelos variantes de mejor ajuste más probables y sus correspondientes niveles de confianza corregidos a los que ha convergido el método propuesto después de 4 iteraciones. La cantidad de iteraciones puede variar según la calidad de los datos de entrada y la cantidad de muestras en el grupo, pero en general, unas pocas iteraciones son suficientes para que el método converja en resultados consistentes en un grupo.

[0078] Más generalmente, los modelos que mejor se ajustan fuera del paso (f) pueden agruparse en *q* diferentes grupos de muestras ($1 \leq q \leq n$-*1)* en base a $[V_{ij.k}^{1}|\ V_{ij.k}^{2}]$ valores tales que dentro de cada grupo de muestras $G_r (1 \leq r \leq q)$, todos los resultados concuerdan entre sí. Por lo tanto, las muestras con el resultado [NA|NA] no se consideran en este paso. El nivel de confianza general para este grupo se puede calcular como: $L_{ij.Gr} = 1 - \Pi_{k \in Gr}(1 - L_{ij.k})$.

[0079] Por ejemplo, si el grupo 1 contiene 3 muestras con niveles de confianza de análisis cruzado 0,5, 0,4 y 0,3 respectivamente, entonces $L_{ij.G1}$ = 1-(1-0,5)(1-0,4)(1-0,3) = 0,79.

[0080] Si hay más de 1 grupo (q>1) fuera del paso e), podemos elegir el grupo G_h con el nivel de confianza más alto $L_{ij.Gh}$, y establecer el valor $[V^1_{ij.Gh}|V^2_{ij.Gh}]$ de este grupo como el nuevo valor en iteración *p* para $[V^1_{ij.p}|V^2_{ij.p}]$ antes de la iteración *p* en el paso f).

[0081] El nuevo nivel de confianza para el par i y j puede entonces calcularse como aumentado o disminuido como:

$$L_{ij.nuevo} = max(0, 1-(1-L_{ij.Gh})*\prod_{1\le r\le q, r\neq h}(1-L_{ij.Gr})^{-1}$$

[0082] Por ejemplo, si hay 3 grupos G 1, G 2, G 3 con nivel de confianza 0,9, 0,8, 0,7 correspondientemente: como el grupo G 1 tiene el nivel de confianza más alto 0.9, establecemos $[V^1_{ij.G1}|V^2_{ij.G1}]$ como nuevo resultado para $[V^1_{ij.p}|V^2_{ij.p}]$. Y el nuevo nivel de confianza puede ser:

$$L_{ij.p} = max(0,1-(1-0.9)(1-0.8)^{-1}(1-0.7)^{-1}) = max(0,-0.67)=0.$$

[0083] Por todos los pasos anteriores por el paso de iteración f) todos los modelos de mejor ajuste de muestras cruzadas $[V^1_{ij}|V^2_{ij}]$ y el nivel de coincidencia L_{ij} correspondiente pueden actualizarse a nuevos resultados para todos los pares de muestras *i* y *j.* Entonces, si hay más de 1 grupo (*q* > 1) para cualquier par i y j (lo que significa que los resultados tienen conflictos), se repetirá toda la iteración, mientras que si todos los pares *i* y *j* generan solo 1 grupo (lo que significa todos los resultados concuerdan entre sí, alcanzando un valor estable), la iteración puede detenerse en el paso f).

Selección del grupo de muestras más probable correspondiente al tipo salvaje

[0084] Para eliminar el sesgo de los datos de todas las muestras, es necesario identificar qué subconjunto de los grupos identificados de modelos variantes que mejor se ajustan corresponden a la referencia de tipo salvaje, que no llevan ninguna mutación pero pueden haber sido sesgados únicamente por los procesos de flujo de trabajo de secuenciación de próxima generación. En el ejemplo de la FIG. 13, todos los modelos convergentes coinciden entre sí como [0|0], es decir, no hay mutación relativa entre sí, por lo que se puede derivar fácilmente que todas las muestras son la misma referencia de tipo salvaje (no se identificó ninguna mutación variante para reportar). Sin embargo, en el ejemplo de la FIG. 14, los modelos de variantes que mejor se ajustan después de la iteración del análisis cruzado se pueden combinar en dos grupos de muestras, a saber, el grupo 1={S1, S3, S4, S5, S6, S7, S8} y el grupo 3={S2}. En el último caso, existe la necesidad de identificar qué grupo es más probable que lleve el tipo salvaje.

[0085] En una posible forma de realización, el método refinado de llamada de variantes 244 y el analizador de datos genómicos 120 para detectar y caracterizar una variante de patrón repetido en la secuencia genómica de una muestra de paciente puede comprender identificar el subconjunto de una o más muestras correspondientes al tipo salvaje referencia en el grupo de muestras de pacientes seleccionando como el tipo salvaje el grupo de modelo de variante homocigota que mejor se ajusta [V_G | V_G] con el que se ha asociado el mayor número de muestras *i,j,…* a partir del análisis cruzado del grupo de muestras.

[0086] De manera más general, el método refinado de llamada de variantes 244 y el analizador de datos genómicos 120 para detectar e informar 250 una variante de homopolímero como el patrón repetido de al menos dos nucleótidos en la secuencia genómica de una muestra de paciente puede comprender identificar el subconjunto de uno o más muestras correspondientes a la referencia de tipo salvaje en el conjunto de muestras de pacientes seleccionando como tipo salvaje el modelo de variante homocigótica que mejor se ajusta [V_G | V_G] para los que se han cumplido hipótesis adicionales.

[0087] Por ejemplo, si el grupo G está asociado con la referencia de tipo salvaje y V_G =-1, el método refinado de llamada de variantes 244 y el analizador de datos genómicos 120 pueden identificar las variantes de patrones repetidos en la muestra del grupo G como las correspondientes a la referencia del genoma humano (longitud relativa [0|0]), independientemente de la longitud del homopolímero medida realmente más corto debido a errores de secuencia.

[0088] En el ejemplo de la FIG. 14, solo un grupo (grupo 1) está asociado con un modelo de variante de mejor ajuste homocigoto. Por lo tanto, este grupo se identificará como coincidente con el tipo salvaje, aquí sin sesgo de observación ([0|0] modelo de variante de mejor ajuste). En el otro grupo 2, una sola muestra 2 se caracterizará por el método refinado de llamada de variantes 244 y el analizador de datos genómicos 120 informará 250 como portadora de una inserción heterocigota en relación con la referencia del genoma humano ([0|1] modelo variante de mejor ajuste en relación con el modelo de variante de tipo salvaje no sesgada [0|0]).

[0089] En una posible forma de realización, cuando hay más de dos candidatos de grupo de tipo salvaje después de clasificarlos según su estado homocigoto y/o su tamaño, se pueden considerar hipótesis adicionales para garantizar que el resultado sea lo más sólido posible desde el punto de vista estadístico. Por ejemplo, todas las variantes identificadas en el grupo de muestras deberían ser biológicamente posibles. Por lo tanto, en una posible forma de realización, el método refinado de llamada de variantes 244 y el analizador de datos genómicos 120 para detectar una variante de homopolímero como el patrón repetido de al menos dos nucleótidos en la secuencia genómica de una muestra de un paciente pueden

comprender además identificar el subconjunto de una o más muestras correspondientes a la referencia de tipo salvaje en el grupo de muestras de pacientes por:

(g1) para cada posible modelo de variante de mejor ajuste $[V_G^1|V_G^2]$ identificado para un grupo G de uno o más pares de muestras, identificando si la variante es homocigota $V_G^1=V_G^2$; de lo contrario, excluir el modelo de variante de mejor ajuste $[V_G^1|V_G^2]$ como la referencia de tipo salvaje para el conjunto de muestras;
(g2) para cada grupo G del modelo de variante de mejor ajuste homocigoto $[V_G^1|V_G^2]$, para cada otro grupo *G'* del modelo de variante de mejor ajuste $[V_{G'}^1|V_{G'}^2]$, verificar si $[V_{G'}^1|V_{G'}^2]$ es una variante posible bajo la hipótesis de que el modelo de variante de mejor ajuste homocigoto $[V_G^1|V_G^2]$ es el tipo salvaje; por ejemplo, si $V_{G'}^1 < V_G^1$ y/o $V_{G'}^2 < V_G^2$, verificando que la longitud del patrón repetido con el modelo de variante de mejor ajuste $[V_{G'}^1|V_{G'}^2]$ es lo suficientemente largo para ser detectado como una variante de eliminación plausible, o si $V_{G'}^1 > V_G^1$ y/o $V_{G'}^2 > V_G^2$ verificando que la longitud del patrón repetido con el modelo de variante de mejor ajuste $[V_{G'}^1|V_{G'}^2]$ es lo suficientemente corto para ser detectado como una variante de inserción plausible; de lo contrario, excluyendo el modelo de variante de mejor ajuste $[V_G^1|V_G^2]$, como la referencia de tipo salvaje;
(g3) para cada grupo restante de modelo de variante de mejor ajuste homocigoto $[V_G^1|V_G^2]$, estimando una tasa de error basada en la longitud promedio del homopolímero h y la desviación estándar DE en cada grupo:

Si $\bar{h}$ está lo suficientemente cerca (dentro de un umbral predefinido *umbral_h*, por ejemplo en el intervalo de 0 a 0,1) al entero más cercano o $\lceil \bar{h} \rfloor$, es decir si abs($\bar{h}$ - o $\lceil \bar{h} \rfloor$) < *umbral_h,* y si DE es lo suficientemente pequeño para estar por debajo de un umbral predefinido *umbral_de* (seleccionado por ejemplo en el intervalo de 0 a 0,1), es decir, y si DE < *umbral_de,* seleccionando el modelo de variante de mejor ajuste homocigoto $[V_G^1|V_G^2]$ como la referencia de tipo salvaje con una tasa de error baja e informando todas las muestras *i,j,.* asociadas con este grupo G como portadoras el patrón de repetición de referencia del genoma humano;
En caso contrario, seleccionando como tipo salvaje el grupo de homocigotos modelo de variante de mejor ajuste $[V_G^1|V_G^2]$ que comprende el mayor número de muestras.

[0090] La FIG. 15 muestra un posible flujo de trabajo simplificado de la lógica de selección de grupos de tipo salvaje anterior que puede aplicarse, con un procesador, mediante el método de refinamiento de llamada de variante propuesto 244.

[0091] El analizador de datos genómicos 120 propuesto permite servir a miles de laboratorios de abastecimiento, procesamiento de los datos de cientos de miles de muestras clínicas procesadas con múltiples tecnologías de enriquecimiento y secuenciadas en una diversidad de plataformas de secuenciación de próxima generación (NGS). Al utilizar este rico conjunto de datos junto con los métodos de análisis de datos genómicos propuestos, se pueden lograr resultados sólidos y precisos de llamada de variantes con la sensibilidad y especificidad del flujo de trabajo automatizado propuesto que coincide con la configuración manual del algoritmo y el ajuste por parte de expertos en bioinformática. Además, el analizador de datos genómicos 120 totalmente automatizado propuesto puede implementarse, probarse y validarse sin necesidad de una configuración individual y un ajuste fino de su flujo de trabajo de análisis genómico NGS específico por parte de los laboratorios de abastecimiento y, por lo tanto, acelerará el acceso a la medicina personalizada y de precisión para cientos de personas de miles de pacientes en Europa y en todo el mundo.

Datos experimentales

[0092] El analizador de datos genómicos propuesto 120 se ha adaptado en la plataforma de software de análisis genómico de Sophia Genetics Data Driven Medicine (DDM) para implementar el método propuesto como un método complementario para mejorar la detección 244 y el informe 250 de variantes de homopolímero en genes humanos sobre los flujos de trabajo NGS de la técnica anterior.

[0093] La FIG. 16) muestra la cobertura de datos de NGS para un patrón de homopolímero repetido ejemplar en el gen CHEK2 de un grupo experimental que comprende una mezcla de muestras de pacientes mutados y muestras de tipo salvaje, y la FIG. 17) muestra la distribución medida correspondiente de las longitudes del patrón. El flujo de trabajo de secuenciación introduce un sesgo de una eliminación, por lo que la longitud del patrón de repetición de CHEK2 para el tipo salvaje se centra en 12 repeticiones en lugar de 13. El grupo de experimentos también comprende muestras centradas en 10 repeticiones. Gracias al método propuesto, se corrige el sesgo y el analizador de datos genómicos informa correctamente tanto de las variantes de tipo salvaje (sin mutación) como de las variantes corregidas (11 repeticiones, es decir, dos deleciones del patrón relativo al tipo salvaje).

[0094] La FIG. 18) muestra la cobertura de datos de NGS para un patrón de homopolímero repetido ejemplar en el gen RAD54L de un conjunto de experimentos que comprende una mezcla de muestras de pacientes mutados y muestras de tipo salvaje, y la FIG. 19) muestra la distribución medida correspondiente de las longitudes del patrón. El flujo de trabajo de secuenciación introduce un sesgo de eliminación de dos en dos, por lo que la longitud del patrón de repetición de RAD54L para el tipo salvaje se centra en 12 repeticiones en lugar de 14. El grupo de experimentos también comprende muestras centradas en 11 repeticiones. Gracias al método propuesto, se corrige el sesgo y el analizador de datos genómicos informa correctamente tanto las variantes de tipo salvaje (sin mutación) como las variantes corregidas (12 repeticiones, es decir, una eliminación del patrón con respecto al tipo salvaje).

