(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 711 163**

(21) Número de solicitud: 201731242

(51) Int. CI.:

G16B 20/50 (2009.01)

(12) SOLICITUD DE PATENTE A1

(22) Fecha de presentación:

23.10.2017

(43) Fecha de publicación de la solicitud:

30.04.2019

(71) Solicitantes:

HEALTH IN CODE, S.L. (100.0%)
Edf. El Fortin, Hospital Marítimo de Oza, s/n
15006 AS XUBIAS (A Coruña) ES

(72) Inventor/es:

DE UÑA IGLESIAS, David

(74) Agente/Representante:

CARPINTERO LÓPEZ, Mario

(54) Título: **Sistema y método de detección de variantes genéticas estructurales.**

(57) Resumen:

Sistema y método de detección de variantes genéticas estructurales.

Método, sistema y programa de ordenador para la detección de variantes estructurales, que permite su aplicación en distintos escenarios incluida la secuenciación dirigida empleando métodos de captura (incluyendo exomas o paneles). La invención puede implementarse mediante hilos paralelos y facilita un buen rendimiento independiente del tamaño de la región de interés o la cantidad de resultados. El método comprende caracterizar (230) cada muestra determinando un género en función de una cobertura cromosómica diferencial; calcular una covariabilidad experimental a través de una o más matrices de correlaciones; seleccionar (406) al menos una estructura de control mediante clusterización (404) iterativa; establecer (213) unos puntos de trabajo en función de unas variaciones respecto a las estructuras de control; y detectar (240) las variantes genéticas estructurales en los puntos de trabajo determinados.

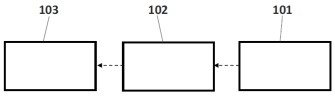

FIG. 1A

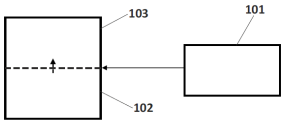

FIG. 1B

ES 2 711 163 A1

DESCRIPCIÓN

Sistema y método de detección de variantes genéticas estructurales.

Objeto de la invención

La presente invención se refiere al sector sanitario y biotecnológico, y más concretamente, a un método y sistema para la detección, evaluación y exploración de variantes genéticas estructurales.

Antecedentes de la invención

La información genética está codificada como una sucesión de nucleótidos. El genoma contiene genes que codifican proteínas específicas y regula distintas funciones del organismo. La información genética es específica de cada organismo, de modo que las distintas variaciones condicionan las diferentes condiciones fisiológicas del mismo. Múltiples enfermedades están asociadas a variantes genéticas con efectos deletéreos o mutaciones. Si bien muchas variantes consisten en alteraciones de unos pocos nucleótidos otras involucran rangos que pueden ir de decenas a miles de estos, ya sea en forma de inversión de secuencias, reubicación en posiciones diferentes, supresión o duplicación de las mismas. A las variantes que involucran un número elevado de posiciones se las denomina variantes estructurales y en particular, cuando suponen un aumento o disminución del número de copias de una determinada secuencia, se denominan variantes de número de copias (CNV, del inglés '*copy number variants*'). Puesto que dichas variantes estructurales se han descrito frecuentemente asociadas a consecuencias adversas para la salud, resulta crítico desarrollar técnicas para su detección y análisis.

Desde la aparición del método de secuenciación Sanger en 1977, los científicos tuvieron la capacidad de secuenciar ácidos nucleicos de manera reproducible y fiable. Una década después Applied Biosystems introdujo el AB370, primer instrumento automático de secuenciación basado en la electroforesis capilar que se convirtió en la principal herramienta para completar el proyecto del genoma Humano, la obtención de una secuencia genética consenso para el homo Sapiens. Desde esta primera etapa en el siglo XXI han emergido nuevas tecnologías de secuenciación (NGS, del inglés '*next generation sequencing*') que

han alcanzado la potencia necesaria para secuenciar rápidamente cantidades inmensas de secuencias genéticas a bajo coste realizando para ello millones de reacciones de secuenciación en paralelo al mismo tiempo. Debido al elevado volumen de información generado por los métodos NGS, actualmente la mayor complejidad se ha desplazado desde los procesos de secuenciación a los procesos de análisis mediante computadoras.

Los métodos NGS actualmente afianzados en el mercado, incluyendo la tecnología líder, Illumina, requieren de una preparación de muestras que comprende la amplificación (obtención de secuencias clónicas a las de la muestra original) y fragmentación, para así disponer de una colección de secuencias, relativamente cortas, que cubren las secuencias originales y que se secuenciarán en paralelo en el instrumento NGS. Una vez resueltos los nucleótidos de cada fragmento (o de una parte de los mismos), se dispone de un conjunto de lecturas (también denominadas habitualmente por su nombre en inglés '*reads*') que proveen la información de la secuencia de un tramo del material genético de la muestra de partida. En el caso de que la secuenciación provea pares lecturas que resuelvan cada uno de los extremos de un fragmento (pudiendo quedar la parte central sin secuenciar) se habla de secuenciación de pares de extremos (también denominada habitualmente por su nombre en inglés '*pair-end*'). Un nucleótido en la posición *p* de la secuencia de referencia habrá sido secuenciado por todas aquellas lecturas que cubran dicha posición *p*. Se dice entonces que *p* tiene una cobertura X igual al número de veces que ha sido leída o equivalentemente el número de lecturas que la cubren.

La secuenciación puede dirigirse a toda la muestra o limitarse a determinado material genético de la misma. La selección del material genético puede realizarse mediante la amplificación selectiva de regiones genéticas de interés ("amplicones") o bien recuperando fragmentos del material de partida que contienen secuencias específicas ("técnicas de captura"). Tanto la secuenciación de exomas (WES, del inglés *'Whole Exome Sequencing'*) como los paneles son aplicaciones de la secuenciación selectiva. La secuenciación de genomas (WGS, del inglés *'Whole Genome Sequencing'* es un caso de secuenciación no selectiva.

Dada la prevalencia de las técnicas NGS en el ámbito industrial y clínico, a la par de la importancia que cobra la determinación de variantes estructurales, en los últimos años se

han desarrollado distintos métodos para dicha determinación a partir de las lecturas producto de la secuenciación NGS. Los distintos enfoques pueden resumirse en cuatro categorías principales, incluyéndose asimismo combinaciones de las mismas. Todas ellas requieren un proceso de alineamiento de los *reads* contra una secuencia consenso conocida, asociada al material genético original, con el fin de conocer qué parte de la secuencia original está cubriendo cada *read* (p.e. el consenso puede ser la secuencia de referencia del genoma humano). Los métodos de la primera categoría (denominados habitualmente por su nombre en inglés ´read-pair´) analizan la orientación y distancia entre las dos secuencias laterales de un fragmento (trabajan sobre secuenciación ´pair-end´). La segunda categoría agrupa métodos (denominados habitualmente por su nombre en inglés ´split-read´) que examinan lecturas que contienen secuencias que no se dan en la secuencia de referencia directamente sino que son combinaciones de secuencias distantes. El tercer grupo de métodos (denominados habitualmente por su nombre en inglés ´read count´) lo componen aquellos que comparan el número de *reads* que cubren determinada región de la secuencia consenso en relación a otras áreas en condiciones análogas (de la misma muestra o de otras muestras control) para detectar regiones diferenciales asociadas a una variación relativa en la cantidad del material de partida (p.ej. existencia de deleciones que incluyan dichas secuencias) o variaciones en la propia secuencia de la muestra en dichas posiciones haciendo imposible encontrar una correspondencia con la referencia. Finalmente otros métodos tratan de reconstruir a partir de las lecturas la secuencia completa (no fragmentada) de partida para posteriormente compararla con la secuencia de referencia detectando así variaciones estructurales.

El enfoque tradicional para identificar variantes estructurales emplea técnicas citogenéticas como cariotipado o fluorescencia hibridación in situ (FISH, del inglés "*fluorescence in situ hibridation"*). En 2003 aparecen los arrays de comparación de hibridación genómica (CGH, del inglés "*comparative genomic hybridation*") y las metodologías de arrays de polimorfismo de nucleótido-único (SNPs arrays, del inglés "*single-nucleotide polymorphism arrays*"). Con el asentamiento de la tecnología NGS las distintas estrategias de detección han permitido superar algunas limitaciones. Se parte de datos orientados a un uso más general y se alcanza una mayor resolución, precisión y capacidad para detectar nuevas variantes estructurales. A pesar de los avances, los resultados obtenidos con los distintos métodos siguen variando y resultando poco fiables, por lo que no existe un estándar universal, o una solución satisfactoria o conjunto de herramientas de uso generalizado.

La secuenciación dirigida se ha asentado como manera coste-efectiva para interrogar regiones de interés de un conjunto grande de muestras, especialmente en la práctica clínica para el estudio de paneles de genes relacionados con enfermedades bajo evaluación. Aunque la secuenciación de paneles de genes hoy en día aporta la mayoría de los estudios realizados, mayormente en la práctica clínica, prácticamente la totalidad de las propuestas de análisis de variantes estructurales para NGS no soportan este tipo de diseño para el análisis. Asimismo, el reducido tamaño, la dispersión y naturaleza discontinua de las regiones de interés para la secuenciación dirigida impone muchas dificultades para aplicar los métodos de detección de variantes estructurales sobre los datos.

Los métodos basados en *pair-end* o *Split-read* son incapaces de detectar variantes estructurales en el contexto de secuenciación dirigida salvo que las regiones limítrofes a los extremos de las variantes estructurales estén cubiertas por zonas de interés, lo cual resulta poco habitual. La discontinuidad de las regiones de interés también hace que los métodos basados en ensamblado no funcionen. En este contexto, los enfoques basados en *read-count* o coberturas, o una combinación con otros, son los más adecuados hasta la fecha. Aun así, incluso los métodos basados en *read-count* tienen dificultades cuando no han sido diseñados específicamente para contemplar esta realidad discontinua, ya que aquellos que se han implementado para secuenciación de genoma completo asumen condiciones como la distribución normal de la cobertura o la continuidad del espacio de búsqueda que no son ciertas en secuenciación dirigida.

Los pocos métodos existentes aplicables al contexto de la secuenciación dirigida no tienen una alta precisión, necesitando en la práctica que uno o varios exones se vean afectados por completo. Tanto la tasa de falsos positivos como negativos es elevada y sin posibilidad de interpretación de los resultados, por lo que no es posible el control en profundidad de los especialistas para la mejora de los resultados. Algunas de las características que contribuyen a esta realidad son:

- La selección de modelos matemáticos para las señales de los datos poco ajustados resultado del desconocimiento de las variables implicadas en las señales respuesta analizadas (sea *read-count*, cobertura u otras) y fruto también de la simplificación de la información en parte al no considerar información relativa la preparación y el contexto biológico.

- Las metodología para la selección de controles de cara a modelar las señales, están orientadas a datos de secuenciación de genoma completo o exoma pero no de paneles. Incluso en las condiciones ideales, los métodos de selección y los modelos no contemplan las variables necesarias, como por ejemplo los efectos de la fragmentación de las muestras o las características de las regiones biológicas.
- Las regiones con alta homología, pseudogenes y el efecto de variantes genéticas especialmente en dichas regiones genera asignaciones incorrectas sistemáticas de *reads* a regiones cromosómicas, que se traducen en alteraciones en la señal en forma de ganancias y pérdidas que son detectadas como variantes estructurales. La falta de control de estas circunstancias en los modelos para su posible revisión hace muy difícil identificar estas situaciones. Además estas alteraciones introducen sesgos en los modelos matemáticos y no son controladas adecuadamente en la selección de controles para la construcción de los modelos.
- Las propuestas resultan muy sensibles a los efectos provocados por la divergencia de las muestras y la variabilidad experimental algo especialmente problemático en la secuenciación de paneles. Dichos efectos son considerados simplemente como ruido no explicable.
- Los intentos de corrección de bias que afectan a la distribución de la señal de análisis a lo largo de las regiones cromosómicas no han conseguido eliminar los sesgos.
- Las estrategias de búsqueda y detección de CNVs son muy dependientes de ciertos parámetros fijados experimentalmente con un conjunto de datos que no resulta reproducible en la práctica, frecuentemente son fijados directamente por el usuario, sucede esto por ejemplo con la elección de ventanas de cálculo o bins.
- Cuando existe una valoración de la credibilidad o calidad con que se ha detectado una variante estructural se valora en función del modelo matemático, sin realizar test adicionales considerando información complementaria y sin incorporar la información necesaria sobre el contexto en el que se ha encontrado la variante de cara a una segunda valoración en profundidad del especialista.
- Las transformaciones que se realizan sobre los datos o señales de cara a su modelización y presentación cuando está disponible algún tipo de gráfico hacen que se pierda la traza de relación señal-característica biológica, ocultando la causalidad.