[0095] La FIG. 20) muestra un diagrama de la cobertura de datos de NGS para un patrón de homopolímero repetido ejemplar en el gen ATM de un grupo de experimentos que comprende una mezcla de muestras de pacientes mutados y muestras de tipo salvaje, y la FIG. 21) muestra la distribución medida correspondiente de las longitudes del patrón. El flujo de trabajo de secuenciación introduce un sesgo de una eliminación, por lo que la longitud del patrón de repetición ATM para el tipo salvaje se centra en 15 repeticiones en lugar de 19. El grupo de experimentos también comprende muestras centradas en 13 repeticiones. Gracias al método propuesto, se corrige el sesgo y el analizador de datos genómicos informa correctamente tanto de las variantes de tipo salvaje (sin mutación) como de las variantes corregidas (17 repeticiones, es decir, dos deleciones del patrón relativo al tipo salvaje).

[0096] La FIG. 22) muestra un diagrama de la cobertura de datos de NGS para un patrón de heteropolímero repetido ejemplar en el gen CFTR de un conjunto de experimentos que comprende una mezcla de muestras de pacientes con diferentes mutaciones, y la FIG. 23) muestra la correspondiente distribución medida de las longitudes del patrón. La longitud del patrón de repetición de CFTR TG para el tipo salvaje se centra en 11 repeticiones, igual que la longitud de referencia en el genoma humano; sin embargo, el flujo de trabajo de secuenciación introdujo un sesgo de una eliminación en aproximadamente el 30 % de las secuencias, que es similar a una supresión heterocigota. El grupo experimental también comprende todas las combinaciones de muestras en las que cada alelo centró entre 10 y 12 repeticiones. Gracias al método propuesto se corrige el sesgo y el analizador de datos genómicos reporta correctamente las variantes de tipo salvaje (sin mutación) así como todas las combinaciones de las variantes.

[0097] Por lo tanto, mediante la optimización de un ensayo NGS basado en un panel de mutación con los métodos bioinformáticos propuestos, es posible proporcionar sensibilidad y especificidad mejoradas comparables a los métodos clásicos de secuenciación de Sanger a un costo significativamente menor para el análisis genómico. Los métodos propuestos también mejoran el rendimiento de los flujos de trabajo NGS de la técnica anterior en la detección de esas mutaciones, independientemente del sesgo experimental introducido por las plataformas de secuenciación de próxima generación. Las aplicaciones actuales incluyen diagnósticos, cribado neonatal y de portadores para una diversidad de enfermedades hereditarias y somáticas como el cáncer.

Otras formas de realización y aplicaciones

[0098] En particular, como será evidente para los expertos en la técnica de la genómica y la medicina personalizada, los métodos propuestos no se limitan a las especificaciones de las construcciones génicas de repeticiones de nucleótidos cortos (homopolímero o heteropolímero) como se describe en ciertos ejemplos a lo largo de esta divulgación. Si bien en este documento se han descrito ejemplos para el análisis de los genes CHEK2, RAD54L, ATM, BRCA y CFTR, otras áreas de la práctica de la genómica médica aún están en curso de investigación para asociar el análisis genómico de ciertas regiones genéticas con ciertas patologías. El método refinado de llamada de variantes propuesto 244 puede así aplicarse para mejorar la detección de variantes en otras regiones genómicas distintas de estos genes ejemplares si tales regiones se caracterizan por patrones repetidos y se asocian con diferentes diagnósticos por futuros trabajos de investigación médica. Este puede ser el caso, por ejemplo, en el campo de las enfermedades neurológicas, donde el desarrollo reciente en modelos animales editados con genoma está acelerando el estudio de múltiples mutaciones, mientras que ya existen asociaciones conocidas entre ciertas repeticiones de heteropolímeros y ciertas enfermedades, como por ejemplo las variantes repetidas de poliglutamina CAG (polyQ) en Huntington, así como enfermedades de ataxia.

[0099] Como será evidente para los expertos en la técnica de las comunicaciones de datos digitales, los métodos descritos en el presente documento pueden aplicarse indiferentemente a varias estructuras de datos tales como archivos de datos o flujos de datos. Los términos "datos", "estructuras de datos", "campos de datos", "archivo" o "flujo" pueden utilizarse indistintamente a lo largo de esta especificación.

[0100] Como será evidente para los expertos en estadísticas de la técnica, los métodos descritos en el presente documento pueden aplicarse indiferentemente a varios métodos estadísticos tales como representaciones de probabilidad y mediciones estadísticas. Los términos "distribución", "verosimilitud", "probabilidad" pueden, por tanto, utilizarse indistintamente a lo largo de esta memoria descriptiva. Aunque la descripción detallada anterior contiene muchos detalles específicos, estos no deben interpretarse como limitantes del alcance de las formas de realización sino simplemente como ilustraciones de algunas de varias formas de realización.

[0101] Aunque anteriormente se han descrito varias formas de realización, debe entenderse que se han presentado a modo de ejemplo y no de limitación. De hecho, después de leer la descripción anterior, será evidente para un experto en la(s) técnica(s) relevante(s) cómo implementar formas de realización alternativas.

[0102] Además, debe entenderse que cualquier figura que destaque la funcionalidad y las ventajas se presenta solo con fines de ejemplo. Los métodos descritos son lo suficientemente flexibles y configurables como para que puedan utilizarse de formas distintas a las mostradas.

[0103] Aunque el término "al menos uno" se puede usar a menudo en la memoria descriptiva, las reivindicaciones y los dibujos, los términos "un", "una", "el", "dicho", etc. también significan "al menos uno" o "el al menos uno" en la memoria descriptiva, reivindicaciones y dibujos.

[0104] A lo largo de esta especificación, varias instancias pueden implementar componentes, operaciones o estructuras descritas como una sola instancia. Aunque las operaciones individuales de uno o más métodos se ilustran y describen como operaciones separadas, una o más de las operaciones individuales pueden realizarse al mismo tiempo, y nada requiere que las operaciones se realicen en el orden ilustrado. Las estructuras y funcionalidades presentadas como componentes separados en configuraciones de ejemplo pueden implementarse como una estructura o componente combinado. De manera similar, las estructuras y la funcionalidad presentadas como un solo componente pueden implementarse como componentes separados. Estas y otras variaciones, modificaciones, adiciones y mejoras caen dentro del alcance del objeto de este documento.

[0105] Ciertas formas de realización se describen aquí como que incluyen lógica o una serie de componentes, módulos, unidades o mecanismos. Los módulos o unidades pueden constituir módulos de software (por ejemplo, código incorporado en un medio legible por máquina o en una señal de transmisión) o módulos de hardware. Un "módulo de hardware" es una unidad tangible capaz de realizar ciertas operaciones y puede configurarse o disponerse de cierta manera física. En varias formas de realización de ejemplo, uno o más sistemas informáticos (*p. ej.,* un sistema informático independiente, un sistema informático cliente o un sistema informático servidor) o uno o más módulos de hardware de un sistema informático (*p. ej.,* un procesador o un grupo de procesadores) puede ser configurado por software (por ejemplo, una aplicación o parte de la aplicación) como un módulo de hardware que opera para realizar ciertas operaciones como se describe aquí.

[0106] En algunas formas de realización, un módulo de hardware puede implementarse mecánicamente, electrónicamente o en cualquier combinación adecuada de los mismos. Por ejemplo, un módulo de hardware puede incluir lógica o circuitos dedicados que están permanentemente configurados para realizar ciertas operaciones. Por ejemplo, un módulo de hardware puede ser un procesador de propósito especial, como una matriz de puertas programables en campo (FPGA) o un ASIC. Un módulo de hardware también puede incluir lógica o circuitos programables que se configuran temporalmente mediante software para realizar ciertas operaciones. Por ejemplo, un módulo de hardware puede incluir software incluido dentro de un procesador de propósito general u otro procesador programable. Se apreciará que la decisión de implementar un módulo de hardware mecánicamente, en circuitos dedicados y permanentemente configurados, o en circuitos configurados temporalmente (*p. ej.,* configurados por software) puede ser impulsada por consideraciones de costo y tiempo.

[0107] Las diversas operaciones de los métodos de ejemplo descritos en este documento pueden ser realizadas, al menos parcialmente, por uno o más procesadores que están configurados temporalmente (por ejemplo, por software) o configurados permanentemente para realizar las operaciones relevantes. Ya sea que estén configurados de manera temporal o permanente, dichos procesadores pueden constituir módulos implementados por el procesador que operan para realizar una o más operaciones o funciones descritas en este documento. Como se usa aquí, "módulo implementado por proceso" se refiere a un módulo de hardware implementado usando uno o más procesadores.

[0108] De manera similar, los métodos descritos en el presente documento pueden implementarse al menos parcialmente en un procesador, siendo un procesador un ejemplo de hardware. Por ejemplo, al menos algunas de las operaciones de un método pueden ser realizadas por uno o más procesadores o módulos implementados por procesador.

[0109] Algunas partes del tema discutido en este documento pueden presentarse en términos de algoritmos o representaciones simbólicas de operaciones en datos almacenados como bits o señales digitales binarias dentro de una memoria de máquina (por ejemplo, una memoria de computadora). Dichos algoritmos o representaciones simbólicas son ejemplos de técnicas utilizadas por los expertos en las técnicas del procesamiento de datos para transmitir la esencia de su trabajo a otros expertos en la materia. Tal como se utiliza aquí, un "algoritmo" es una secuencia autoconsistente de operaciones o un procesamiento similar que conduce a un resultado deseado. En este contexto, los algoritmos y las operaciones involucran la manipulación física de cantidades físicas.

[0110] Por ejemplo, varias formas de realización o características de las mismas pueden mezclarse y combinarse o hacerse opcionales por un experto en la materia. Dichas formas de realización de la materia inventiva pueden ser referidas en este documento, individual o colectivamente, por el término "invención" simplemente por conveniencia y sin la intención de limitar voluntariamente el alcance de esta solicitud a una sola invención o concepto inventivo si hay más de uno, de hecho, revelado.

[0111] Se cree que las formas de realización ilustradas en el presente documento se describen con suficiente detalle para permitir que los expertos en la técnica pongan en práctica las enseñanzas descritas. Se pueden usar otras formas de realización y derivarse de ellas, de modo que se puedan realizar sustituciones y cambios estructurales y lógicos sin apartarse del alcance de esta divulgación. La Descripción Detallada, por lo tanto, no debe tomarse en un sentido limitativo, y el alcance de varias formas de realización está definido únicamente por las reivindicaciones adjuntas, junto con la gama completa de equivalentes a los que tales reivindicaciones tienen derecho. Además, se pueden proporcionar instancias plurales para recursos, operaciones o estructuras descritas en este documento como una sola instancia. Además, los límites entre varios recursos, operaciones, módulos, motores y almacenes de datos son algo arbitrarios, y las operaciones particulares se ilustran en un contexto de configuraciones ilustrativas específicas. Se contemplan otras asignaciones de funcionalidad y pueden caer dentro del alcance de varias formas de realización de la presente invención. En general, las estructuras y funcionalidades presentadas como recursos separados en las configuraciones de ejemplo pueden

implementarse como una estructura o recurso combinado. De manera similar, las estructuras y funcionalidades presentadas como un solo recurso pueden implementarse como recursos separados. Estas y otras variaciones, modificaciones, adiciones y mejoras se encuentran dentro del alcance de las formas de realización de la presente invención como se representa en las reivindicaciones adjuntas. Por consiguiente, la especificación y los dibujos deben considerarse en un sentido ilustrativo más que restrictivo.

[0112] Finalmente, es la intención del solicitante que solo las reivindicaciones que incluyen el lenguaje expreso "significa para" o "paso para" se interpreten bajo 35 USC 112, párrafo 6. Las reivindicaciones que no incluyen expresamente la frase "significa para" o "paso para" no debe interpretarse según 35 USC 112, párrafo 6.

REIVINDICACIONES

1. Un método para detectar y reportar, con un procesador, una variante como una repetición de al menos dos patrones de nucleótidos en una secuencia genómica de una muestra de un paciente, el método comprende:

(a) identificar un patrón repetido de referencia $P_{ref} = N*I$ como la repetición de $I(I>=2)$ patrones genómicos N en una región genómica de una secuencia de referencia del genoma humano;
(b) obtener, con un secuenciador de próxima generación, n conjuntos de pacientes de lecturas de datos de secuenciación de próxima generación que cubran la región genómica del patrón repetido de referencia $S= \{S_1, S_2, ..., S_i, ..., S_n\}$ de un conjunto de n muestras de pacientes genómicos enriquecidos, estando asociado cada conjunto Si con una muestra de paciente, siendo el número n de muestras de pacientes genómicos enriquecidos al menos 4;
(c) para cada muestra de paciente i en el conjunto S de muestras de pacientes, medir la distribución P_i de la longitud del patrón repetido en el conjunto de lecturas de secuenciación de próxima generación S_i;
(d) para un posible par de muestras de pacientes i y j, $j>i$, estimando un modelo de mejor ajuste $[V_{ij}^1 | V_{ij}^2]$ de las dos variantes alélicas para la muestra i relativo a la muestra j, con un nivel de confianza L_{ij};
(e) para cada posible triplete de muestras de pacientes i, $j>i$, $k>j$, comparar sus respectivos modelos de mejor ajuste $[V_{ji}^1 | V_{ji}^2]$ $[V_{jk}^1 | V_{jk}^2]$, $[V_{ik}^1 | V_{ik}^2]$, agrupando grupos de modelos variantes de mejor ajuste con un mayor nivel de confianza, e iterar la comparación hasta que se formen grupos estables de modelos variantes que mejor se ajusten;
(f) identificar el grupo más probable que lleva la variante de tipo salvaje;
(g) para cada muestra en el grupo que lleva la variante de tipo salvaje, reportar la variante de la muestra como el patrón repetido de referencia de tipo salvaje $P_{ref} = N*I$.
(h) para cada muestra del grupo que lleva la variante de tipo salvaje, eliminar el sesgo del modelo de variante de mejor ajuste del grupo que comprende esta muestra en función del modelo de variante de mejor ajuste para el grupo de tipo salvaje identificado e informar la variante de la muestra como la variante del modelo de mejor ajuste no sesgada.