Las propuestas actuales además de tener las limitaciones que se han presentado entre otras, no se integran bien con la práctica clínica y como resultado:

- Procuran una alta especificidad sacrificando la sensibilidad, pues no contemplan el proceso desde una perspectiva general, donde se debe primar en una primera fase la sensibilidad, pues la herramienta constituye un soporte cuyos resultados serán evaluados y confirmados posteriormente por los expertos ya que se cuenta con mucha información suplementaria, tanto de tipo biológico como clínico. Los límites de detección son problemáticos y hay un gran riesgo de perder eventos.
- Existen restricciones importantes en la elección de controles, por ejemplo en cuanto a la inclusión de familiares o la detección de CNVs polimórficos.
- No hay la posibilidad de explorar los resultados contemplando información individualizada de cada caso-muestra, se pierde la trazabilidad. No se proporcionan filtros aplicables de manera individual a posteriori o tratamientos específicos según las circunstancias del estudio.

En definitiva, sigue existiendo en el estado de la técnica la necesidad de una herramienta de detección de variantes genéticas fiable y capaz de soportar secuenciación de paneles, que facilite asimismo la gestión y explotación para el análisis integral de los resultados obtenidos.

Descripción de la invención

La presente invención soluciona todos los problemas mencionados mediante una técnica de determinación de variantes genéticas de perspectiva global, aplicable a diversos escenarios entre los que se encuentra (de manera no limitativa) el de la genética médica y la secuenciación por paneles.

En un primer aspecto de la invención, se presenta un método de detección de variantes genéticas estructurales a partir de datos de secuenciación, preferentemente procedentes de sistemas NGS, pero que en determinadas realizaciones pueden ser obtenidos mediante cualquier otra tecnología de secuenciación conocida en el estado de la técnica. El método comprende los siguientes pasos para cada muestra:

- Determinar un género de la muestra en función de, al menos, una cobertura cromosómica diferencial del cromosoma X. Preferentemente, dicha determinación de género se realiza mediante comparación de cobertura en el cromosoma X y los

cromosomas autosómicos, si bien pueden implementarse otros métodos de determinación de género conocidos en el estado de la técnica. Esta estrategia permite realizar una determinación de género robusta ante muestras de baja calidad o mal secuenciadas.

- Calcular, a partir de los datos de secuenciación, al menos una matriz de correlaciones que representa la covariabilidad experimental de la pluralidad de muestras. Preferentemente, se calculan dos matrices de correlaciones. La primera matriz de correlaciones está preferentemente basada en perfil de cobertura, pero puede no obstante basarse en un rango de posiciones, un número de lecturas mapeadas, una métrica derivada de una región cromosómica, o cualquier combinación de las anteriores. La segunda matriz de correlaciones está preferentemente basada en la distribución del tamaño de fragmentos secuenciados, con o sin incluir adaptadores y otros fragmentos de longitud conocida, pero puede no obstante basarse en separación de lecturas en lecturas ~~pareja final~~ de parejas de extremos, o cualquier otra métrica de tamaños conocida en el estado de la técnica.
- Seleccionar al menos una estructura de control mediante clusterización iterativa. La división en clústeres se realiza preferentemente de manera iterativa hasta alcanzar un resultado convergente, y más preferentemente, se basa en la aplicación de un algoritmo kmeans, si bien otras realizaciones de la invención pueden comprender otros algoritmos de clusterización conocidos en el estado de la técnica.
- Establecer unos puntos de trabajo (o "working points") en función de la variabilidad y los valores de referencia de la estructura de control (es decir, del sustituto de modelo) y de las muestras de estudio. Preferentemente la determinación de puntos de trabajo se hace sobre posiciones con cobertura superando un umbral, sin que esto excluya la posibilidad de trabajar con rangos en lugar de posiciones o métricas alternativas a la cobertura como un número de reads mapeados. También preferentemente se estudian dichas características considerando el contexto local, es decir, la información de las posiciones próximas. Dichos puntos de trabajo representan posiciones donde el modelo captura adecuadamente los efectos de las potenciales variantes estructurales, eliminando posibles sesgos. Preferentemente, los puntos de trabajo se establecen en función de las variaciones respecto a un valor de referencia. Más preferentemente, dicho valor de referencia es una media, mediana u otra medidas de tendencia central, calculada sobre la cobertura, contaje de lecturas (o "reads"), o cualquier otra métrica derivada de dichas variables. También preferentemente, la variación se calcula mediante el rango intercuartílico, la

varianza, la desviación típica u otras medidas de variación respecto a dicho valor de referencia.

- Preferentemente, normalizar los datos en función del género determinado para cada muestra, eliminando así sesgos de enriquecimiento. Preferentemente, los datos se normalizan empleando la cobertura total en las regiones autosómicas, si bien puede emplearse uno o más rangos determinados. Alternativamente, los datos pueden normalizarse mediante un valor de lecturas alineadas.
- Detectar las variantes genéticas estructurales en los puntos de trabajo establecidos, en función de las desviaciones respecto a las estructuras de control en los puntos de trabajo determinados. Preferentemente, este paso comprende además asignar un valor indicativo de un grado confianza a cada variante estructural detectada, en función de, al menos, la desviación respecto a la estructura de control y la región cromosómica donde se sitúa variantes genéticas estructurales. También preferentemente, este paso comprende agrupar múltiples variantes estructurales detectadas en una única, en función de, al menos, la desviación respecto a la estructura de control y la región cromosómica donde se sitúa variantes genéticas estructurales.

- También preferentemente, cribar los resultados en función al menos de un factor como los ratios (número de copias) característicos de las variantes estructurales a detectar, información sobre zonas codificantes o paneles u otras anotaciones sobre las regiones afectadas por los rangos alterados detectados. También preferentemente, los resultados se pueden cribar l en función de la información proporcionada por la estructura del alineamiento de la muestra de estudio, con la posibilidad de analizar eventos asociados a técnicas de “split-read”, tales como la búsqueda de “puntos de ruptura” (habitualmente citados por el correspondiente término inglés “breakpoints.

Preferentemente, el método comprende almacenar los resultados obtenidos siguiendo una codificación que permite acceso aleatorio y que comprende, al menos, los siguientes campos:

- Una cabecera (1812) con una sección de tamaño invariable (1807) y una sección de tamaño variable (1808). La sección de tamaño invariable (1807) comprende

información del tamaño de datos, mientras que la sección de tamaño variable (1808) comprende información de señalización de metadatos;

- Un cuerpo (1813) que comprende bloques de igual tamaño con información de coordenada cromosómica, señal de referencia (1803) reescalada y señales reescaladas (1804, 1805, 1806) asociadas a cada muestra.
- Una cola (1814) que comprende información de localización.

También preferentemente, el método comprende acceder a los resultados mediante un acceso aleatorio, facilitado por la codificación descrita, y que permite la visualización fluida de los resultados independientemente del tamaño total de archivo. En particular el método de acceso comprende:

- Acceder a la cabecera.
- Obtener metadatos de los resultados.
- Obtener información asociada a cada punto a representar accediendo al cuerpo (1813) y recuperando bloques de información de tamaño constante.

En un segundo aspecto de la invención se presenta un sistema de detección de variantes genéticas estructurales, que comprende unos medios de detección que implementan los pasos de cualquier implementación del método del primer aspecto de la invención. Los medios de detección almacenan los resultados generados (es decir, las variantes genéticas detectadas, información vinculada a la mismas, pudiendo incorporar información sobre las estructuras de control así como metainformación sobre los propios datos incluidos) en unos medios de almacenamiento de datos, a los que a su vez acceden unos medios de exploración. Dichos medios de exploración sirven de interfaz con el usuario o con otros sistemas, obteniendo los datos que se necesitan visualizar o transmitir en cada momento. Dependiendo de la implementación particular del sistema, los medios de detección, medios de almacenamiento y medios de exploración pueden estar integrados en un mismo dispositivo o estar implementados en múltiples dispositivos conectados por cualquier conexión alámbrica o inalámbrica conocida en el estado de la técnica. En una opción de implementación, los medios de almacenamiento y medios de exploración pueden estar integrados en un mismo archivo informático independiente generado por los medios de detección, utilizable desde un sistema externo, ya sea local o remoto.

Finalmente, en un tercer aspecto de la invención se presenta un programa de ordenador que comprende medios de código de programa de ordenador adaptados para implementar el método descrito, al ejecutarse en un procesador digital de la señal, un circuito integrado específico de la aplicación, un microprocesador, un microcontrolador o cualquier otra forma de hardware programable. Nótese que cualquier opción preferente e implementación particular del dispositivo y sistema de la invención puede ser aplicado al método y al programa de ordenador de la invención, y viceversa.

El método, sistema y programa de ordenador de la invención permiten por lo tanto detectar variantes genéticas de manera fiable y eficiente, compatible con secuenciación de paneles, y facilitando asimismo la gestión y explotación para el análisis integral de los resultados obtenidos. La organización de los datos permite asimismo independizar la carga computacional del tamaño de los mismos, agilizando dicho cómputo y permitiendo su visualización mediante acceso aleatorio a los resultados a tiempo constante.

Descripción de las figuras

Con objeto de ayudar a una mejor comprensión de las características de la invención de acuerdo con un ejemplo preferente de realización práctica de la misma, y para complementar esta descripción, se acompañan como parte integrante de la misma las siguientes figuras, cuyo carácter es ilustrativo y no limitativo:

Las figura 1A y 1B muestran esquemáticamente los elementos principales de sendas realizaciones preferentes del sistema de la invención

La figura 2 presenta un diagrama de flujo de los pasos realizados por el subsistema de detección, de acuerdo con una realización preferente del método de la presente invención.

La figura 3 es un diagrama de flujo del proceso de asignación de sexo, de acuerdo con una realización preferente del método de la presente invención.

La figura 4 ejemplifica un diagrama de flujo del proceso de asignación de conjuntos de muestras de control a cada muestra de estudio, de acuerdo con una realización preferente del método de la presente invención.

La figura 5 ilustra un diagrama de flujo del cálculo de la matriz de correlación en base al perfil de coberturas, de acuerdo con una realización preferente del método de la presente invención.

La figura 6 presenta un diagrama de flujo del procedimiento de división en clústeres de las muestras y la asignación de muestras control para cada muestra de estudio a partir de los datos de correlación en base a cobertura y fragmentación, de acuerdo con una realización preferente del método de la presente invención.

La figura 7 es un diagrama de flujo del proceso de construcción del modelo de referencia vinculado a cada conjunto de muestras de control, para un conjunto de regiones de interés asignadas a un bloque de trabajo, de acuerdo con una realización preferente del método de la presente invención.

La figura 8 ejemplifica un diagrama de flujo para establecer si una posición de una región de interés pertenece a los puntos de trabajo para el modelo asociado a un conjunto de referencia, de acuerdo con una realización preferente del método de la presente invención.

La figura 9 ilustra un diagrama de flujo del estudio del comportamiento de una señal de cobertura frente al modelo de referencia que le corresponde (algoritmo de búsqueda), de acuerdo con una realización preferente del método de la presente invención.

La figura 10 presenta un diagrama de flujo de procesado de un rango outlier, de acuerdo con una realización preferente del método de la presente invención.

La figura 11 es un diagrama de flujo de ajuste de los límites del outlier, de acuerdo con una realización preferente del método de la presente invención.

La figura 12 ejemplifica un diagrama de flujo del proceso de caracterización de un outlier, de acuerdo con una realización preferente del método de la presente invención.

La figura 13 ilustra un diagrama de flujo para ampliación del límite inferior del intervalo, de acuerdo con una realización preferente del método de la presente invención.

La figura 14 presenta un diagrama de flujo para determinar si un outlier puede reflejar un una variante estructural, de acuerdo con una realización preferente del método de la presente invención.

La figura 15 es un diagrama de flujo de detección y valoración de puntos de ruptura (“*breakpoints”*) compatibles con una variante estructural, de acuerdo con una realización preferente del método de la presente invención.

La figura 16 ejemplifica un diagrama de flujo de cálculo y registro de una puntuación que refleja el grado de confianza en que efectivamente el comportamiento de las señales para el CNV candidato refleja una variante estructural subyacente, de acuerdo con una realización preferente del método de la presente invención.

La figura 17 ilustra un diagrama de flujo de fusión de CNVs, de acuerdo con una realización preferente del método de la presente invención.

La figura 18 presenta una realización preferente de la codificación con la que los medios de exploración almacenan la información en los medios de almacenamiento de datos.

La figura 19 ilustra un diagrama de flujo de generación de flujo de datos, de acuerdo con una realización preferente del método de la presente invención.

Realización preferente de la invención

En este texto, el término "comprende" y sus derivaciones (como "comprendiendo", etc.) no deben entenderse en un sentido excluyente, es decir, estos términos no deben interpretarse como excluyentes de la posibilidad de que lo que se describe y define pueda incluir más elementos, etapas, etc. Asimismo, las descripciones de funciones y elementos conocidos en el estado del arte pueden haber sido omitidos por claridad y concisión

Nótese que las realizaciones preferentes de la invención han sido descritas para el caso de información extraída mediante técnicas NGS, pero puede ser aplicada de manera general a cualquier otra técnica de análisis genético conocida en el estado de la técnica. Asimismo, las realizaciones preferentes han sido descritas utilizando nombres específicos de ficheros y variables para facilitar la comprensión de la invención, pero que en ningún caso limitan su alcance tal y como ha sido reivindicado, pudiendo implementarse la invención con cualquier otra organización o nomenclatura de datos, ficheros y/o bases de datos que permita implementar el proceso descrito. De la misma manera, el experto en la materia podrá entender que pueden introducirse modificaciones en el orden y/o distribución de los pasos descritos, así como en las funciones matemáticas particulares implementadas, dentro de dicho alcance tal y como ha sido reivindicado.