2. El método de la reivindicación 1, en el que estimar un modelo de mejor ajuste $[V_{ij}^1 | V_{ij}^2]$ de las dos variantes alélicas para la muestra i en relación con la muestra j comprende:

(d1) estimación para la muestra j, bajo el supuesto de que la muestra i lleva el patrón de homopolímero de referencia del genoma humano de tipo salvaje $P_{ref}=N*I$ para cada alelo, un modelo de mejor ajuste $[V_{ji}^1 | V_{ji}^2]$ de las dos variantes alélicas para la muestra j, con un nivel de confianza L_{ji}, así como la distancia más pequeña D_{ji} entre la distribución medida P_j para la muestra j y la distribución unimodal o bimodal predicha para el modelo de variante de mejor ajuste $[V_{ji}^1 | V_{ji}^2]$;
(d2) estimando para la muestra i, bajo el supuesto de que la muestra j lleva el patrón de homopolímero de referencia del genoma humano de tipo salvaje $P_{ref} =N*I$ para cada alelo, un modelo de mejor ajuste $[V_{ji}^1 | V_{ji}^2]$ de las dos variantes de alelo para la muestra i, con un nivel de confianza L_{ij}, así como la distancia más pequeña D_{ij} entre la distribución medida P_i para la muestra i y la distribución unimodal o bimodal pronosticada para el modelo de variante de mejor ajuste $[V_{ji}^1 | V_{ji}^2]$;
(d3) si $D_{i|j} \geq D_{j|i}$, seleccionando para el par de muestras (i,j) el modelo de mejor ajuste $[V_{ij}^1 | V_{ij}^2] = [-V_{ji}^1 | -V_{ji}^2]$ como el modelo variante de mejor ajuste de las dos variantes alélicas y el nivel de confianza $L_{ij} = L_{j|i}$ como el valor del nivel de confianza para esta coincidencia de mejor ajuste con la muestra i como la muestra de referencia del par (i,j);
(d4) de lo contrario si $D_{i|j} < D_{j|i}$ seleccionando para el par de muestras (i,j) el modelo $[V_{ij}^1 | V_{ij}^2] = [V_{ij}^1 | V_{ij}^2]$ como el modelo de variante de mejor ajuste de las dos variantes de alelo y el nivel de confianza $L_{ij} = L_{i|j}$ como el valor del nivel de confianza para esta coincidencia de mejor ajuste con la muestra j como la muestra de referencia del par (i,j);

3. El método de la reivindicación 2, que comprende además estimar un modelo secundario de mejor ajuste $[V_{ji}^{1\prime} | V_{ji}^{2\prime}]$ de las dos variantes alélicas para la muestra j, bajo el supuesto de que la muestra i lleva el patrón de homopolímero de referencia del genoma humano de tipo salvaje $P_{ref} = N*I$ para cada alelo, así como la segunda distancia más pequeña D_{ji}' entre la distribución medida P_j para la muestra j y la distribución unimodal o bimodal pronosticada para el modelo de variante de mejor ajuste secundario $[V_{ji}^1 | V_{ji}^2]$, estimando un modelo de mejor ajuste secundario $[V_{ij}^{1\prime} | V_{ij}^{2\prime}]$ de las dos variantes alélicas para muestra i, bajo el supuesto de que la muestra j lleva el patrón de homopolímero de referencia del genoma humano de tipo salvaje $P_{ref} = N*I$ para cada alelo, así como la segunda distancia más pequeña D_{ij}' entre la distribución medida P_i para la muestra i y la distribución unimodal o bimodal predicha para el modelo de variante de mejor ajuste secundario $[V_{ji}^1 | V_{ji}^2]$, y calculando el nivel de confianza L_{ij} de la estimación $[V_{ij}^1 | V_{ij}^2]$ como:

$$L_{ij} = \begin{cases} 1 - D_{i|j}/D_{i|j}', if\ D_{i|j} > D_{j|i} \\ 1 - D_{j|i}/D_{j|i}', if\ D_{i|j} \leq D_{j|i} \end{cases}$$

4. El método de las reivindicaciones 1, 2 o 3, que comprende además agrupar en q diferentes grupos de muestras ($1 \leq q \leq n$-1) en base a $[V_{ij.k}^1 | V_{ij.k}^2]$ valores tales que dentro de cada grupo de muestras G_r ($1 \leq r \leq q$), todos los resultados concuerden entre sí, y calculando el nivel de confianza general para este grupo como $L_{ij.Gr} = 1-\prod_{k \in Gr}(1 - L_{ij.k})$.

5. El método de la reivindicación 4, en el que los modelos de mejor ajuste $[V_{ij}^1 | V_{ij}^2]$ correspondientes a diferentes tipos de

mutaciones heterocigóticas se excluyen de la agrupación de modelos de mejor ajuste coincidentes.

6. El método de las reivindicaciones 4 o 5, que comprende seleccionar el grupo G_h con el nivel de confianza más alto $L_{ij.Gh}$, estableciendo el valor $[V^1_{ij.Gh} | V^2_{ij.Gh}]$ de todas las muestras en este grupo y calcular el nuevo nivel de confianza para el par *i,j* como $Lij._{nuevo} = max(0, 1-(1- L_{ij.Gh}) * \Pi_{1 \leq r \leq q, r \neq h}(1 - L_{ij.Gr})^{-1}$.

7. El método de cualquiera de las reivindicaciones 4 a 6 que comprende además agrupar de nuevo grupos de muestras hasta que los resultados sean estables

8. El método de cualquiera de las reivindicaciones 1 a 7, que comprende para cada triplete posible de muestras de pacientes i, *j>i, k>j,* comparar sus respectivos modelos de mejor ajuste $[V^1_{ij} | V^2_{ij}]$, $[V^1_{jk} | V^2_{jk}]$, $[V^1_{ik} | V^2_{ik}]$, y si los tres modelos de mejor ajuste para el triplete de muestras de pacientes coinciden entre sí, aumentando sus niveles de confianza L_{ij}, L_{jk}, L_{ik}; de lo contrario, los tres modelos de mejor ajuste no coinciden entre sí, reemplazando el modelo de mejor ajuste con el nivel de confianza más bajo fuera del subconjunto por un modelo de mejor ajuste calculado a partir de las otras dos muestras del subconjunto, y la disminución de los niveles de confianza L_{ij}, L_{jk}, L_{ik} de todos los modelos de mejor ajuste

9. El método de la reivindicación 8 en el que los niveles de confianza para cada par en un subconjunto triplete *i, j, k* de modelos coincidentes de mejor ajuste se incrementan como $L_{ij}' = 1-(1- L_{ij})(1 - L_{jk} * L_{ik})$, $L_{jk}' = 1-(1 - L_{jk})(1 - L_{ij} * L_{ik})$ y $L_{ik}' = 1-(1- L_{ik})(1- L_{ij} * L_{jk})$.

10. El método de las reivindicaciones 8 o 9, en el que el nivel de confianza inicial más bajo es L_{ik} para el par *j,k* dentro del triplete, los niveles de confianza para cada par en un subconjunto de triplete *i, j, k* de modelos de mejor ajuste no coincidentes son disminuido como $L_{ij}' = L_{ij}-(1 - L_{jk}) * L_{ik}$, $L_{jk}' = L_{jk}-(1 - L_{ij}) * L_{ik}$ y $L_{ik}' = \max(0, L_{ij} * L_{jk} - L_{ik})$ y el modelo de mejor ajuste para el par *j, k* con el nivel de confianza más bajo fuera del subconjunto se reemplaza por $[V^1_{ik}' | V^2_{ik}'] = [V^1_{ij} + V^1_{ij} | V^2_{jk} + V^2_{jk}]$.

11. El método de la reivindicación 10, en el que identificar el subconjunto de una o más muestras correspondientes a la referencia de tipo salvaje en el conjunto de muestras de pacientes consiste en seleccionar como tipo salvaje el grupo de modelo de variante homocigótica que mejor se ajusta $[V_G | V_G]$ con el que se ha asociado el mayor número de muestras *i, j,*… a partir del análisis cruzado del grupo de muestras.

12. El método de la reivindicación 10 u 11, en el que identificar el subconjunto de una o más muestras correspondientes a la referencia de tipo salvaje en el grupo de muestras de pacientes comprende verificar para cada grupo G del modelo de variante de mejor ajuste homocigoto $[V^1_G | V^2_G]$, si $[V^1_{G'} | V^2_{G'}]$ es una variante posible para cada otro grupo G' del modelo de variante de mejor ajuste $[V^1_{G'} | V^2_{G'}]$, bajo la hipótesis de que el modelo de variante de mejor ajuste homocigoto $[V^1_G | V^2_G]$ es el tipo salvaje, y si ese no es el caso, excluyendo el grupo G como portador del patrón de tipo salvaje;

13. El método de la reivindicación 12, en el que $V^1_{G'} < V^1_G$ y/o $V^2_{G'} < V^2_G$ comprende verificar que la longitud del patrón repetido con el modelo de variante de mejor ajuste $[V^1_{G'} | V^2_{G'}]$ es lo suficientemente larga para ser detectada como una variante de eliminación plausible;

14. El método de la reivindicación 12 o 13, en el que $V^1_{G'} > V^1_G$ y/o $V^2_{G'} > V^2_G$ comprende verificar que la longitud del patrón repetido con el modelo de variante de mejor ajuste $[V^1_{G'} | V^2_{G'}]$ es lo suficientemente corto para ser detectado como una variante de inserción plausible;

15. El método de las reivindicaciones 10, 12, 13 o 14, que comprende además estimar una tasa de error basada en la longitud de homopolímero promedio h y la desviación estándar DE para cada grupo plausible de modelo de variante de mejor ajuste homocigoto $[V^1_G | V^2_G]$, y si $\bar{h}$ está dentro de un umbral predefinido *umbral_h* al más número entero más cercano $\lceil \bar{h} \rfloor$, es decir, si $\text{abs}(\bar{h} - \lceil \bar{h} \rfloor) <$ *umbral_h,* y si DE es lo suficientemente pequeño como para estar por debajo de un umbral predefinido umbral_de, es decir, si DE < *umbral_de,* seleccionando el modelo de variante de mejor ajuste homocigota $[V^1_G | V^2_G]$ que mejor se ajusta como modelo referencia de tipo salvaje con la tasa de error más baja.

16. El método de la reivindicación 15, en el que el *umbral_h* se elige en el intervalo de 0 a 0,1.

17. El método de la reivindicación 15 o 16, en el que *umbral_de* se elige en el intervalo de 0 a 0,1.

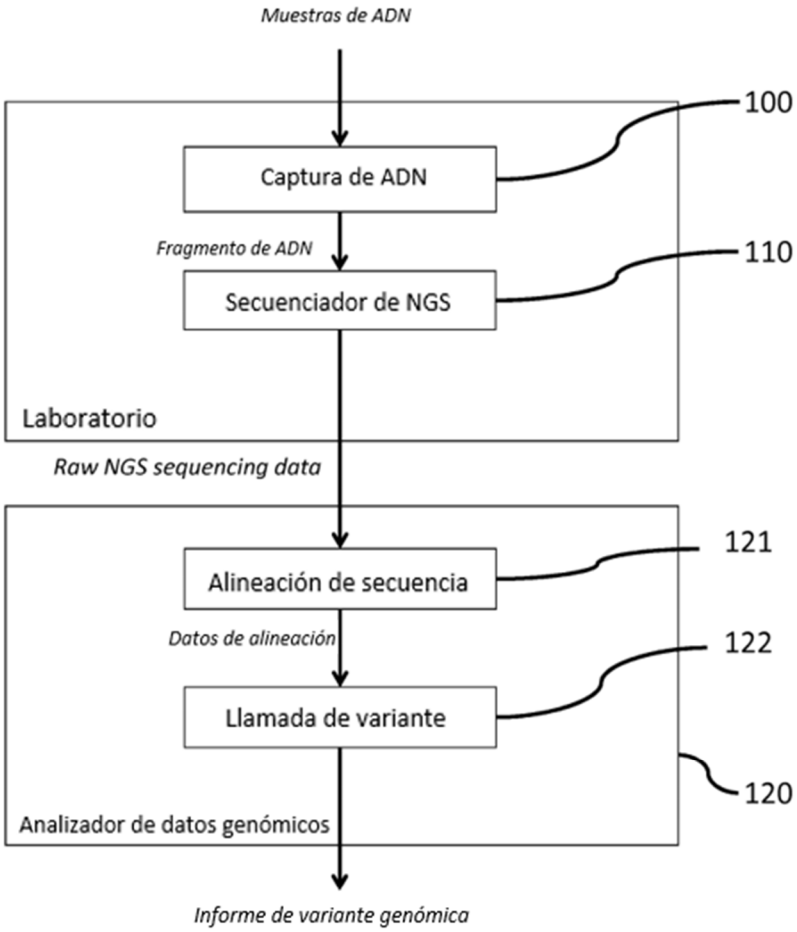


FIG.1 (técnica anterior)

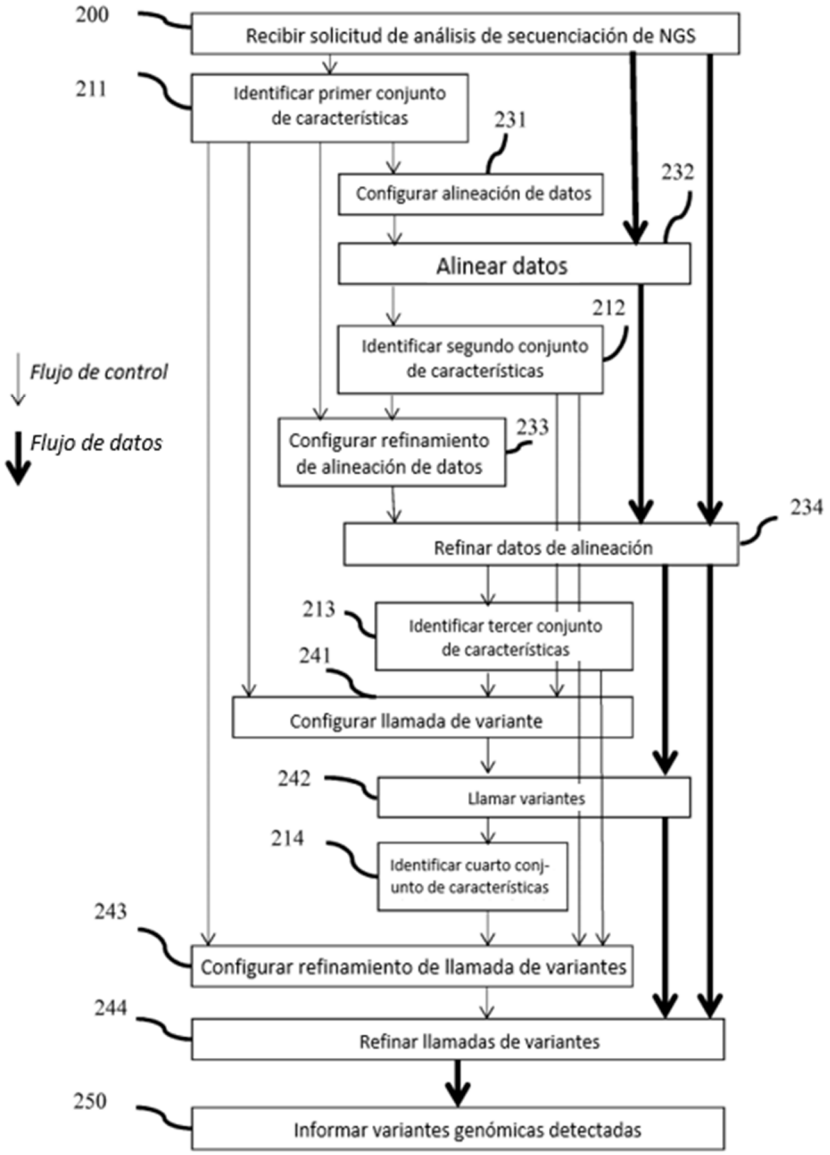

200
Recibir solicitud de análisis de secuenciación de NGS
211
Identificar primer conjunto de características
231
Configurar alineación de datos
232
Alinear datos
212
Identificar segundo conjunto de características
233
Configurar refinamiento de alineación de datos
234
Refinar datos de alineación
213
Identificar tercer conjunto de características
241
Configurar llamada de variante
242
Llamar variantes
214
Identificar cuarto conjunto de características
243
Configurar refinamiento de llamada de variantes
244
Refinar llamadas de variantes
250
Informar variantes genómicas detectadas
Flujo de control
Flujo de datos

FIG.2

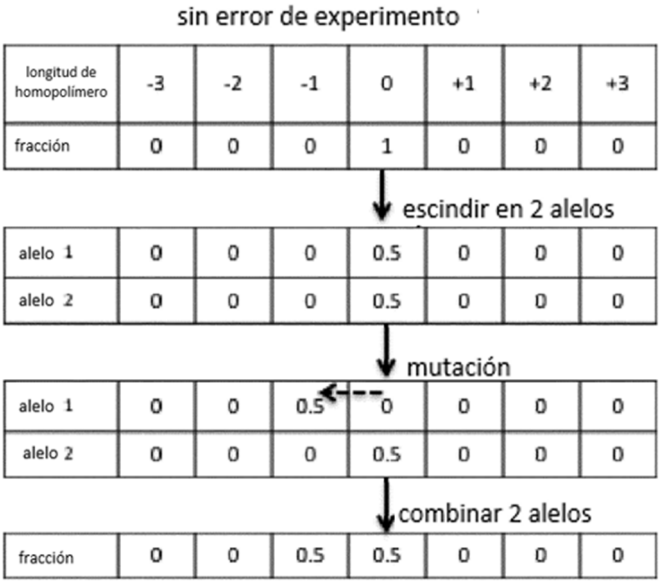

sin error de experimento
longitud de homopolímero
-3
-2
-1
0
+1
+2
+3
fracción
0 0 0 1 0 0 0
escindir en 2 alelos
alelo 1
0 0 0 0.5 0 0 0
alelo 2
0 0 0 0.5 0 0 0
mutación
alelo 1
0 0 0.5 0 0 0 0
alelo 2
0 0 0 0.5 0 0 0
combinar 2 alelos
fracción
0 0 0.5 0.5 0 0 0

FIG.3

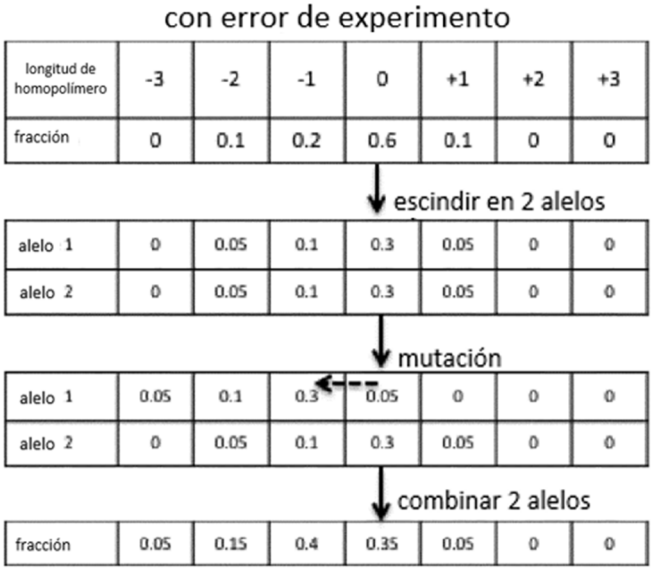

con error de experimento
longitud de homopolímero
-3
-2
-1
0
+1
+2
+3
fracción
0
0.1
0.2
0.6
0.1
0
0
escindir en 2 alelos
alelo 1
0
0.05
0.1
0.3
0.05
0
0
alelo 2
0
0.05
0.1
0.3
0.05
0
0
mutación
alelo 1
0.05
0.1
0.3
0.05
0
0
0
alelo 2
0
0.05
0.1
0.3
0.05
0
0
combinar 2 alelos
fracción
0.05
0.15
0.4
0.35
0.05
0
0

FIG.4

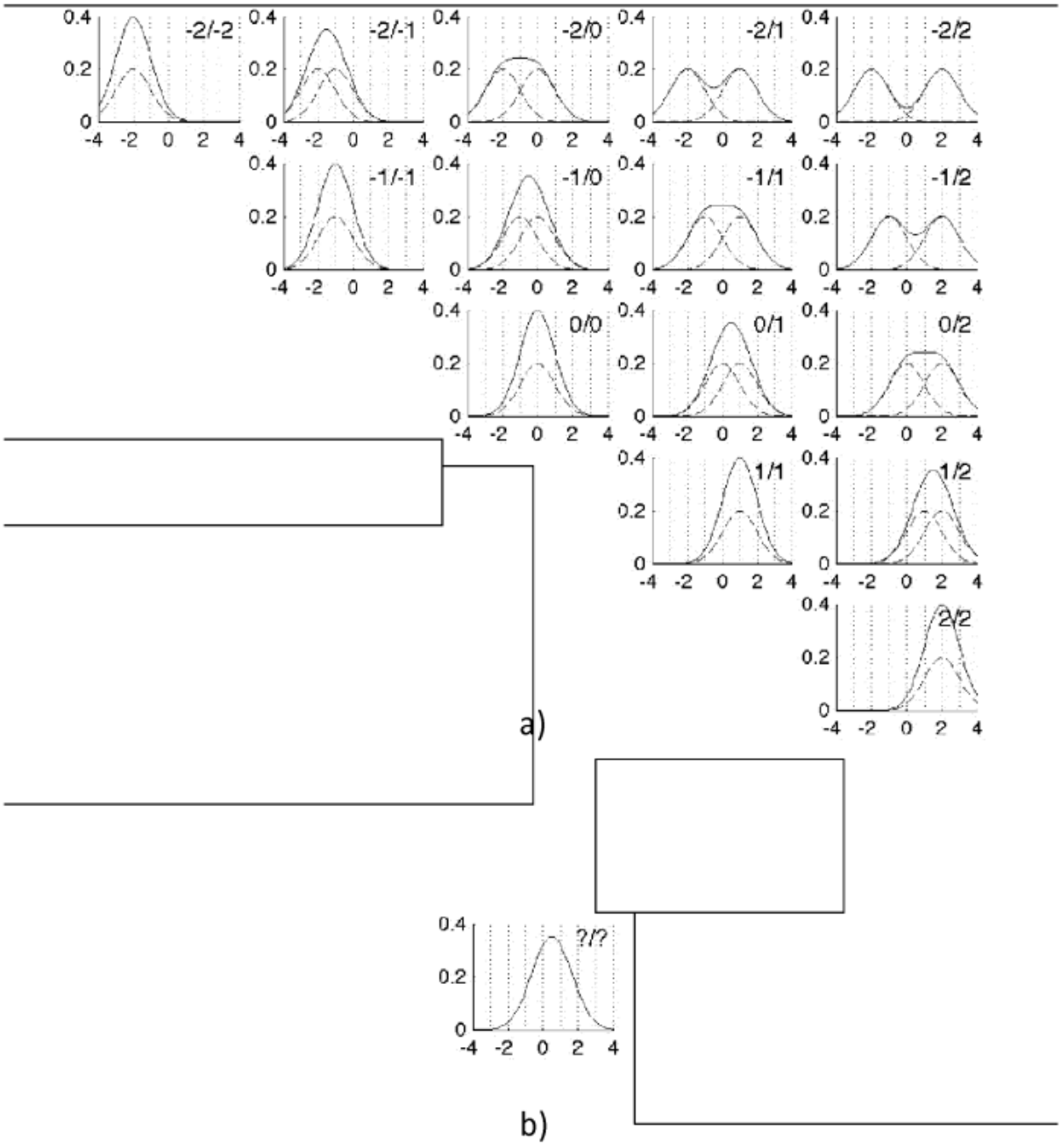


FIG.5

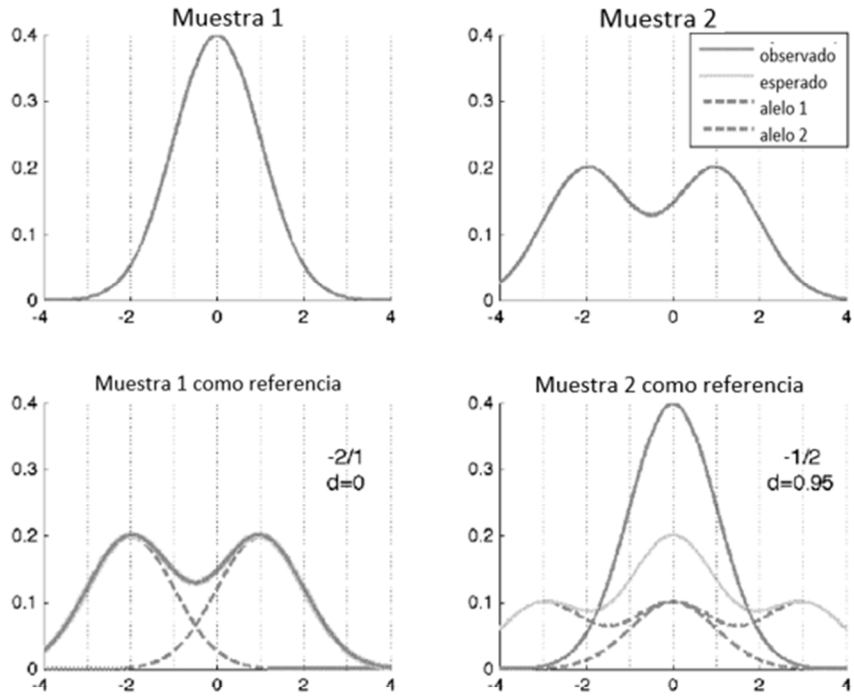

Muestra 1
Muestra 2
observado
esperado
alelo 1
alelo 2
Muestra 1 como referencia
-2/1
d=0
Muestra 2 como referencia
-1/2
d=0.95