La figura 1A presenta esquemáticamente los elementos de una primera realización preferente de la invención, que comprende un subsistema de detección (101) y anotación (también denominado medios de detección), un subsistema contenedor de datos (102) (también denominado medios de almacenamiento de datos), y un subsistema de exploración (103) (también denominado medios de exploración). El subsistema de detección (100) detecta las variantes genéticas estructurales y codifica la información resultante de manera estructurada en el subsistema contenedor de datos (102). A su vez, el subsistema de

exploración (103) actúa de interfaz con el usuario, recibiendo sus comandos y mostrando la información correspondiente almacenada en el subsistema contenedor de datos (102).

Si bien el formato de la información almacenada en el subsistema contenedor de datos (102) puede variar entre implementaciones, el flujo de bytes para todo el intervalo de exploración asociado a cada CNV candidato comprende preferentemente los siguientes metadatos: las posiciones incluidas según la resolución de exportación configurada en el subsistema de detección y anotación, así como la señal de referencia y de cada muestra para dichas posiciones. Para un determinado intervalo de exploración, el flujo de bytes está formado por una sucesión de bloques de igual tamaño que depende del número de muestras. Cada bloque contiene información asociada a una posición cromosómica, preferentemente indicando en sus dígitos superiores de la coordenada cromosómica (la parte entera del resultado de dividir dicha coordenada entre 10000), los 5 dígitos menos significativos de dicha coordenada (el resto de la división de la coordenada entre 10000), la señal de referencia según el modelo para dicha posición reescalada a la señal de la muestra de estudio para dicha posición, y las señales reescaladas asociadas a dicha posición para cada una de las muestras consideradas en el estudio.

Todos los metadatos de un flujo de datos asociados a un CNV candidato se proveen directamente al subsistema de exploración (102), sin necesidad de examinar todos los datos del contenedor. Una vez localizado un flujo de datos, el sistema de exploración (102) solicita bloques de datos asociados a coordenadas concretas, los cuales se proveen mediante un acceso aleatorio, es decir sin necesidad de acceder al resto de bloques del flujo.

Nótese que los bloques principales del sistema pueden implementarse siguiendo diversas configuraciones alternativas, independientes de la técnica de detección implementada en el subsistema de detección (101). Por ejemplo, la figura 1B presenta un ejemplo en el que el subsistema de detección (101) se configura para generar un resultado que integra el subsistema contenedor de datos (102) y el subsistema de exploración (103) en un único fichero. De la misma manera, el subsistema de detección (101), el subsistema contenedor de datos (102) y el subsistema de exploración (103) pueden estar integrados en un mismo equipo, estar implementados en diversos equipos conectados mediante cualquier subsistema de comunicaciones alámbricas o inalámbricas conocidas en el estado de la técnica, o comprender cualquier medio adicional de procesado, almacenaje de datos,

interacción con el usuario, etc conocidos de manera general en el estado de la técnica. Por ejemplo, el subsistema de exploración (103) puede comprender un visor propio, o bien medios de comunicación que suministran una interfaz gráfica a un explorador web. De acuerdo con otras alternativas, el subsistema de exploración (103) puede ser una aplicación cliente de un servidor de la información del subsistema contenedor de datos (102); o bien el subsistema de exploración (103) y el subsistema contenedor de datos (102) pueden estar integrados en único resultado autónomo, que actúa como proveedor de datos para el subsistema de exploración (103) embebido en el propio resultado, de manera que el usuario puede explorar los resultados fuera de línea con un navegador.

La figura 2 presenta esquemáticamente los pasos realizados por el subsistema de detección (101), de acuerdo con una realización preferente del método de la invención. Una vez inicializado (200), los pasos pueden agruparse en leer (210) los parámetros de configuración para el proceso de detección, procesar (220) la información e inicializar el proceso, caracterizar (230) las muestras y asignar modelos de referencia, detectar y caracterizar (240) CNVs y generar (250) resultados. El paso de leer (210) los parámetros de configuración engloba la configuración e inicialización del proceso propiamente dicho y la estructuración del trabajo a realizar, comprendiendo a su vez los siguientes pasos: cargar (211) los parámetros de procesado (que denominaremos configuración del análisis), determinar (212) qué muestras se van a analizar y dónde localizar sus datos NGS para el análisis, establecer (213) los intervalos cromosómicos sobre los que se desea realizar el proceso de detección de CNVs (que denominaremos región de interés o ROI, del inglés “region of interest”) y obtener (214) las anotaciones vinculadas a dicha región de interés así como información contextual.

La información de configuración del sistema puede proveerse mediante un fichero de configuración, sin que esto excluya hacerlo mediante una interfaz de usuario para tal fin con campos para fijar los parámetros deseados. El listado de muestras, así como la localización de los datos NGS vinculados, puede suministrarse de igual manera, también la región de interés o las anotaciones. Como caso particular no excluyente, y debido a su popularidad, la región de interés puede especificarse en formato .bed, al igual que las distintas anotaciones cromosómicas (también se da soporte entre otros a ficheros .gff u otros formatos como el empleado por el USCS, Universidad California Santa Cruz debido a su popularidad en el ámbito de la bioinformática).

Los datos NGS básicos para el análisis, relacionados con las muestras, consisten típicamente en un fichero de alineamientos (denominado *bam*) por muestra. Dicho fichero es un estándar que puede obtenerse alineando los *read*s generados durante la secuenciación empleando diversos programas, estando el proceso de obtención de los *bam* ampliamente descrito en el estado de la técnica. Una vez secuenciadas las muestras, basta con configurar la ejecución de un flujo de datos o cadena de ejecución (también denominado "*pipeline*") previo al proceso de detección de CNVs para contar con dichos datos.

La región de interés suministrada al sistema de exploración y anotación consta de uno o más intervalos cromosómicos, para los que se especifica: cromosoma, posición cromosómica de inicio del intervalo, posición cromosómica de fin del intervalo y dos etiquetas. La primera etiqueta da nombre a una serie de regiones relacionadas (por ejemplo puede ser el nombre de un gen), la segunda, al combinarse con la primera, identifica unívocamente a la región (por ejemplo el número de región contigua codificante asociada al gen). Otras implementaciones pueden comprender niveles de identificación alternativos y/o adicionales. Durante la carga de la región de interés, aquellos rangos que se solapan son fusionados en rangos nuevos que los comprenden, manteniendo un registro de las regiones originales. Un parámetro de configuración establece un número de pares de bases adicionales a los intervalos subministrados al subsistema y que se incorporan a la región de interés (se añaden por los extremos de los intervalos). Cuando tras dicha adicción los intervalos lleguen a ser contiguos o solaparse, se fusionan en uno nuevo, conservando también el registro de sus regiones originales. Los intervalos fusionados resultantes se ordenan según el cromosoma que ocupan y sus coordenadas iniciales. Al final del proceso se tiene una lista de intervalos $ROI = \{R_1 \ldots R_n\}$. Para cada elemento R_i de la lista de intervalos, se tiene: un nombre (identificador), un grupo o categoría, un cromosoma, posiciones cromosómicas inicial y final del intervalo que abarca, una lista de coordenadas iniciales de cada uno de los intervalos (no repetidos) iniciales fusionados y una lista de igual modo con las coordenadas finales correspondientes.

Las anotaciones están referidas a posiciones o rangos cromosómicos, indicando de manera equivalente a la región de interés: cromosoma, posiciones cromosómicas inicial y final, y tipo de anotación de que se trata. Dichas anotaciones incluyen las regiones exónicas (en coordenadas cromosómicas), tanto las regiones codificantes como las no traducidas o UTRs

(del inglés Untraslated Regions de los transcritos a considerar y que abarcan la región de interés.

A continuación, el paso de procesar (220) la información e inicializar el proceso comprende generar (221) un plan de ejecución y preparar la carga de trabajo de cara a su paralelización. El subsistema de detección (101) de CNVs divide el trabajo en bloques que se pueden ejecutar en paralelo. Se estudia el número de bases que abarca la región de interés y se divide en bloques que son asignados a múltiples ramas de trabajo, configuradas en forma de hilos, subprocesos, etc. Cada rama trabaja secuencialmente sobre los bloques que le son asignados, siendo cada bloque un conjunto de intervalos de ROI de manera que todos los intervalos pertenecientes a un mismo grupo (por ejemplo, al mismo cromosoma) están contenidos en un único bloque. La asignación de bloques a las ramas se realiza de manera que la diferencia en pares de bases (pb) entre los bloques es mínima. El tamaño de los bloques está controlado por un parámetro del sistema que define los pb por iteración y que denominaremos *Cfg.pbPerIteration*, de forma que intervalos de un nuevo gen no serán asignados a un bloque de trabajo si previamente dicho bloque ya rebasa el número de pares de bases establecido por dicho parámetro. Este modelo de trabajo optimiza los recursos disponibles manteniendo la máxima información de contexto (ej. genes contiguos) durante el procesado de los datos.

Previamente a la búsqueda de CNVs el subsistema de detección y anotación, se deben modelar las muestras de estudio. El paso de caracterizar (230) las muestras y asignar modelos de referencia comprende asignar (231) perfiles de coberturas, sexo y variantes; y asignar (232) conjuntos de referencia (muestras de control) para cada muestra. Este proceso conlleva la caracterización de las muestras, y la asignación de conjuntos de muestras control, para cada muestra de estudio a partir del cual se genera un modelo de control. Para cada muestra se registra: identificador de la muestra, placa de secuenciación, carril dentro la placa (conocido habitualmente por el término en inglés ´lane´) e índice asignado a la muestra, fichero de alineamiento asociado, fichero de variantes (por ejemplo, un fichero .vcf) asociado. Durante la caracterización de las muestras se registran también: el total de cobertura en las regiones de interés (en ROI), los cromosomas autosómicos, el cromosoma X y el cromosoma Y; el sexo asociado (hombre, mujer); el listado de variantes vinculadas indicando: cromosoma, posición cromosómica, frecuencia de alelo alternativo y calidad. La obtención de coberturas puede realizarse de acuerdo con lo conocido de manera

general en el estado de la técnica, como por ejemplo a partir de informes de cobertura fruto de un pipeline de análisis ejecutado previamente (ej. pipeline de detección de variantes no estructurales), mediante la ejecución de un programa externo, o empleando un código para requerir dicha información a partir de los ficheros *bam*.

Tras caracterizar (230) las muestras y asignar modelos de referencia, el método comprende ejecutar (241) el proceso de detección y anotación, y generar (251) los resultados correspondientes, antes de su finalización (260).

La figura 3 muestra con mayor detalle el proceso de asignación de sexo implementado en el paso de caracterizar (230) las muestras y asignar modelos de referencia. La asignación de sexo se basa en la proporción diferenciada que presentan hombres y mujeres en cuanto a cobertura en el cromosoma X frente a los cromosomas autosómicos. Los pasos de dicho protocolo tras su inicio son los siguientes:

1) Calcular (302) las coberturas totales en la región de interés ROI para cada una de las muestras, en el cromosoma X y en los cromosomas autosómicos

$$RX = \{ r \in ROI / crom(r)=X \} , RA = \{ r \in ROI / crom(r) \notin \{X,Y\} \}$$

$$Cx = \{ Cx_1 \dots Cx_n / Cx_i = \sum_{r \in RX} \sum_{p=pini(r)}^{pfin(r)} cob(p, M_i) \}$$

$$Ca = \{ Ca_1 \dots Ca_n / Ca_i = \sum_{r \in RA} \sum_{p=pini(r)}^{pfin(r)} cob(p, M_i) \}$$

donde:

crom(r) es una función que devuelve el cromosoma sobre el que se localiza una región de interés r.

RX representa el conjunto de regiones de interés localizadas sobre el cromosoma X.

RA representa el conjunto de regiones de interés localizadas sobre los cromosomas autosómicos (es decir todos excluyendo el cromosoma X y el cromosoma Y).

M representa el conjunto de muestras de estudio. M_i a la muestra “i” en estudio.

cob(p,m) es una función que devuelve la cobertura en la posición *p* para la muestra *m*

pini(r) y *pfin(r)* son funciones que dada una región r perteneciente al conjunto de regiones de interés ROI devuelven respectivamente la posición de inicio de dicha región y la posición final.

C_x es el conjunto de coberturas totales en el cromosoma X para cada una de las muestras. Cada elemento de conjunto (Cx_i) es la cobertura total en el cromosoma X para la muestra "i" en estudio calculada como la suma de las coberturas en cada posición abarcada por cada una de las regiones de interés localizadas sobre el cromosoma X para la muestra "i".

Ca es el conjunto de coberturas totales en los cromosomas autosómicos (todos con excepción del cromosoma X y el cromosoma Y) para cada una de las muestras. Cada elemento de conjunto (Ca_i) es la cobertura total en los cromosomas autosómicos para la muestra "i" en estudio calculada como la suma de las coberturas en cada posición abarcada por cada una de las regiones de interés localizadas sobre algún cromosoma autosómico para la muestra "i".

2) Calcular (303) para cada muestra los ratios de cobertura en X frente a los cromosomas autosómicos y ordenarlos

$$Cr = \{ Cr_1 \dots Cr_n / Cr_i < Cr_{i+1} \ \wedge Cr_j \in \bigcup_{j=1}^{n} \frac{Cx_j}{Ca_j} \}$$

donde:

Cx_j es la cobertura total en el cromosoma X para la muestra "j".

Ca_j es la cobertura total en los cromosomas autosómicos para la muestra "j".

Cr_j es el ratio para la muestra "j" entre la cobertura total en el cromosoma X y en los cromosomas autosómicos para dicha muestra.

Cr es el conjunto de ratios de coberturas totales asociados a las distintas muestras de estudios ordenados de menor a mayor.

3) Tomar (304) una pareja de elementos de Cr que no haya sido explorada previamente, y tomar (305) el valor inferior como representante del ratio asociado a hombres y el superior como representante del asociado a mujeres. Si es la primera pareja que se toma, inicializar valores de representación previos a -1.

$$P = \{ (Cr_i, Cr_j) \;/\; j > i \wedge Cr_i \in Cr \wedge Cr_j \in Cr \}$$

$$r_H = Cr_i\,, \qquad r_M = Cr_j\,, \qquad r_H^{'} = r_M^{'} = -1$$

donde:

P es el conjunto de posibles parejas, formado por todos los posibles posibles pares de ratios de coberturas totales distintos (sin importar el orden de los dos términos) de las distintas muestras. Cr_j es el ratio para la muestra "j" entre la cobertura total en el cromosoma X y en los cromosomas autosómicos para dicha muestra.

r_H es una variable que representa el ratio de referencia para hombres, que será recalculada en las distintas iteraciones del método. r'_H es el valor que tomaba r_H en la iteración previa a la actual (o -1 en el caso de la primera iteración).

r_M es una variable que representa el ratio de referencia para mujeres, que será recalculada en las distintas iteraciones del método. r'_M es el valor que tomaba r_M en la iteración previa a la actual (o -1 en el caso de la primera iteración).

4) Dividir (306) los ratios en pertenecientes a hombres o mujeres. Si la distancia de un ratio al valor de referencia para hombres es menor que la distancia respecto del de referencia para mujeres tomarlo como hombre, en otro caso como mujer.

$$H = \{Cr_i \;/\; |Cr_i - r_H| < |Cr_i - r_M|\}$$

$$M = \{Cr_i \;/\; |Cr_i - r_H| \geq |Cr_i - r_M|\}$$

donde:

Cr_i es el ratio para la muestra “i” entre la cobertura total en el cromosoma X y en los cromosomas autosómicos para dicha muestra.

r_H es una variable que representa el ratio de referencia para hombres.

r_M es una variable que representa el ratio de referencia para mujeres.

H es el conjunto de ratios de cobertura más próximos a $r_{H.}$

M es el conjunto de ratios de cobertura más próximos a r_M o a la misma distancia que r_H.

5) Eliminar (307) los ratios atípicos asignados a los grupos. En el clúster de hombres serán atípicos aquellos valores para los cuales, al dividir el valor mínimo en el grupo de mujeres entre dichos valores, el resultado es superior a 3 unidades. En el clúster de mujeres serán atípicos, aquellos valores que, al dividirlos por el valor máximo en el grupo de hombres, sean superiores a 3 unidades.

$$H_{ok} = \{\, Cr_i /\ \ Cr_i \in H \ \wedge\ \frac{\min(M)}{Cr_i} < 3 \,\}$$

$$M_{ok} = \{\, Cr_i /\ \ Cr_i \in M \ \wedge\ \frac{Cr_i}{\max(H)} < 3 \,\}$$

donde:

Cr_i es el ratio para la muestra “i” entre la cobertura total en el cromosoma X y en los cromosomas autosómicos para dicha muestra.

min(X),max(X) son funciones que devuelven el elemento con menor valor y mayor valor de un conjunto X respectivamente.

H_{ok} es el conjunto de ratios pertenecientes a H tal que la distancia entre el menor valor en M y ellos es inferior a 3 veces el valor de dichos ratios. Los elementos de H_{ok} representan valores válidos en H (quedando excluidos valores atípicos).

M_{ok} es el conjunto de ratios pertenecientes a M tal que la distancia entre el mayor valor en H y ellos es inferior a 3 veces el valor dicho valor máximo en H. Los elementos de M_{ok} representan valores válidos en M (quedando excluidos valores atípicos).

6) Si alguno de los conjuntos, tras sacar los valores atípicos, es vacío y quedan parejas de valores por explorar volver al paso 3, tomando una nueva pareja. Si no quedan parejas por explorar, asignar (308) sexo desconocido a cada muestra concluyendo (309) el protocolo.

7) En el caso de que los conjuntos de valores para hombres y mujeres sin valores atípicos no son vacíos, recalcular (310) los valores representantes para cada grupo como el promedio de dichos conjuntos. Asimismo, los valores actuales pasan a tomarse como valores previos.

$$r_H^{'} = r_H\ ,\ \ r_M^{'} = r_M\ ,\quad r_H = \overline{H_{ok}}\ \ ,r_M = \overline{M_{ok}}$$

donde:

r_H es una variable que representa el ratio de referencia para hombres, que será recalculada en las distintas iteraciones del método. r'_H es el valor que tomaba r_H en la iteración previa a la actual (o -1 en el caso de la primera iteración).

r_M es una variable que representa el ratio de referencia para mujeres, que será recalculada en las distintas iteraciones del método. r'_M es el valor que tomaba r_M en la iteración previa a la actual (o -1 en el caso de la primera iteración).

$\overline{H_{ok}}\ y\ \overline{M_{ok}}$ representan el valor promedio de los elementos pertenecientes al conjunto H_{ok} y al conjunto M_{ok} respectivamente.

8) Si los valores recalculados coinciden con los previos:

   a. Si el cociente entre el ratio representativo de mujeres y el de hombres está dentro del intervalo [1.8, 2.2], dividir (311) las muestras en base a la proximidad de los ratios asociados a los valores de referencia para hombres y mujeres.

$$H = \{Cr_i\ /\ \ |Cr_i - r_H|\ \ <\ \ \ |Cr_i - r_M|\}$$

$$M = \{Cr_i\ /\ \ |Cr_i - r_H|\ \ \geq\ \ \ |Cr_i - r_M|\}$$

donde:

Cr_i es el ratio para la muestra “i” entre la cobertura total en el cromosoma X y en los cromosomas autosómicos para dicha muestra.

r_H es una variable que representa el ratio de referencia para hombres.

r_M es una variable que representa el ratio de referencia para mujeres.

H es el conjunto de ratios de cobertura más próximos a r_H considerados ratios pertenecientes a muestras de sexo masculino.

M es el conjunto de ratios de cobertura más próximos a r_M *o a la misma distancia que* r_H, considerados ratios pertenecientes a muestras de sexo femenino.

b. Si el cociente no está dentro del intervalo [1.8, 2.2]: si quedan parejas volver al paso 3 tomando (304) una nueva pareja; en caso contrario, asignar (308) sexo desconocido a cada muestra concluyendo (309) el protocolo.

Nótese que el sexo registrado es en función del número de copias del cromosoma X, y permite corregir el desbalance natural para dicho cromosoma entre ambos sexos a la hora de comparar las señales vinculadas a las muestras en dichos cromosomas. Una muestra sin señal en el cromosoma X se tomará como Hombre y una muestra con 2 o más copias del cromosoma X se tomará como mujer (independientemente de que el sexo fenotípico sea de hombre –ej. síndrome de Klinefelter –XXY-).

La figura 4 muestra el proceso de asignación de conjuntos de muestras de control a cada muestra de estudio, realizado mediante clústering iterativo sobre las muestras de análisis (hasta alcanzar una condición de parada para determinar los grupos finales. El procedimiento comienza (401) tomando (402) las regiones de interés normalizadas *ROI*, ordenadas por cromosoma y posición inicial. Se calcula (403) una primera matriz de correlaciones (MRC) para el conjunto de muestras analizadas en base a su perfil de cobertura y se calcula (404) una segunda matriz de correlaciones (MRT) para el conjunto de muestras analizadas en base a la distribución del tamaño de los fragmentos alineados asociados a cada muestra. Para cada muestra bajo estudio (M_E) se dividen (405) en clústeres las muestras en base al par de valores de correlación de cobertura y fragmentación ($R = \{R_i \ / \ R_i = (RC_{i\text{-}E}, RT_{i\text{-}E})\}$, donde R es el conjunto de pares de valores

posibles tal que cada par representa la correlación según la matriz MRC y MRT de la muestra i con la muestra de estudio ($_E$)). Nótese que la muestra bajo análisis consigo misma es el par (1,1). Realizando el proceso de clustering iterativo que se describe más adelante se selecciona (406) un conjunto de muestras de control, que denominaremos M.set. Cuando el proceso se ha realizado iterativamente para todas las muestras bajo estudio se finaliza (407) este proceso.

La figura 5 muestra con mayor detalle el paso de calcular (403) la matriz de correlación en base al perfil de coberturas. El procedimiento comienza (501) cargando (502) las regiones de interés normalizadas, el número de sitios de muestreo de cobertura deseados para calcular las correlaciones en base a al perfil de cobertura y las muestras de estudio. A continuación se calcula (503) el tamaño en pares de bases de la región de interés (PBR) como:

$$\mathrm{PBR} = \sum_{i=1}^{n} (pfin(R_i) - pini(R_i) + 1).$$

donde:

PRB es el tamaño en pares de bases de la región e interés.

pini(r) y *pfin(r)* son funciones que dada una región r perteneciente al conjunto de regiones de interés ROI devuelven respectivamente la posición de inicio de dicha región y la posición final.

También se calcula un intervalo de muestreo (IM) igual a la parte entera del cociente entre PBR y el parámetro de configuración que indica el número de sitios de muestreo deseados (Cfg.MUESTREO) para realizar la correlación de las muestras en base a la cobertura IM := *foor*(PBR/Cfg.MUESTREO, donde la función floor devuelve la parte entera de un número). En caso de que el cociente sea inferior a uno, se tomará como intervalo de muestreo 1. Con esta información se construye la lista de posiciones cromosómicas para las que se calculan la cobertura en cada muestra y sobre las que se calculan la matriz de correlaciones. Para cada uno de los intervalos R_i contenidos en ROI se calcula (505) el número de sitios a muestrear (np), parte entera del cociente entre el tamaño del intervalo R_i en pares de bases y el intervalo de muestreo $np = floor(\frac{(pfin(R_i) - pini(R_i) + 1)}{IM})$. Si para R_i np es 0, se calcula (506) la posición cromosómica intermedia del intervalo R_i como $p = posinicial(R_i) + floor\ \frac{(pfinal(R_i) - pi + 1)}{2}$ y se añade a la lista de posiciones de muestreo las posiciones en el

intervalo [p-3,p+3] para el cromosoma vinculado a R_i. Si para R_i np es mayor que 0, se añade (507) a la lista las posiciones en el intervalo: [pini(R_i) + (np·i) – 3, pini(R_i) + (np·i) + 3] variando el valor para i en 1 desde 0 hasta np y para el cromosoma vinculado a R_i. Una vez obtenida la lista de posiciones cromosómicas a muestrear, se obtiene (508) la cobertura en dichas posiciones para cada una de las muestras de estudio.

Para cada una de las muestras m se calcula (509) el conjunto de tasas de variación para cada una de las posiciones de muestreo como D ={ $\frac{NC_{m,p}-\mathrm{med}(NCM_p)}{iqr(NCM_p)}$ / p $\in$ SM}, donde NCm,p es la cobertura obtenida para la muestra m en la posición p dividida entre el total de cobertura para dicha muestra en los cromosomas autosómicos, y NCMp es el conjunto de valores NCm,p para cada muestra m del conjunto total de muestras (M). Considerando δ = med(D) donde med representa la mediana, ε = iqr (D) donde iqr representa el rango intercuartílico, se calcula la tasa de correlación de cada muestra con las demás considerando las coberturas para aquellas posiciones en SM tales que Dm,p < δ + ε, conformándose de esta manera la matriz correlaciones.

Para al cálculo de la matriz de correlación en base al perfil de fragmentación de las muestras, se procesan los ficheros de alineamiento para cada muestra, registrando para cada una de ellas, el número de fragmentos alineados de un determinado tamaño. Dicho tamaño se corresponde con el dato que consta en los propios ficheros de alineamiento. Obtenidas las frecuencias por tamaño de fragmento, para cada muestra se calcula la matriz de correlación de dichas frecuencias para las muestras de análisis.