FIG.6

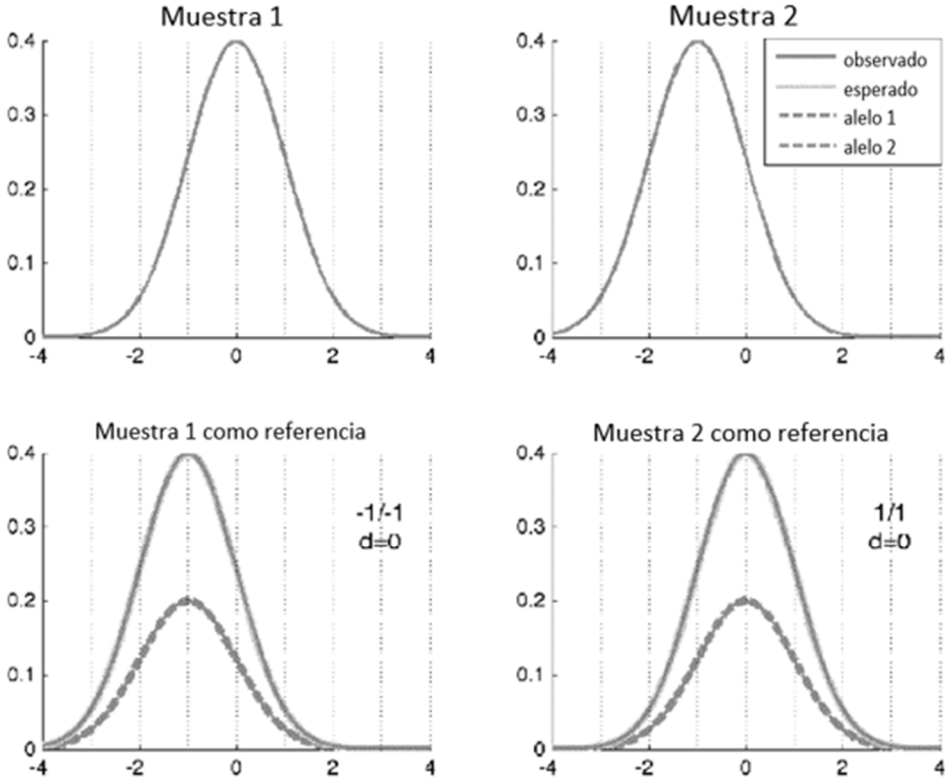

Muestra 1
Muestra 2
observado
esperado
alelo 1
alelo 2
Muestra 1 como referencia
-1/-1
d=0
Muestra 2 como referencia
1/1
d=0

FIG.7

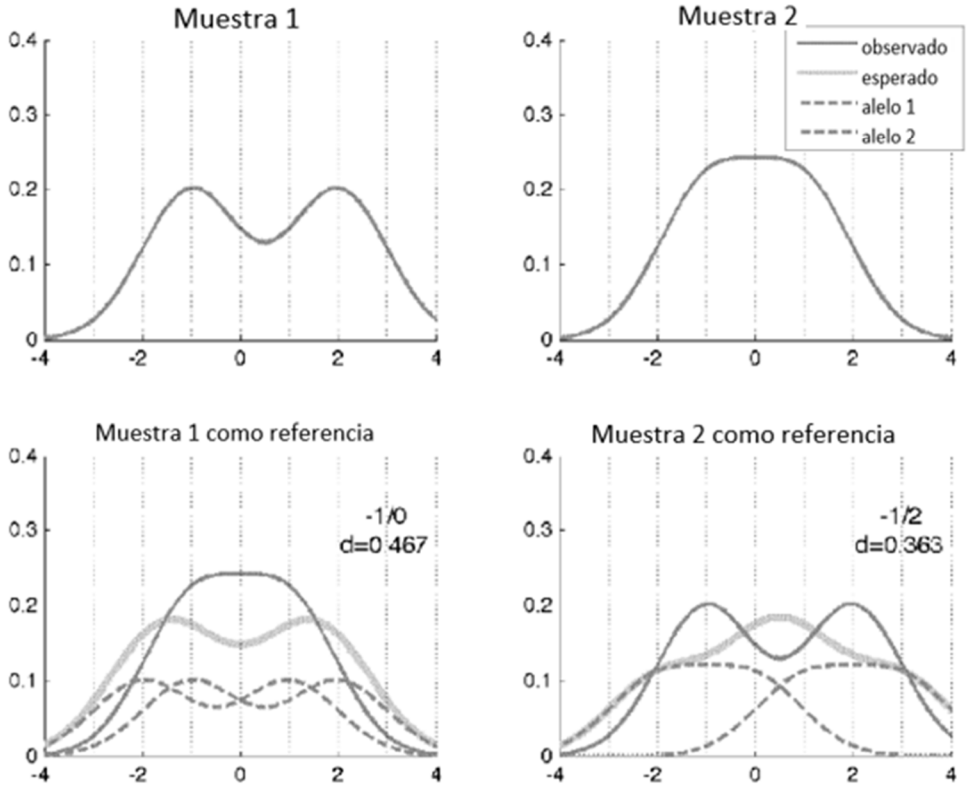

Muestra 1
Muestra 2
observado
esperado
alelo 1
alelo 2
Muestra 1 como referencia
-1/0
d=0.467
Muestra 2 como referencia
-1/2
d=0.363

FIG.8

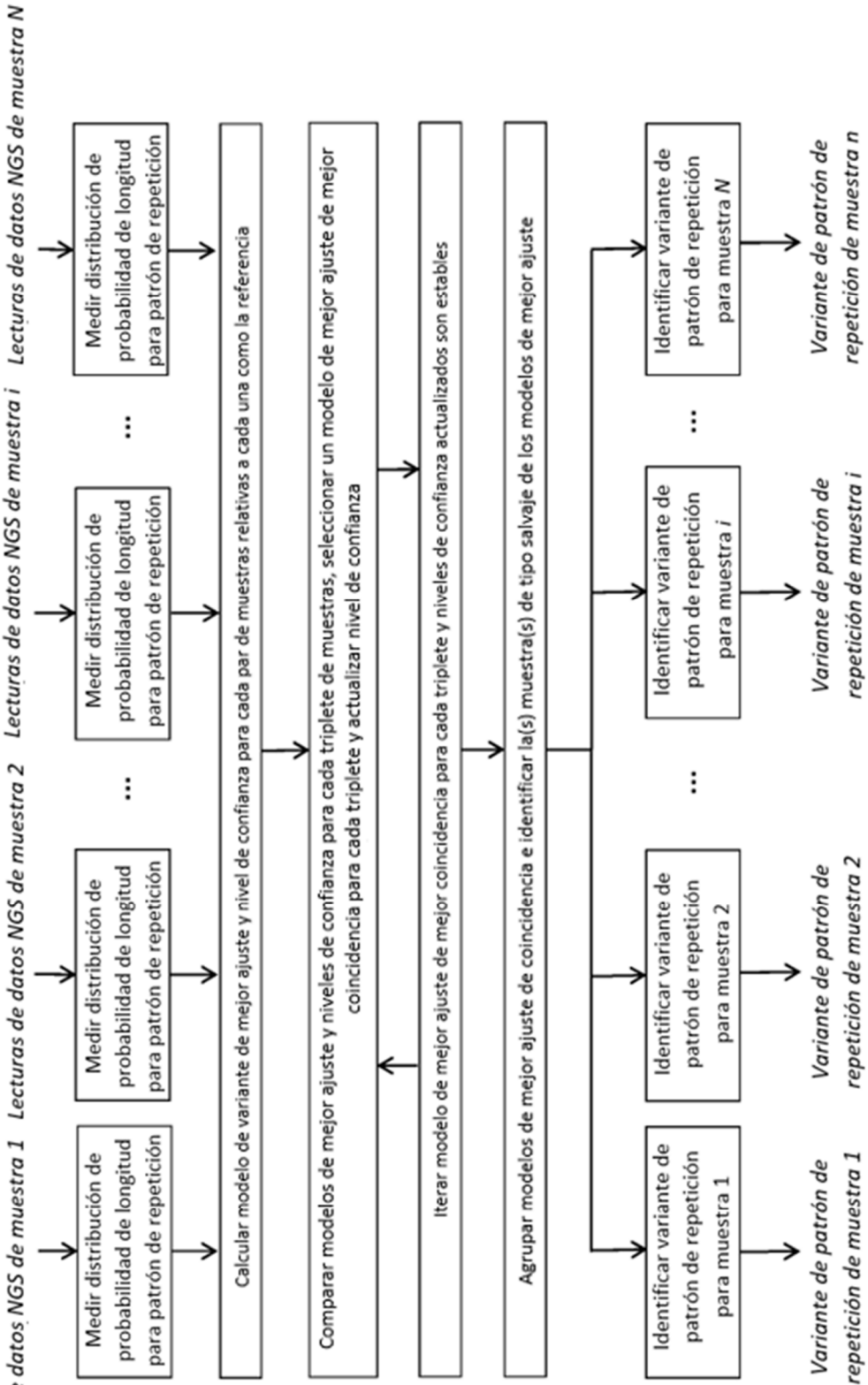

Lecturas de datos NGS de muestra 1
Lecturas de datos NGS de muestra 2
...
Lecturas de datos NGS de muestra i
...
Lecturas de datos NGS de muestra N
Medir distribución de probabilidad de longitud para patrón de repetición
Medir distribución de probabilidad de longitud para patrón de repetición
Medir distribución de probabilidad de longitud para patrón de repetición
Medir distribución de probabilidad de longitud para patrón de repetición
Calcular modelo de variante de mejor ajuste y nivel de confianza para cada par de muestras relativas a cada una como la referencia
Comparar modelos de mejor ajuste y niveles de confianza para cada triplete de muestras, seleccionar un modelo de mejor ajuste de mejor coincidencia para cada triplete y actualizar nivel de confianza
Iterar modelo de mejor ajuste de mejor coincidencia para cada triplete y niveles de confianza actualizados son estables
Agrupar modelos de mejor ajuste de coincidencia e identificar la(s) muestra(s) de tipo salvaje de los modelos de mejor ajuste
Identificar variante de patrón de repetición para muestra 1
Identificar variante de patrón de repetición para muestra 2
...
Identificar variante de patrón de repetición para muestra i
...
Identificar variante de patrón de repetición para muestra N
Variante de patrón de repetición de muestra 1
Variante de patrón de repetición de muestra 2
Variante de patrón de repetición de muestra i
Variante de patrón de repetición de muestra n

FIG. 9

	S2	S3	S4	S5	S6	S7	S8
S1	0/0 0.46	0/0 0.60	0/0 0.67	0/0 0.64	0/0 0.67	0/0 0.91	0/0 0.69
S2		0/0 0.32	-1/0 0.08	-1/0 0.10	NA/NA NA	0/0 0.18	-1/0 0.10
S3			0/0 0.12	0/0 0.29	NA/NA NA	0/0 0.38	0/0 0.32
S4				0/0 0.76	0/0 0.92	0/0 0.74	0/0 0.79
S5					0/0 0.64	0/0 0.68	0/0 0.96
S6						0/0 0.75	0/0 0.72
S7							0/0 0.68