La figura 6 muestra en detalle el procedimiento de división en clústeres de las muestras y la asignación de muestras control para cada muestra de estudio a partir de los datos de correlación en base a cobertura y fragmentación. El proceso se inicia (601) tomando (602) el número mínimo de muestras de control para una muestra dada (Cfg.SMIN), *M* = { M_1 … M_n } como las muestras analizadas, M_E = M_i como la muestra de estudio, y *R* = {R_i / R_i = (RC_{i-E},RT_{i-E})} como los pares de correlaciones (RC,RT) de cada muestra con la de estudio. A continuación se calcula (603) la distancia euclídea de cada par R_i respecto al que corresponde a la muestra de estudio. Para la muestra i la distancia euclídea es entonces $\sqrt{(RC_{i-E}-1)^2+(RT_{i-E}-1)^2}$.

donde:

$RC_{i\text{-}E}$ es el coeficiente de correlación en cuanto al perfil de cobertura según los puntos de muestreo correspondientes entre la muestra i y la muestra de estudio ($_E$). En la implementación preferente el coeficiente de correlación es el coeficiente de correlación pearson.

$RT_{i\text{-}E}$ es el coeficiente de correlación en cuanto al perfil de fragmentación entre la muestra i y la muestra de estudio ($_E$). En la implementación preferente el coeficiente de correlación es el coeficiente de correlación pearson.

Empezando por dos clústeres (604), se aplica un algoritmo de agrupamiento, como por ejemplo kmeans, para agrupar (605) las muestras en base a los datos de correlación de fragmentación y cobertura. Se inicializan los centroides a los valores de correlación correspondientes a los k pares con menor distancia euclídea respecto del (1,1) sin incluir este par, y se aplica la misma estrategia incrementando en uno de cada vez el número de clústeres hasta que el par (1,1) correspondiente con la muestra objeto de estudio sea el único miembro del clúster al que ha sido asignado al hacer la partición en k clústeres. Se toma k-1 como el número de clústeres ideal. Si en la iteración con el número de clústeres ideal el clúster asignado a la muestra de estudio tiene un número de miembros superior al número de muestras de control mínimo configurado en el subsistema (por defecto dos muestras), se registran como muestras de control para la muestra de estudio las muestras cuyos correspondientes pares de correlación han sido asignados al mismo grupo que el de la muestra de control (sin incluir a la propia muestra de estudio como control). En caso de que el número de miembros en la iteración k-1 no supere el mínimo configurado, se reporta (606) una excepción que alerta esta situación y recupera la iteración con el número de clústeres más próximo a k-1 que contenga un número miembros superior al requerido para el clúster correspondiente a la muestra de estudio. El resto de muestras de dicho grupo se toma (607) como controles para la muestra de estudio y se finaliza el proceso (608). Nótese que otras técnicas de clusterización en bases as sus perfiles de cobertura y fragmentación pueden ser utilizadas alternativamente.

Tras la caracterización y modelado de las muestras, cada rama de ejecución, en paralelo, lleva a cabo secuencialmente el proceso de detección y anotación de CNVs para cada uno de los bloques de trabajo que le corresponde. Este proceso consta de dos fases:

a) Generación de modelos de referencia para la comparación de señales asociadas a las muestras.

b) Para cada muestra, estudio del comportamiento de la señal de cobertura asociada para detectar desviaciones respecto al modelo de referencia; registro de los intervalos afectados, caracterizándolos y seleccionando entre ellos aquellos candidatos a estar afectados por variantes estructurales; valoración de dicha afectación y recompilado de la información relevante para la generación de resultados.

La figura 7 muestra el proceso de construcción del modelo de referencia vinculado a cada conjunto de muestras de control S_i para un conjunto de regiones de interés R_1 … R_n asignadas a un bloque de trabajo. El proceso se inicia (701) cargando (702) las siguientes variables :

ROI = { R_1 , R_2, … , R_R } : Regiones de la iteración ordenadas por cromosoma y posición inicial.

M = { M_1 , M_2, … , M_M }: Conjunto de muestras de estudio.

B = { B_i / B_i = bam(M_i) }: Conjunto de alineamientos para las muestras, B_i es el alineamiento para la muestra i.

S = { S_1, S_2, … , S_S}: Conjunto de conjuntos de muestras de control (conjuntos de referencia) distintos asignados durante el procedimiento correspondiente ya descrito.

RC , RNC : Matrices de cobertura cruda y cobertura normalizada respectivamente (m x posiciones(R_i)) por región.

RRF, RUP, RDW, RVAR : Matrices asociadas a la estructura de control o modelo (s x posiciones(R_i.)), respectivamente: Cobertura normalizada de referencia para el modelo, límite superior para considerar una cobertura normal (no atípica) según el modelo, límite inferior para considerar una cobertura normal según el modelo, variabilidad asociada al modelo. s es el número de conjuntos de control generados.

r: región activa, inicializada a 1.

$_R$: número de regiones

A continuación se realizan los siguientes pasos:

1. Para cada muestra m:

a. Para cada posición cromomosómica *p* en el intervalo correspondiente a la primera región R_1, obtener (702) la cobertura para cada muestra de estudio consultando el alineamiento (fichero *bam*) correspondiente (B_m). Registrar dichas coberturas vinculadas a muestra, posición y región como coberturas crudas.

$$RC_r = C_{M \times pos(Rr)} \; / \; C_{m,p} = getCov\,(B_m, p)$$

donde:

getCov(Bm,p) es una función que devuelve la cobertura en la posición p de acuerdo al alineamiento Bm.

$RC_r = C_{M \times pos(Rr)}$ es una matriz con tantas filas como muestras de estudio y tantas columnas como posiciones tenga una región r donde cada valor $Cm_{,p}$ se calcula como *getCov(Bm,p)*.

b. Normalizar (703) las coberturas crudas. La cobertura normalizada para una posición y muestra de estudio se calcula dividiendo la cobertura cruda entre la cobertura total en las regiones autosómicas de interés para dicha muestra y adicionalmente entre 2 cuando la región a la que pertenece la posición se localiza en uno de los cromosomas sexuales y el sexo asociado a la muestra es mujer. Registrar dichas coberturas vinculadas a muestra, posición y región como coberturas normalizadas.

$$RNC_r = NC_{M \times pos(Rr)} \; / \; NC_{m,p} = \frac{C_{m,p}}{cobAutosomal(M_m) * \varphi}$$

$$\varphi = 2 \text{ si } sexo(M_m) = \text{Mujer} \wedge crom(R_r) \in \{X,Y\}$$

$$\varphi = 1 \text{ otro caso.}$$

donde:

$RNC_r = NC_{M\ x\ pos(Rr)}$ es una matriz con tantas filas como muestras de estudio y tantas columnas como posiciones tenga una región r cuyos valores son las correspondiente coberturas normalizadas.

$C_{m,p}$ representa la cobertura cruda en la posición p para la muestra m.

cobAutosomal(m) es una función que devuelve la cobertura total en los cromosomas autosómicos y la región de interés para la muestra m.

sexo(m) es una función que devuelve el sexo asociado a una muestra m.

crom(r) es una función que devuelve el cromosoma asociado a la función r.

φ es un factor de corrección del número de copias diferenciales para X en hombres y mujeres, toma valor dos en el caso de regiones sobre el cromosoma X cuando el género de la muestra bajo cálculo es mujer, en otro caso vale 1.

2. Para cada conjunto de referencia S_i generar (705) los datos asociados a su modelo:
   a. Vinculado a posición y región se calcula y registra:
      i. La señal de referencia (cobertura normalizada) para el modelo. La señal de referencia para una posición dada es la mediana de las coberturas normalizadas asociadas a dicha posición (una por cada muestra de estudio). No obstante, otras realizaciones podrían trabajar con otros parámetros estadísticos como media, media truncada, etc.

$$RRF_r = RF_{s\ x\ pos(Rr)} / RF_{s,p} = med(\{ NC_{m,p} / M_m \in S_s \})$$

donde:

$RRC_r = FR_{s\ x\ pos(Rr)}$ es una matriz con tantas filas como conjuntos de referencia y tantas columnas como posiciones tenga una región r, para una fila asociada al conjunto de referencia s, los valores de las columnas se corresponden con la mediana de las coberturas

normalizadas de las muestras pertenecientes a s para cada una de las posiciones de la región r.

med(X) representa la mediana de los valores del conjunto X.

ii. Límite superior de variación típica de la señal de referencia, calculado como el valor que determina el primer cuartil de los valores asociados a las coberturas normalizadas para una posición dada. No obstante, otras realizaciones podrían trabajar con otros parámetros estadísticos como desviación típica, varianza, etc.

$$RUP_r = UP_{s\ x\ pos(Rr)} \ / \ UP_{s,p} = Q1(\{ NC_{m,p} / M_m \in S_s \})$$

donde:

$RRC_r = UP_{s\ x\ pos(Rr)}$ es una matriz con tantas filas como conjuntos de referencia y tantas columnas como posiciones tenga una región r, para una fila asociada al conjunto de referencia s, los valores de las columnas se corresponden con la mediana de las coberturas normalizadas de las muestras pertenecientes a s para cada una de las posiciones de la región r.

Q1(X) representa el primer cuartil de los valores del conjunto X.

iii. Límite inferior de variación típica de la señal de referencia, calculado como el valor que determina el tercer cuartil de los valores asociados a las coberturas normalizadas para una posición dada.

$$RDW_r = DW_{s\ x\ pos(Rr)} \ / \ DW_{s,p} = Q3(\{ NC_{m,p} / M_m \in S_s \})$$

donde:

$RDW_r = DW_{s\ x\ pos(Rr)}$ es una matriz con tantas filas como conjuntos de referencia y tantas columnas como posiciones tenga una región r, para una fila asociada al conjunto de referencia s, los valores de las columnas se corresponden con la el tercer cuartil de las coberturas normalizadas de las muestras pertenecientes a s para cada una de las posiciones de la región r.

Q3(X) representa el tercer cuartil de los valores del conjunto X.

iv. Variación típica de la señal de referencia para una posición calculada como el rango intercuartílico de los valores asociados a las coberturas normalizadas para dicha posición.

$$RVR_r = VR_{s\ x\ pos(Rr)} \ / \ VR_{s,p} = UP_{s\ x\ pos(Rr)} - DW_{s\ x\ pos(Rr)}$$

donde:

$RVR_r = VR_{s\ x\ pos(Rr)}$ es una matriz con tantas filas como conjuntos de referencia y tantas columnas como posiciones tenga una región r, para una fila asociada al conjunto de referencia s, los valores de las columnas se corresponden con rango intercuartílico de las coberturas normalizadas de las muestras pertenecientes a s para cada una de las posiciones de la región r.

v. Si la posición pertenece a las posiciones de trabajo, se registran (706) puntos de trabajo (también citados como *working points*) vinculados al modelo, es decir las señales en dicha región cumplen los criterios necesarios para considerar su medida satisfactoria para el procedimiento de detección.

3. Repetir los pasos anteriores para cada una de las regiones incluidas en el bloque de trabajo, registrando los datos de cada modelo para cada posición de cada región. Cuando se finalizan dichas regiones (r=R), el proceso finaliza.

La figura 8 presenta el proceso para establecer si una posición *p* de una región de interés *R* pertenece a los puntos de trabajo para el modelo asociado a un conjunto de referencia S_i. La determinación de puntos de trabajo supone una estrategia de selección de posiciones de análisis de las señales, para controlar el efecto de la variabilidad experimental y ciertos artefactos que se producen por el diseño y la tecnología empleada. Una posición se considera un punto de trabajo si cumple ciertas condiciones, algunas de las cuales conllevan el estudio de su contexto cromosómico local. El rango de posiciones incluidas en el contexto cromosómico local está determinado por una ventana de posiciones prefijada (denominaremos a esta variable Cfg.WPwindow). El proceso se inicia (801) cargando (802) las siguientes variables:

$M = \{ M_1 , M_2, \dots , M_M \}$: Muestras de estudio.

$MS = \{M_i / M_i \in S\}$ controles para el modelo S.

$R=[p_1 \dots p_n]$, S : Región que abarca una sucesión de posiciones $p_1 ... p_n$ y muestras control para el modelo para el que se calculan los Puntos de trabajo.

RF, VR : Vectores modelo de cobertura normalizada de referencia y variación para la región R y el modelo S.

WP: Puntos de trabajo asociados a la región R y el modelo S.

WP se inicializa a 0 y p se inicializa a 1.