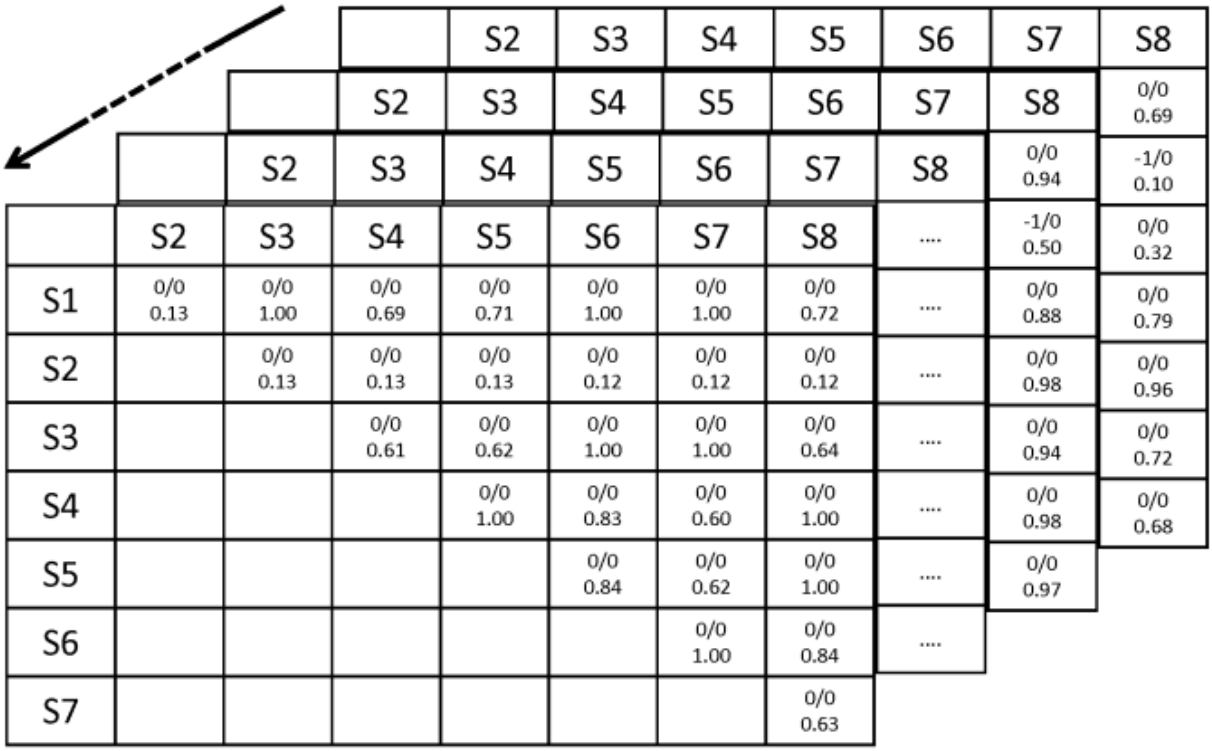
FIG.10

	S2	S3	S4	S5	S6	S7	S8
S1	0/0 0.46	0/0 0.60	0/0 0.67	0/0 0.64	0/0 0.67	0/0 0.91	0/0 0.69
S2		0/0 0.32	-1/0 0.08	-1/0 0.10	NA/NA NA	0/0 0.18	-1/0 0.10
S3			0/0 0.12	0/0 0.29	NA/NA NA	0/0 0.38	0/0 0.32
S4				0/0 0.76	0/0 0.92	0/0 0.74	0/0 0.79
S5					0/0 0.64	0/0 0.68	0/0 0.96
S6						0/0 0.75	0/0 0.72
S7							0/0 0.68

FIG.11

	S2	S3	S4	S5	S6	S7	S8
S1	0/0 0.46	0/0 0.60	0/0 0.67	0/0 0.64	0/0 0.67	0/0 0.91	0/0 0.69
S2		0/0 0.32	-1/0 0.08	-1/0 0.10	NA/NA NA	0/0 0.18	-1/0 0.10
S3			0/0 0.12	0/0 0.29	NA/NA NA	0/0 0.38	0/0 0.32
S4				0/0 0.76	0/0 0.92	0/0 0.74	0/0 0.79
S5					0/0 0.64	0/0 0.68	0/0 0.96
S6						0/0 0.75	0/0 0.72
S7							0/0 0.68

FIG.12



			S2	S3	S4	S5	S6	S7	S8
		S2	S3	S4	S5	S6	S7	S8	0/0 0.69
	S2	S3	S4	S5	S6	S7	S8	0/0 0.94	-1/0 0.10
S1	0/0 0.13	0/0 1.00	0/0 0.69	0/0 0.71	0/0 1.00	0/0 1.00	0/0 0.72	....	-1/0 0.50	0/0 0.32
S2		0/0 0.13	0/0 0.13	0/0 0.13	0/0 0.12	0/0 0.12	0/0 0.12	....	0/0 0.88	0/0 0.79
S3			0/0 0.61	0/0 0.62	0/0 1.00	0/0 1.00	0/0 0.64	....	0/0 0.98	0/0 0.96
S4				0/0 1.00	0/0 0.83	0/0 0.60	0/0 1.00	....	0/0 0.94	0/0 0.72
S5					0/0 0.84	0/0 0.62	0/0 1.00	....	0/0 0.98	0/0 0.68
S6						0/0 1.00	0/0 0.84	....	0/0 0.97	
S7							0/0 0.63			

FIG.13

	S2	S3	S4	S5	S6	S7	S8
S1	-1/0	0/0	0/0	0/0	0/0	0/0	0/0
S2		0/1	0/1	0/1	0/1	0/1	0/1
S3			0/0	0/0	0/0	0/0	0/0
S4				0/0	0/0	0/0	0/0
S5					0/0	0/0	0/0
S6						0/0	0/0
S7							0/0

Grupo 1: {S1, S3, S4, S5, S6, S7, S8}
Grupo 2: {S2}

FIG.14

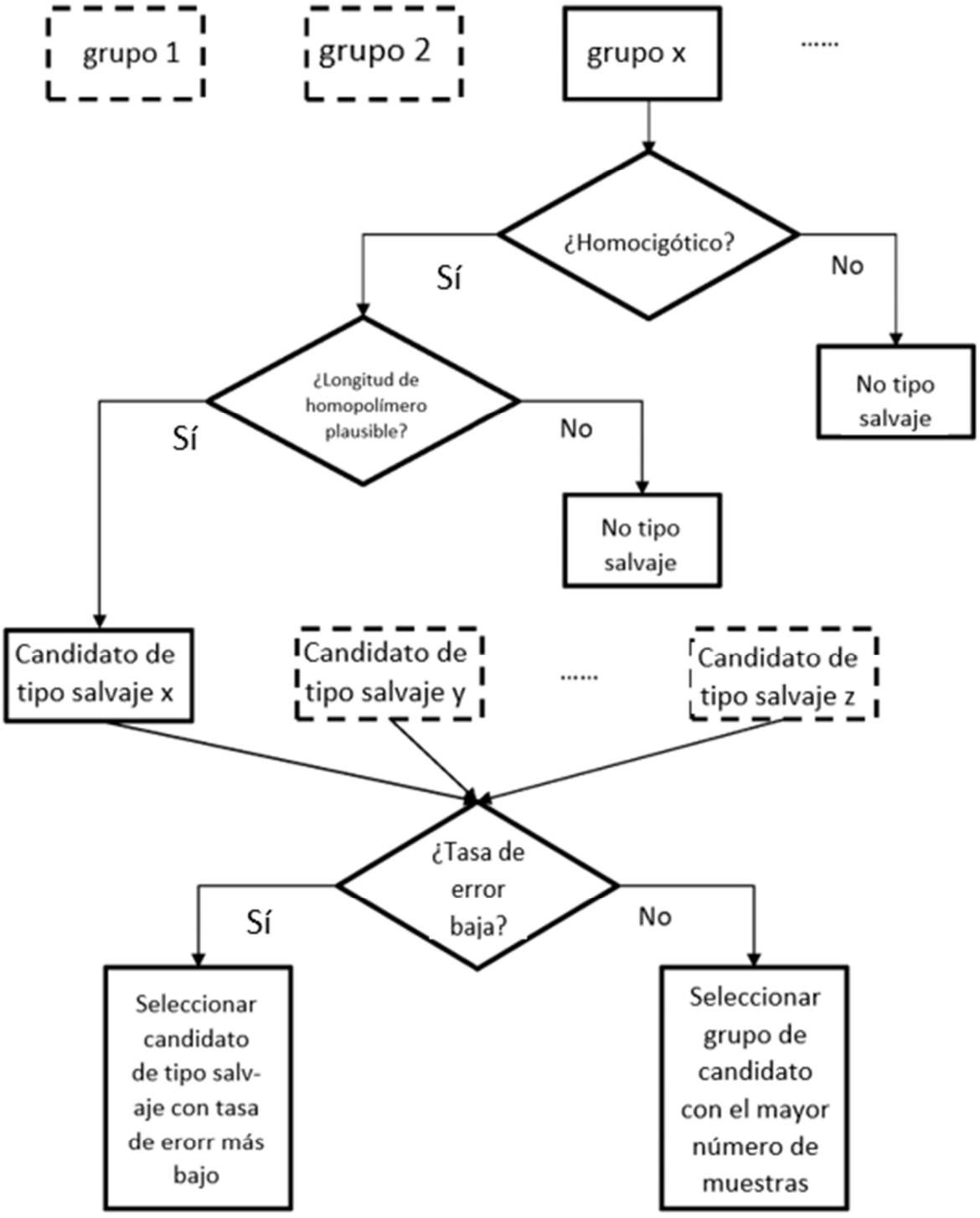

grupo 1
grupo 2
grupo x
.....
¿Homocigótico?
Sí
No
No tipo salvaje
¿Longitud de homopolímero plausible?
Sí
No
No tipo salvaje
Candidato de tipo salvaje x
Candidato de tipo salvaje y
.....
Candidato de tipo salvaje z
¿Tasa de error baja?
Sí
No
Seleccionar candidato de tipo salvaje con tasa de erorr más bajo
Seleccionar grupo de candidato con el mayor número de muestras

FIG.15

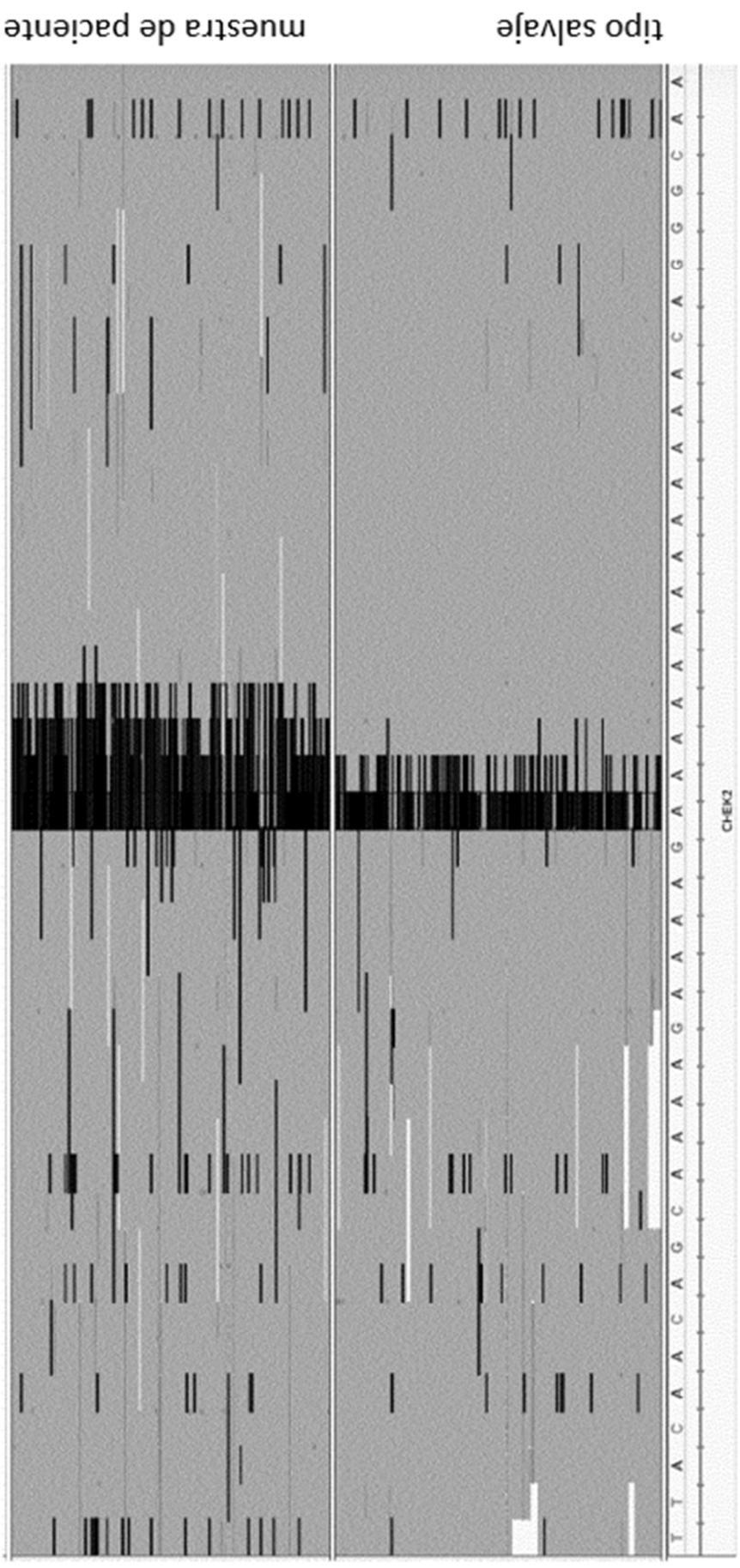


FIG.16

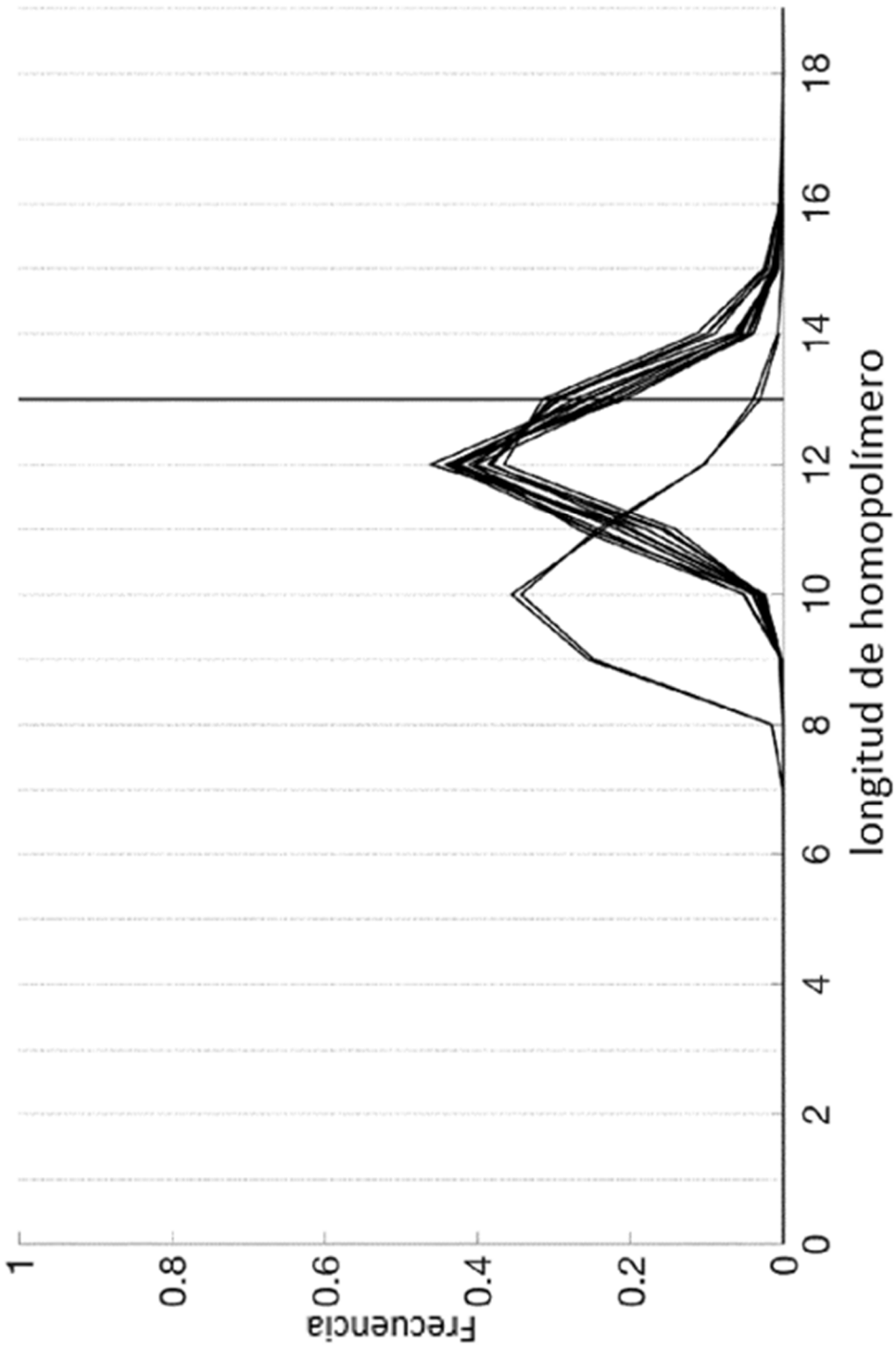

Frecuencia
1
0.8
0.6
0.4
0.2
0
0
2
4
6
8
10
12
14
16
18
longitud de homopolímero

FIG.17

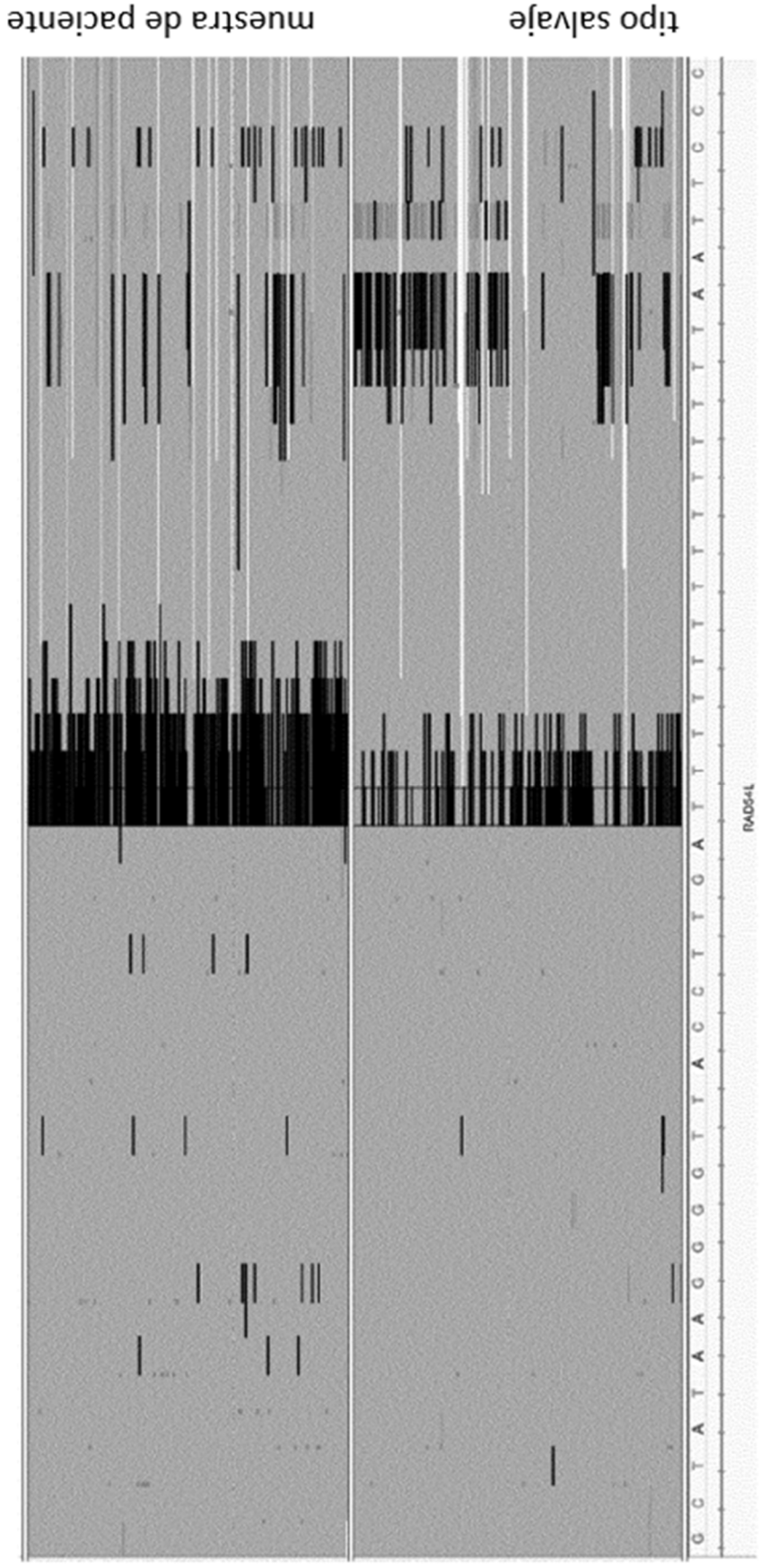

tipo salvaje
muestra de paciente
GCTAATAAGGGGCTTACCTTGATTTTTTTTTTTTTTAATTCCC
RAD641