A continuación se ejecutan los siguientes pasos

1. Para un modelo, se calcula (803) el promedio de los factores de corrección de cobertura (fc) para la normalización de las muestras de control:

$$fc = \frac{\sum_{i=1}^{|MS|} cobAuto(\mathrm{MS}_i).\varphi}{|\mathrm{MS}|}$$

$\varphi=2$ si *sexo*(MS_i) = Mujer $\wedge$ crom$(R_r) \in \{X,Y\}$

$\varphi=1$ en otro caso.

donde:

MS = Es el conjunto de muestras que forman el conjunto de referencia S_i. SM_i representa la muestra i incluida en dicho conjunto y |MS| es el cardinal del mismo (número de muestras incluidas en este).

cobAutosomal(m) es una función que devuelve la cobertura total en los cromosomas autosómicos y la región de interés para la muestra m.

sexo(m) es una función que devuelve el sexo asociado a una muestra m.

crom(r) es una función que devuelve el cromosoma asociado a la función r.

φ es un factor de corrección del número de copias diferenciales para X en hombres y mujeres, toma valor dos en el caso de regiones sobre el cromosoma X cuando el género de la muestra bajo cálculo es mujer, en otro caso vale 1.

2. Si el tamaño en pares de bases de la región en estudio es igual o inferior al determinado por el parámetro de configuración Cfg.WPwindow, toda posición de la región es considerada local para otra de la misma:
    a. Obtener (804) las tasas de variación de señal asociadas al modelo en la región, considerando únicamente los valores correspondientes a posiciones *"bien medidas"*. Por posiciones "*bien medidas*" se entiende aquellas posiciones que pertenecen a los intervalos de interés originales registrados para la región R y para los que la cobertura normalizada de referencia para el modelo multiplicada por fc (el factor de corrección promedio) es igual o superior a la cobertura límite establecida por el parámetro de configuración Cfg.WPmincover (por defecto 50). La tasa de variación para una posición viene determinada por el cociente entre la variación y la señal de referencia asociadas al modelo para dicha posición.

$$\text{Tasas Variación}_{ok} = \{ \frac{VR_i}{RF_i} / (RF_i * fc \geq cfg.WPmcb) \wedge (p_i \subset \cup R.icore_j) \}$$

donde:

R.icorej son las regiones originalmente seleccionadas como target para el análisis asociadas a una región de estudio que a partir de ellas el método ha generado. $\cup R.icore_j$ representa el conjunto (unión) de todas estas regiones.

p_i representa una posición dada "i" incluida en la región en estudio.

RF_i es el valor de referencia para el modelo en la posición "i"

VR_i es el la variación asociada al modelo para la posición "i"

$TasasVariación_{ok}$ = Es el conjunto de cocientes de la variación entre el valor de referencia (es decir las tasas de variación) para el modelo calculados para las posiciones "i" que cumplen las condiciones citadas.

b. Registrar (805) como puntos de trabajo aquellas posiciones *"bien medidas"* de la región, cuya tasa de variación es inferior al promedio calculado sobre el conjunto $TasasVariación_{ok}$ definido anteriormente más la desviación estándar calculada también para $TasasVariación_{ok}$ y multiplicada por un factor (establecido con Cfg.WPstringence, por defecto 2), siempre y cuando exista más de una posición *"bien medida"* en la región, sino ninguna posición se considerará *"punto de trabajo"*.

$$\text{WP si } TV_p < \overline{TV_{ok}} + \sigma(TV_{ok}) \ . \ \text{cfg.WPstringence}$$

donde:

TV_p representa la tasa de variación asociada a un modelo de referencia para la posición p.

TV_{ok} es una abreviatura de la $TasaVariación_{ok}$ definida anteriormente. $\overline{TV_{ok}}$ representa la media del conjunto de tasas de variación ok calculadas para la región en estudio, $\sigma(TV_{ok})$ representa la desviación típica para dicho conjunto.

WP es el conjunto de puntos de trabajo para la región y modelo en estudio.

3. Si el tamaño de la región es superior a Cfg.WPwindow.

   a. Calcular (806) la tasa de variación inicial límite lim_{izq} como el promedio de las tasas de variación vinculadas a las posiciones “bien medidas” entre las primeras Cfg.WPwindow posiciones de la región + la desviación típica por el coeficiente establecido en Cfg.WPstringence. En caso de que no existan dos o más posiciones “bien medidas” entre las primeras posiciones, se tomará lim_{izq} como 0.

$$lim_{izq} = \overline{TV_{ok-izq}} + \sigma(TV_{ok-izq}) \quad .\text{cfg.WPstringence}, \quad lim_{izq} = 0 \text{ si } TV_{ok\text{-}izq} = \emptyset$$

donde:

$TV_{ok\text{-}izq}$ representa el conjunto de Tasas de Variación$_{ok}$ calculadas únicamente considerando las primeras cfgWPwindow posiciones de la región en estudio. $\overline{TV_{ok-izq}}$ representa el promedio de dichas tasas y $\sigma(TV_{ok-izq})$ la desviación típica.

lim_{izq} es el valor límite para la tasa de variación de una posición localizada entre las cfg.WPwindow primeras de la región en estudio para que sea considerada un punto de trabajo o (WP, working point).

   b. Calcular la tasa de variación final límite lim_{der} como el promedio de las tasas de variación vinculadas a las posiciones “bien medidas” entre las últimas Cfg.WPwindow posiciones de la región + la desviación típica por el coeficiente establecido en Cfg.WPstringence. En caso de que no existan dos o más posiciones “bien medidas” entre las últimas posiciones, se tomará lim_{izq} como 0.

$$lim_{der} = \overline{TV_{ok-der}} + \sigma(TV_{ok-der}) \quad .\text{cfg.WPstringence}, \; lim_{der} = 0 \text{ si } TV_{ok\text{-}der} = \emptyset$$

donde:

$TV_{ok\text{-}der}$ representa el conjunto de Tasas de Variación$_{ok}$ calculadas únicamente considerando las últimas cfgWPwindow posiciones de la región en estudio. $\overline{TV_{ok-der}}$ representa el promedio de dichas tasas y $\sigma(TV_{ok-der})$ la desviación típica.

lim_{izq} es el valor límite para la tasa de variación de una posición localizada entre las cfg.WPwindow últimas de la región en estudio para que sea considerada un punto de trabajo o (WP, working point).

c. Tomar como WP aquellas posiciones “bien medidas” de entre las primeras ceil(Cfg.WPwindow/2) posiciones de la región cuya tasa de variación es inferior a lim_{izq}, donde la función ceil(x) redondea al número entero superior o igual a x más pequeño posible.
d. Tomar como WP aquellas posiciones “bien medidas” de entre las últimas ceil(Cfg.WPwindow/2) posiciones de la región cuya tasa de variación es inferior a lim_{der}.
e. Para cada posición intermedia *p* de la región [no incluída entre las primeras ceil(Cfg.WPwindow/2) ni últimas ceil(Cfg.WPwindow/2) posiciones de la región]:
    i. Calcular la tasa de variación límite para la posición p lim_p. Lim_p se calcula tomando las posiciones “bien medidas” en el intervalo [p- Cfg.WPwindow/2, p + Cfg.WPwindow/2], cuando este conjunto sean dos o más posiciones lim_p es el promedio de las tasas de variación vinculadas a dichas posiciones más la desviación típica multiplicada por Cfg.WPstringence, en otro caso es 0.
    ii. Incorporar (807) la posición p como “punto de trabajo” cuando su tasa de variación es inferior a lim_p.

$$lim_p = \overline{TV_{ok-p}} + \sigma(TV_{ok-p}) \,.\, \text{cfg.WPstringence},\ lim_p = 0 \text{ si } TV_p = \emptyset$$

donde:

$TV_{ok\text{-}p}$ representa el conjunto de Tasas de Variación$_{ok}$ calculadas únicamente considerando las posiciones en el rango [p- Cfg.WPwindow/2, p + Cfg.WPwindow/2] de la región en estudio. $\overline{TV_{ok-p}}$ representa el promedio de dichas tasas y $\sigma(TV_{ok-p})$ la desviación típica.

lim_p es el valor límite para la tasa de variación de una posición localizada dentro del rango [p- Cfg.WPwindow/2 , p + Cfg.WPwindow/2] de la región en estudio para que sea considerada un punto de trabajo o (WP, working point).

El proceso finaliza (808) cuando se verifica p=n, donde n es el número de posiciones de la región en estudio.

La figura 9 presenta el estudio, para cada muestra, del comportamiento de su señal de cobertura frente al modelo de referencia que le corresponde, constituye el algoritmo de búsqueda, una vez obtenidas las coberturas para las muestras de estudio en las regiones pertenecientes a un bloque de trabajo, y calculados los modelos correspondientes a los distintos conjuntos de control. El protocolo de estudio revisa las señales de las muestras a lo largo de las posiciones válidas (“puntos de trabajo”), en las regiones de interés, para la detección de CNVs candidatos, registrando los rangos cromosómicos con señales alteradas respecto a la señal de referencia para el correspondiente modelo y caracterizando dichos rangos para finalmente fusionar (907) aquellos que, se considera, forman parte de una zona más amplia que estaría alterada y los incluye.

El proceso comienza (901) inicializando (902) las siguientes variables:

ROI = { R1 , R2, … , Rn } : Regiones de la iteración ordenadas por cromosoma y posición inicial.

extensiones := Cfg.maxExten : Variable que indica el número de extensiones disponibles para controlar el flujo del algoritmo y que inicialmente se establece según un parámetro de configuración. Una extensión supone considerar una posición que no tiene una señal

alterada en base al correspondiente modelo como una candidata a un evento de variante estructural.

outlierActivo := NO : Variable de tipo lógico que durante el proceso de búsqueda indica si en un momento dado el examen de una posición constituye la continuación de una alteración detectada en caso de que esta también lo esté o es la primera de un nuevo bloque.

wpexten=0; wpTam:=0: Se trata de variables contadores, la primera indica el número de extensiones que se han usado y la segunda el número de puntos de trabajo abarcados.

A continuación, el proceso se lleva a cabo de la siguiente forma:

1. Llevar un contador de posiciones no alteradas ignorables (*extensiones*) que se inicializa al valor máximo que puede alcanzar según se ha configurado en Cfg.maxExtensiones y un registro de situación (*outlierActivo*) que inicialmente está a estado “outlier inactivo”. Además inicializar unos registros contadores de puntos de trabajo: extendidas (wpexten) y abarcadas (wpTam), ambos registros se inicializar con valor 0.
2. Posicionarse en la primera posición de la primera región asignada al bloque de trabajo (según el orden cromosoma-coordenada cromosómica).
3. Si el registro de situación está activo (outlierActivo = SI), cuando el grupo al que pertenece la región actual sea distinto del grupo asignado a dicho outlier, cerrar (903) el registro del outlier (pasar estado a “*outlier inactivo*”) asignando como posición final la última posición significativa y como región final la que la contiene, después si el número de puntos de trabajo que abarca (wptam-wpext) es igual o superior a Cfg.minTamaño, procesar (904) dicho outlier (el proceso para procesar un outlier se describe más adelante).
4. Verificar si la posición actual es válida (punto de trabajo), en caso negativo actualizar la posición actual y la región actual para avanzar a la siguiente posición de las regiones a procesar del bloque de trabajo y volver al paso 3.
5. Incrementar el valor de wpTam en una unidad (la posición actual es WP).
6. Calcular (905) la distancia, entre la señal de la muestra y referencia en el modelo para la posición. La distancia se calcula como la diferencia en valor absoluto entre la cobertura normalizada de la muestra y la de referencia para su modelo de control dividida entre la variación del modelo. Cuando la cobertura normalizada de la muestra es superior o igual a la de referencia para el modelo, el signo de la distancia es positivo, sino negativo.

$$d = distancia(p) = \left|\frac{NC_{m,p} - RF_{s,p}}{VR_{s,p}}\right| \quad d.d.\ m\text{: } muestra, s\text{: } cto\ control$$

$$sd = signodistancia(p) = " + " si\ \ NC_{m,p} > \ RF_{s,p}\ \ \ sino = " - "$$

donde:

$NC_{m,p}$ es la cobertura normalizada en la posición p para la muestra m.

$RF_{s,p}$ es el valor de referencia para ls posición p según conjunto de control s (asignado para la muestra m).

$VF_{s,p}$ es el valor de variación para ls posición p según conjunto de control s (asignado para la muestra m).

7. En caso de que la situación actual sea “outlier activo”:
   7.1. Si el signo asociado al outlier activo difiere del signo asociado a la posición actual (sd) cerrar el registro del outlier (pasar estado a “*outlier inactivo*”) asignando como posición final la última posición significativa y como región final la que la contiene, después si el número de puntos de trabajo que abarca (wptam-wpext) es igual o superior a Cfg.minTamaño, procesar dicho outlier y después continuar en el paso 9.
   7.2. Si la distancia muestra-referencia (*d*) es inferior al límite establecido por el parámetro de configuración Cfg.distmin (por defecto 1.5), cuando el valor del contador *extensiones* es mayor que cero disminuirlo en una unidad, incrementar el contador de puntos de trabajo extendidas y continuar en el paso 9, en otro caso continuar cerrar el registro del outlier (pasar estado a “*outlier inactivo*”) asignando como posición final la última posición significativa y como región final la que la contiene, después si el número de puntos de trabajo que abarca (wptam-wpext) es igual o superior a Cfg.minTamaño, procesar dicho outlier y después continuar en el paso 9.
8. En el caso de que la situación actual sea “outlier inactivo”, si la distancia muestra-referencia (d) es igual o supera el límite establecido (Cfg.distmin), abrir (906) un nuevo registro de outlier (estado a “outlier activo”) asociándole como región inicial, posición inicial, cromosoma, grupo y signodistancia los vinculados a la posición actual y establecer a Cfg.maxExtensiones.