FIG.18

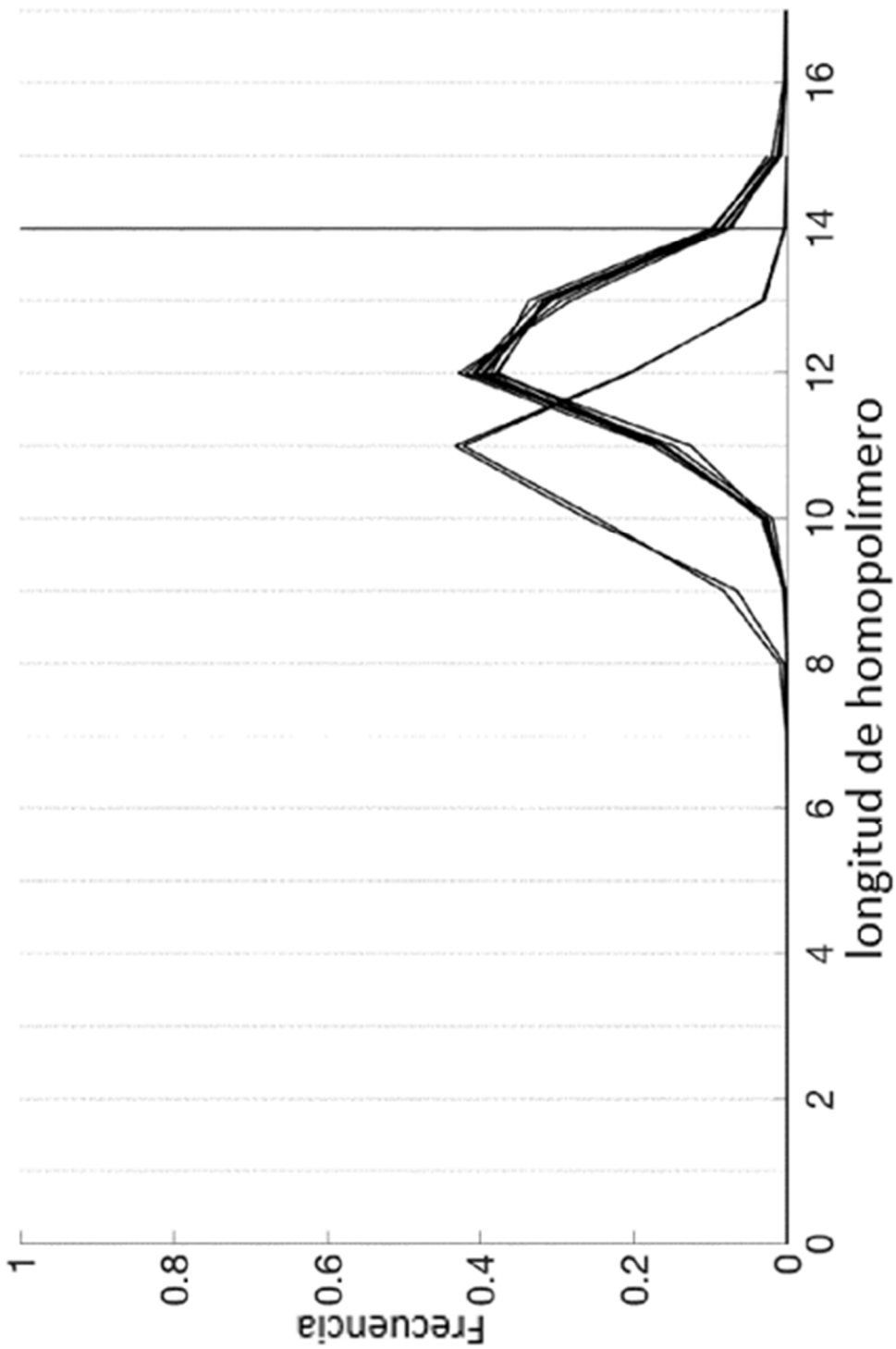

Frecuencia
1
0.8
0.6
0.4
0.2
0
0
2
4
6
8
10
12
14
16
longitud de homopolímero

FIG.19

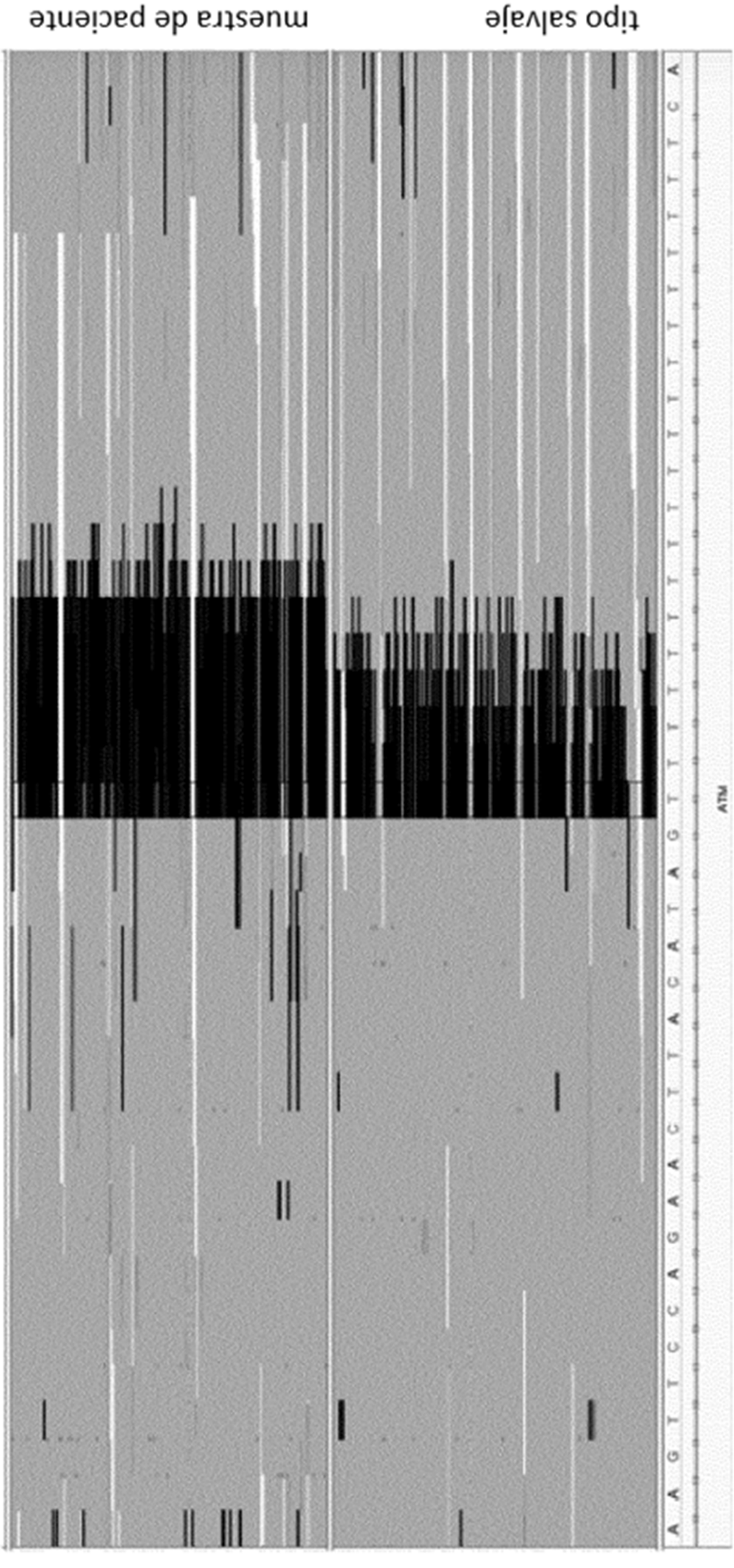


FIG.20

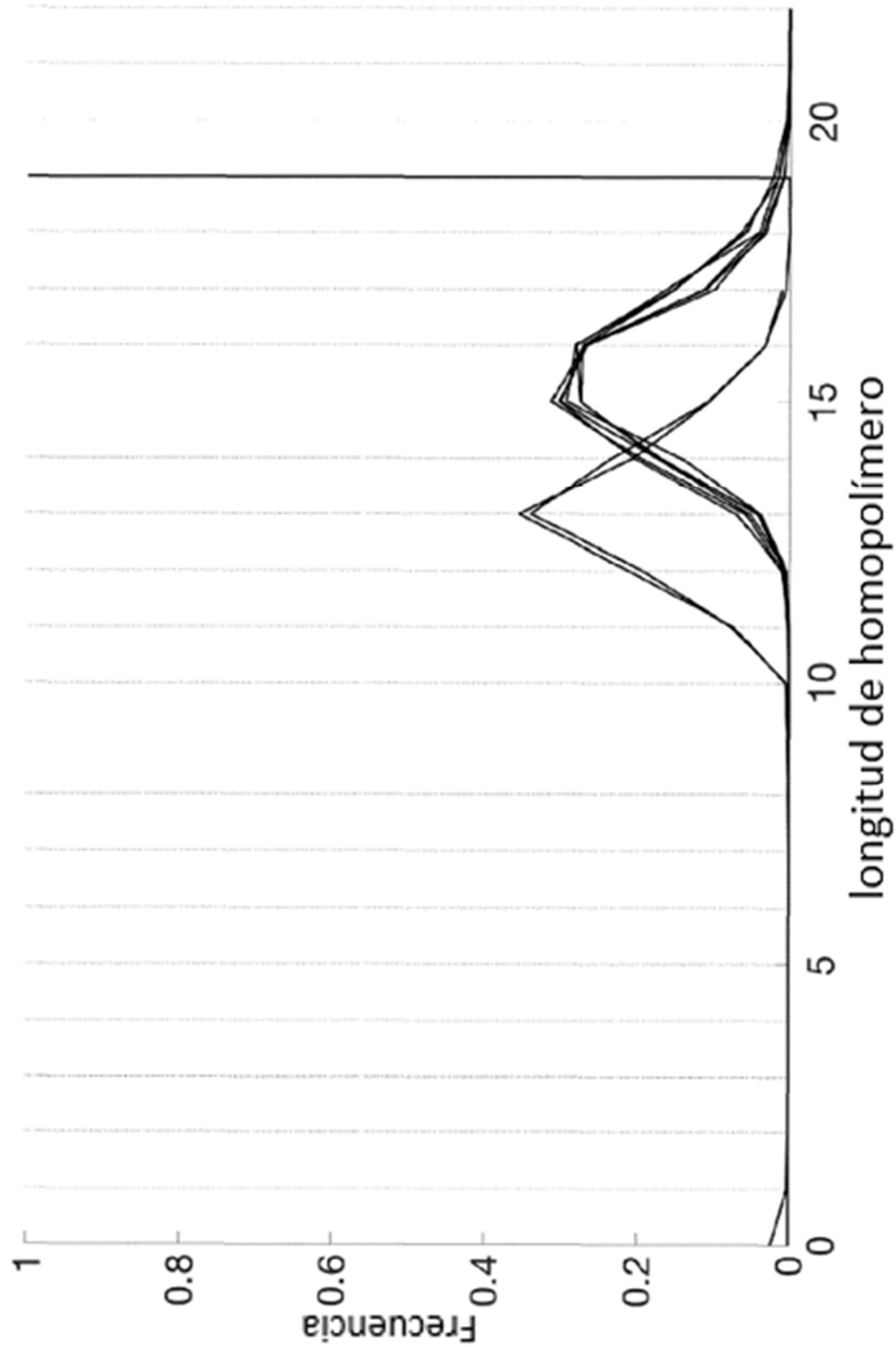

Frecuencia
1
0.8
0.6
0.4
0.2
0
0
5
10
15
20
longitud de homopolímero

FIG.21

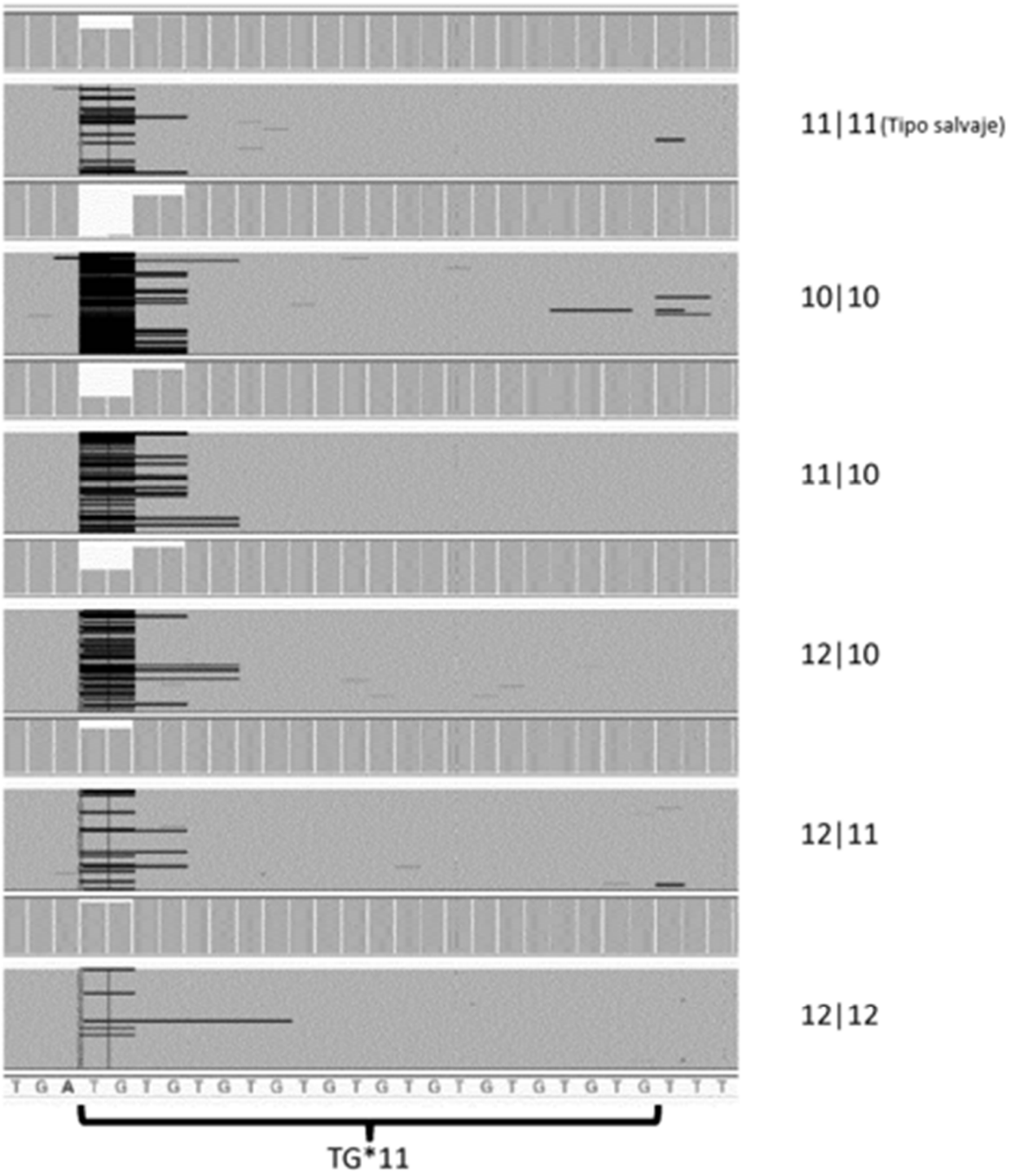

11|11 (Tipo salvaje)
10|10
11|10
12|10
12|11
12|12
T G A T G T G T G T G T G T G T G T G T G T G T T T
TG*11

FIG.22

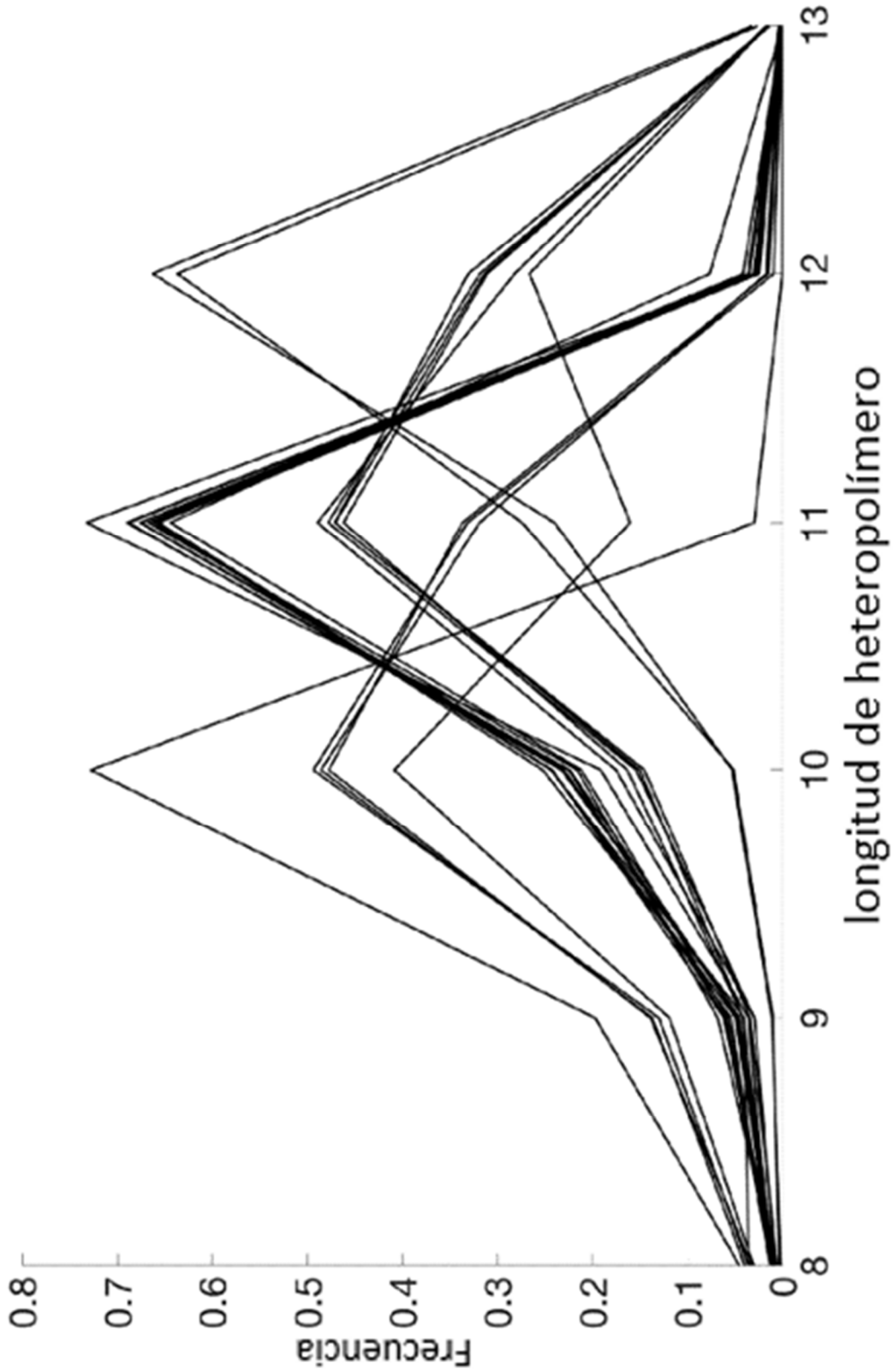

0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Frecuencia
8
9
10
11
12
13
longitud de heteropolímero

FIG.23