9. En caso de que queden posiciones incluidas en las regiones asignadas al bloque de trabajo por explorar tomar la siguiente a la actual, en la región siguiente si es necesario, y continuar en el paso 3. En caso de que no queden posiciones válidas por explorar, cuando la situación actual sea de “outlier activo” cerrar el registro del outlier (pasar estado a “*outlier inactivo*”) asignando como posición final la última posición significativa y como región final la que la contiene, después si el número de puntos de trabajo que abarca (wptam-wpext) es igual o superior a Cfg.minTamaño, procesar dicho outlier.
10. Fusionar (907) aquellos outliers que forman parte de una zona más amplia que estaría alterada y los incluye, y finalizar (908) el protocolo.

La figura 10 muestra el protocolo de procesado de un rango outlier, cuyo objetivo es revelar la presencia de un CNV, asignar una valoración de confianza en que el rango cromosómico con señales alteradas se corresponda con un rango afectado por un CNV, y recuperar y registrar información contextual para informar el CNV candidato. El proceso se inicia (1001) ajustando (1002) los límites de la región outlier. En caso de que el rango siga superando la longitud establecida en un umbral que denominaremos Cfg.minTamano, se procede a su caracterización (1003), en otro caso se desestima como CNV. La caracterización está basada en la medida de ciertos parámetros del modelo y de la muestra, así como de la región cromosómica donde se sitúa el outlier. Una vez el rango outlier ha sido caracterizado, se determina (1004) si es candidato a CNV, en otro caso se desestima. Los rangos candidatos a CNVs son anotados (1005) con información del contexto experimental y cromosómico (por ejemplo: zonas génicas afectadas, regiones cromosómicas con características especiales, variantes en dicha región, veces que se ha visto la región alterada, información de bases de datos sobre dicha región). Una vez obtenidos todos los datos, se exportan (1006) codificadas las señales en el contexto del área afectada y vecina, para recuperarlas durante la creación del informe final. El último paso del proceso se encarga de asignar (1007) un grado de confianza en que el CNV candidato sea una variante estructural realmente, tras lo cual finaliza (1008) el proceso.

La figura 11 presenta el ajuste de los límites del outlier, que comienza (1101) con la estimación (1102) del ratio muestra-modelo característico del outlier, $\hat{R}$. $\hat{R}$ se calcula como la mediana del conjunto de cocientes de cobertura normalizada de la muestra entre la

referencia para el modelo que le corresponda, considerando las posiciones que son punto de trabajo dentro del intervalo outlier inicialmente detectado.

$$\hat{R} = med\left(\left\{ \frac{\mathrm{NC}_{m,p}}{\mathrm{RF}_{s,p}} \ / \ p \subset [\mathrm{O}_{\mathrm{ini}}, \mathrm{O}_{\mathrm{fin}}] \wedge p \subset \mathrm{WP} \right\}\right)$$

donde:

$NC_{m,p}$ es la cobertura normalizada en la posición p para la muestra m.

$RF_{s,p}$ es el valor de referencia para ls posición p según conjunto de control s (asignado para la muestra m).

WP es el conjunto de puntos de trabajo para la muestra en estudio y p una posición incluida en dicho conjunto.

O_{ini} y O_{fin} son respectivamente la posición de inicio y fin asociadas al outlier.

Si $\hat{R}$ es mayor que 1 el outlier se correspondería con una ganancia de señal. Cuando el ratio asociado a la posición inicial es menor o igual que (1+$\hat{R}$)/2 se recalcula la posición de inicio disminuyéndola en una posición hasta que se alcance la posición de inicio de la región en el que comienza el outlier inicial o bien el ratio $\frac{\mathrm{NC}_{m,p}}{\mathrm{RF}_{s,p}}$ para la nueva posición no supere (1+$\hat{R}$)/2. Cuando el ratio asociado a la posición inicial supera (1+$\hat{R}$)/2 se recalcula la posición de inicio aumentándola en una posición mientras el ratio $\frac{\mathrm{NC}_{m,p}}{\mathrm{RF}_{s,p}}$ para la nueva posición sea inferior a (1+$\hat{R}$)/2 y ls podición inferior a la última perteneciente a la región inicial para el outlier. Para el recálculo de la posición final se procede a la inversa. Cuando el ratio asociado a la posición final es mayor que (1+$\hat{R}$)/2 se recalcula la posición de fin incrementándola en una posición hasta que se alcance la posición de fin de la región en el que termina el outlier inicial o bien el ratio $\frac{\mathrm{NC}_{m,p}}{\mathrm{RF}_{s,p}}$ para la nueva posición no supere (1+$\hat{R}$)/2. Cuando el ratio asociado a la posición final no supera (1+$\hat{R}$)/2 se recalcula la posición de fin decrementándola en una posición mientras el ratio $\frac{\mathrm{NC}_{m,p}}{\mathrm{RF}_{s,p}}$ para la nueva posición sea menor que (1+$\hat{R}$)/2 y no se haya alcanzado la última posición asignada a la última región afectada por el outlier inicial.

Cuando $\hat{R}$ no supera la unidad el outlier se correspondería con una pérdida de señal. Cuando el ratio asociado a la posición inicial es menor que (1+$\hat{R}$)/2 se recalcula la posición

de inicio disminuyéndola en una posición hasta que se alcance la posición de inicio de la región en el que comienza el outlier inicial o bien el ratio $\frac{\mathrm{NC}_{m,p}}{\mathrm{RF}_{s,p}}$ para la nueva posición no sea inferior a (1+$\hat{R}$)/2. Cuando el ratio asociado a la posición inicial es igual o supera (1+$\hat{R}$)/2 se recalcula la posición de inicio aumentándola en una posición mientras el ratio $\frac{\mathrm{NC}_{m,p}}{\mathrm{RF}_{s,p}}$ para la nueva posición supere (1+$\hat{R}$)/2 y no se haya alcanzado la posición final de la región inicial asignada al outlier. Para el recálculo de la posición final se procede a la inversa. Cuando el ratio asociado a la posición final es menor que (1+$\hat{R}$)/2 se recalcula la posición de fin incrementándola en una posición hasta que se alcance la posición de fin de la región en la que termina el outlier inicial o bien el ratio $\frac{\mathrm{NC}_{m,p}}{\mathrm{RF}_{s,p}}$ para la nueva posición no sea inferior a (1+$\hat{R}$)/2. Cuando el ratio asociado a la posición final no es inferior a (1+$\hat{R}$)/2 se recalcula la posición de fin decrementándola en una posición mientras el ratio $\frac{\mathrm{NC}_{m,p}}{\mathrm{RF}_{s,p}}$ para la nueva posición sea mayor que (1+$\hat{R}$)/2 y no se haya alcanzado la posición final de la última región afectada por el outlier. Tras la actualización (1103) de O_{ini} y O_{fin}, finaliza (1104) el proceso.

Completado el ajuste de los límites del outlier si el número de pares que abarca es inferior al valor definido por Cfg.minTanaño se descarta como CNV el intervalo y se prosigue la búsqueda si es el caso, según ha sido descrita de otros intervalos afectados, cuando dicho tamaño mínimo es satisfecho comienza la caracterización del outlier, tal como se presenta en la figura 12.

El proceso comienza (1201) cargando (1202) las siguientes variables:

O,m,s : Outlier que está siendo caracterizado, muestra y modelo vinculados.

WP : Puntos de trabajo para s en el intervalo cromosómico que abarca O.

$NC_{m,p}$: Cobertura normalizada para la muestra (m) en la posición p.

$RF_{s,p}$: Cobertura normalizada de referencia para s en la posición p.

$VR_{s,p}$: Variación (Rango intercuartílico) asociada en s a la posición p.

A continuación, se mantiene (1203) o se corrige (1204) el sexo:

$$\varphi=2 \text{ si sexo(Mm)} = \text{Mujer} \wedge \text{crom(Rr)} \in \{X,Y\},\ \varphi=1 \text{ otro caso.}$$

La caracterización consiste en registro de los siguientes parámetros vinculados al outlier O para una muestra m y su correspondiente modelo de referencia s:

a) Cobertura (1205) del modelo característica para O, O.cob. Este valor se calcula como el producto de la cobertura total autosómica de la muestra por la mediana de las coberturas normalizadas de referencia para el modelo vinculadas a cada una de las posiciones incluidas como Puntos de trabajo en el intervalo cromosómico que abarca O. El resultado del producto anterior se multiplica adicionalmente por 2 en caso de que el sexo asignado a la muestra m sea mujer y el rango cromosómico vinculado al outlier esté localizado sobre un cromosoma sexual.

$$\mathrm{O.cob} = \varphi \,.\, cobAuto(m) \,.\, md(\{RF_{s,p} \,/\, p \;\in WP\})$$

donde:

φ es un factor de corrección del número de copias diferenciales para X en hombres y mujeres, toma valor dos en el caso de regiones sobre el cromosoma X cuando el género de la muestra bajo cálculo es mujer, en otro caso vale 1.

cobAuto(m) es una función que devuelve la cobertura total en los cromosomas autosómicos y la región de interés para la muestra m.

p es una posición perteneciente al conjunto de puntos de trabajo.

md(X) es mediana para el conjunto X.

$RF_{s,p}$ es el valor de referencia para la posición p según conjunto de control s (asignado para la muestra m).

WP es el conjunto de puntos de trabajo para la muestra en estudio y p una posición incluida en dicho conjunto.

b) Variación (1206) del modelo característica para O, O.var. Este valor se calcula como la mediana de las variaciones de cobertura normalizada para el modelo vinculadas a cada una de las posiciones incluidas como Puntos de trabajo en el intervalo cromosómico que abarca O

$$\mathrm{O.var} = md(\{VR_{s,p} \,/\, p \;\in WP\})$$

donde:

p es una posición perteneciente al conjunto de puntos de trabajo.

md(X) es mediana para el conjunto X.

$VR_{s,p}$ es la variación vinculada a la posición p según el conjunto de control s (asignado para la muestra m).

WP es el conjunto de puntos de trabajo para la muestra en estudio y p una posición incluida en dicho conjunto.

c) Tasa de variación (1207) del modelo característica para O y su variación a lo largo de O, O.var y O.vtvar. O.var se calcula como la mediana de los cocientes entre variaciones de cobertura normalizada y los valores de cobertura normalizada de referencia para el modelo vinculados a cada una de las posiciones incluidas como *Puntos de trabajo* en el intervalo cromosómico que abarca O. O.vtvar se calcula como el rango intercuartílico para el conjunto de los cocientes considerados para el cálculo de O.var.

$$0.\text{tvar} = md(\{\frac{RV_{s,p}}{RF_{s,p}} \; / \; p \; \in WP\})$$

$$\text{O.vtvar} = iqr(\{\frac{RV_{s,p}}{RF_{s,p}} \; / \; p \; \in WP\})$$

donde:

p es una posición perteneciente al conjunto de puntos de trabajo.

md(X), iqr(X) son respectivamente la mediana y el rango intercuartílico para el conjunto X.

$RF_{s,p}$ es el factor de referencia para la posición p según el conjunto de control s (asignado para la muestra m).

$VR_{s,p}$ es la variación vinculada a la posición p según el conjunto de control s (asignado para la muestra m).

WP es el conjunto de puntos de trabajo para la muestra en estudio y p una posición incluida en dicho conjunto.

d) Ratio de cobertura (1208) característico para O y su variación a lo largo de O, O.ratio y O.ratiovar. O.ratio se calcula como la mediana de los cocientes entre las coberturas normalizadas para la muestra y las de referencia para el modelo vinculado, considerando las posiciones incluídas como *Puntos de trabajo* en el intervalo cromosómico que abarca O. O.ratiovar se calcula como el rango intercuartílico para el conjunto de los cocientes considerados para el cálculo de O.ratio.

$$0.\text{ratio} = md(\{\frac{NC_{m,p}}{RF_{s,p}} / \, p \in WP\})$$

$$0.\text{ratiovar} = iqr(\{\frac{NC_{m,p}}{RF_{s,p}} / \, p \in WP\})$$

donde:

p es una posición perteneciente al conjunto de puntos de trabajo.

md(X), iqr(X) son respectivamente la mediana y el rango intercuartílico para el conjunto X.

$RF_{s,p}$ es el factor de referencia para la posición p según el conjunto de control s (asignado para la muestra m).

$NC_{m,p}$ es la cobertura normalizada en la posición p para la muestra m.

WP es el conjunto de puntos de trabajo para la muestra en estudio y p una posición incluida en dicho conjunto.

e) Distancia característica (1209) para O y su variación a lo largo de O, O.distancia y O.distanciavar. O.distancia se calcula como la mediana de los cocientes entre las diferencias de coberturas normalizadas, para la muestra y el valor de referencia según el modelo vinculado, y las variaciones de cobertura normalizada para el modelo en las posiciones consideradas *Puntos de trabajo* en el intervalo cromosómico que abarca O. O.distanciavar se calcula como el rango intercuartílico para el conjunto de los cocientes considerados para el cálculo de O.distancia.

$$\text{O.distancia} = md(\{\frac{NC_{m,p} - RF_{s,p}}{RV_{s,p}} \ / \ p \ \in WP\})$$

$$\text{O.distanciavar} = iqr(\{\frac{NC_{m,p} - RF_{s,p}}{RV_{s,p}} \ / \ p \ \in WP\})$$

donde:

p es una posición perteneciente al conjunto de puntos de trabajo.

md(X), iqr(X) son respectivamente la mediana y el rango intercuartílico para el conjunto X.

$RF_{s,p}$ es el factor de referencia para la posición p según el conjunto de control s (asignado para la muestra m).

$VR_{s,p}$ es la variabilidad asociada a la posición p según el conjunto de control s (asignado para la muestra m).

$NC_{m,p}$ es la cobertura normalizada en la posición p para la muestra m.

WP es el conjunto de puntos de trabajo para la muestra en estudio y p una posición incluida en dicho conjunto.

Además de calcular los valores anteriores, al outlier se le asocian (1210) las posiciones cromosómicas detectadas pero también las posiciones extremas compatibles con una variante estructural, es decir tanto el rango cromosómico más amplio que podría llegar a abarcar como el mínimo a modo de intervalo de confianza para los extremos del outlier detectado (considerar por ejemplo que hay zonas que no pertenecen a la región de interés o para las cuales no se tiene cobertura). El protocolo para determinar dichas las posiciones extremas correspondientes al intervalo más amplio que podría verse afectado parte de los límites detectados para el outlier (O.posini y O posfin) ampliándolos mientras se recorren las posiciones adyacentes no incluidas en O, hasta encontrar una agregación de posiciones (establecida mediante una variable Cfg.BRKnoCNV, por defecto tomada como 20) consideradas puntos de trabajo *y* con una tasa de variación no superior a la característica para la región outlier más de tres veces su variación característica, para las cuales el incremento/decremento en ratio de cobertura muestra-modelo es igual o inferior a 1/3 característico para el outlier inicial (siendo el ratio superior a 1 cuando lo sea el característico del outlier o inferior a este cuando lo sea para el outlier). Para determinar las posiciones extremas correspondientes al intervalo mínimo que se considera que estaría

afectado sigue un proceso similar pero recorriendo las posiciones adyacentes a los extremos incluidas en el outlier hasta encontrar una afregación de posiciones consideradas *Puntos de trabajo* y con una tasa de variación no superior a la característica para la región outlier más tres veces su variación característica, para las cuales el incremento/decremento en ratio cobertura muestra-modelo es igual o superior a 2/3 del característico para el outlier inicial (siendo el ratio superior a 1 cuando lo sea el característico del outlier o inferior a este cuando lo sea para el outlier). Tras registrar las posiciones extremas, afectadas, observadas y potenciales, finaliza (1211) el proceso.

La figura 13 ilustra una posible implementación para la ampliación del límite inferior del intervalo. El proceso parte (1301) de la inicialización (1302) a 0 de un contador de posiciones no alteradas *noCNVs*. Considerando la posición p, la de inicio el outlier, así como r la región de inicio del outlier:

1) Si p es la posición inicial de la primera de las regiones disponibles en el bloque de regiones analizado ir al paso 6 para terminar el protocolo. En otro caso establecer p a la posición previa dentro de las regiones de interés, esto es: si p es mayor que la posición de inicio de la región actual r decrementar p en una unidad, en otro caso decrementar r en una unidad (explorar la región previa) estableciendo p al valor de su posición final.
2) Si para la posición explorada, p, no se satisface que: p es un Punto de trabajo para el modelo vinculado a la muestra que está siendo analizada y la tasa de variación para el modelo (variación entre valor de referencia) no supera para el outlier a su tasa de variación más 3 veces la variación asociada a dicha tasa entonces volver al paso 1.
3) Cuando no se satisface la proposición ω, decrementar en una unidad el contador noCNVs cuando este tenga un valor superior a 0 y volver al paso 1.

$$\omega \equiv \left(\frac{NC_{m,p}}{RF_{s,p}} - 1\right).\left(\frac{O.ratio - 1}{3}\right) \leq \left(\frac{O.ratio - 1}{3}\right)^2$$

donde:

$RF_{s,p}$ es el factor de referencia para la posición p según el conjunto de control s (asignado para la muestra m).

$NC_{m,p}$ es la cobertura normalizada en la posición p para la muestra m.

O.ratio es el ratio de cobertura característico para el outlier bajo revisión.

4) Si el contador noCNVs es inferior a Cfg.BRKnoCNV ir al paso 1.

5) Asignar (1303) como posición mínima posición inicial, O.imin (límite inferior del intervalo), la posición p menos el número de posiciones establecidos en el valor de configuración Cfg.BRKdecay (por defecto 25), como región mínima de inicio, O.irmin, asignar r y finalizar (1304) el protocolo.

Para calcular la posición final máxima puede seguirse un procedimiento análogo, incrementando posiciones desde la final asignada al outlier. Para el cálculo del intervalo mínimo se procede también de manera análoga pero empleando como condición ω la siguiente (y en este caso el significado en lugar de contabilizar posiciones *noCNVs* se contabilizan posiciones *siCNVs*:

$$\omega \equiv \left(\frac{NC_{m,p}}{RF_{s,p}} - 1\right).\left(\frac{2(O.ratio-1)}{3}\right) \geq (\frac{2(O.ratio-1)}{3})^2$$

donde:

$RF_{s,p}$ es el factor de referencia para la posición p según el conjunto de control s (asignado para la muestra m).

$NC_{m,p}$ es la cobertura normalizada en la posición p para la muestra m.

O.ratio es el ratio de cobertura característico para el outlier bajo revisión.

La figura 14 muestra el proceso de determinar si un outlier puede reflejar un una variante estructural, una vez completada su caracterización. El proceso se inicia (1401) seleccionando individualmente (1402) cada outlier. Se considera que el outlier seleccionado puede reflejar una variante estructural en los siguientes casos:

1) Si el tamaño en pares de bases que abarca la región detectada como outlier es igual o mayor el mínimo exigido (Cfg.minTamaño), la cobertura modelo asignada al outlier es igual o mayor al límite establecido para la posiciones *Puntos de trabajo* (cfg.WPmcb) y la distancia característica del outlier es igual o mayor a la mínima exigida (cfg.dmin) evaluar el ratio característico (punto 2) en otro caso desestimar el outlier como candidato.

2) Cuando el ratio característico no está comprendido en el intervalo [0.65,1.35] considerar el CNV como candidato, en otro caso únicamente si supera la distancia

mínima exigida, exponencial con base cfg.dmin y máximo en 1. La distancia mínima para ratio anómalo ψ se calcula (1403) de acuerdo con la función:

$$\psi = cfg.dmin^{((0.6-|O.\widehat{ratio}-1|)\ .\ 15)}$$

donde:

$cfg.dmin$ es la distancia mínima exigida (por configuración) para considerar un outlier como candidato a CNV.

$O.\widehat{ratio}$ es el ratio de cobertura característico para el outlier bajo revisión.

Los outliers candidatos se registran (1404) en el registro correspondiente y finaliza el proceso. Para aquellos outliers que son considerados candidatos a CNVs se procede a la recuperación de información contextual, anotándolos y evalando una serie de filtros que les son etiquetados. La anotación implica recuperar las anotaciones en las distintas fuentes de información suministradas al sistema para todas aquellas posiciones incluidas dentro de las regiones abarcadas por el outlier (aunque sea parcialmente) así como dentro las regiones previas y siguientes (siempre que estén disponibles). Se recuperan también todas las anotaciones hechas sobre rangos cromosómicos que estén contenidos o solapen el intervalo abarcado por el outlier. Como fuentes de información se incluyen las bases de datos, los ficheros .bed, .gff etc y los ficheros de variantes .vcf y de alineamientos *.bam*.

Entre las fuentes de información requeridas y usadas para resolver algunos filtros están los mapeos cromosómicos de las distintas isoformas de los distintos genes que se quieren anotar. Para las regiones en el contexto del outlier mencionadas se recuperan y asocian al registro de el outlier candidato a CNV: las coordenadas de inicio y fin de cada región UTR, intrónica y codificante.

Una vez recuperada la información sobre regiones génicas calcula una isoforma compendio donde las regiones codificantes de las distintas isoformas se fusionan, lo mismo los UTR para el resto de bases no afectadas, y entonces se registra para el outlier el número de bases codificantes que abarca potencialmente (es decir incluidas entre los extremos máximos O.mini y Omfin), el número de regiones codificantes, el tamaño en pares de bases de dichas regiones codificantes y la distancia mínima de alguno de los extremos del outlier a

una base codificante (cuando la distancia sea a una base abarcada por outlier tomará signo negativo y en otro caso positivo). Cuando el valor de distancia a cdna sea inferior al determinado por Cfg.FILTdcdna (por defecto 10) se etiquetará el outlier como filtrado por cdna (esto permite activar dicho filtro a la hora de mostrar los resultados en el subsistema de exploración), también puede implementarse esta característica como un filtro duro que elimina el outlier de los resultados directamente.

Cuando se proporcione al sistema los genes/rangos cromosómicos pertenecientes a distintos paneles, se procesa dicha información para establecer un filtro por panel, estando un outlier activo para cada panel en concreto si alguna de las regiones de interés para dicho panel solapa el la región cromosómica potencialmente abarcada por el outlier.

En cuanto a los recursos de anotación de variantes estructurales de distintas fuentes (ej. la base de datos dgv) se anotará el máximo porcentaje de solapamiento entre alguna de las variantes estructurales descritas del mismo tipo (ganancias o pérdidas de copias) y el rango abarcado por el outlier (para cada variante estructural el máximo vendrá determinado por el porcentaje de bases de la variante descrita en la base de datos cubiertas potencialmente por el outlier o bien por el porcentaje de bases del outlier determinadas que están cubiertas por la variante estructural descrita).

El subsistema de detección y anotación establecerá una comunicación con una base de datos o un fichero o conjunto de ellos para registrar cada uno de los CNVs candidatos detectados en sus distintas ejecuciones guardando información sobre los rangos detectados y potenciales de cada outlier así como los valores característicos vinculados al ratio y el grado de credibilidad calculado. Además queda registrada la muestra para la que se ha detectado dicho outlier. Este repositorio de información es consultado durante esta fase de anotación recuperando el número de veces que el rango afectado por el outlier (el detectado, no el potencial) es solapado por algún registro en dicho repositorio para una muestra distinta a la que está siendo procesada e igualmente el número de veces que esto se produce únicamente considerando aquellas anotaciones que tengan asignado un grado de credibilidad igual o superior al característico para el outlier en estudio. En base a estos valores de contaje se asignan dos filtros por frecuencia de aparición entre las muestras a las

que se le ha estudiado la región cromosómica afectada por el outlier, la frecuencia límite quedará fijada por el parámetro Cfg.FILTCNVfreq (por defecto 5%).

Además de recuperar información de las distintas fuentes que aportan anotaciones sobre posiciones o intervalos cromosómicos también se recupera a partir de los ficheros de variantes de cada muestra (.vcfs) en caso de estar disponibles aquellas variantes que solapen o estén contenidas en las regiones del contexto del outlier que está siendo anotado, se registran sus coordenadas cromosómicas así como su calidad y frecuencia del alelo alternativo.

La figura 15 ilustra el proceso de detección de rupturas (también denominadas habitualmente por su nombre en inglés "*breakpoints*") o su ausencia a partir de los *reads* secuenciados:

1) El proceso se inicia (1501) con la recuperación del fichero de alineamiento de la muestra de estudio los *reads* que solapan el intervalo [0.imin,O.imax] (1502A), es decir que solapen el rango cromosómico entre las dos posiciones extremas para el límite inferior del outlier, estos reads forman el conjunto $READS_{5'}$. Para cada uno de los *reads* recuperados, procesar su información de alineamiento (1503A) para trabajar sobre los que tengan en su correspondiente campo de operaciones del fichero de alineamiento (campo conocido como CIGAR, del inglés *"Compact Idiosyncratic Gapped Alignment Repor"*) las bases laterales enmascaradas –tienen aplicado un *borrado lógico*-, (o como se conoce habitualmente por el término en inglés, presenten "*softclipping*") afectando a más de 10 bases (o un alineamiento quimérico asociado con un número superior a 10 bases segregadas, sin prejuicio de que en lugar de 10 bases puede configurarse el sistema para exigirse un número distinto), recuperar la secuencia que no se puede alinear de forma contigua a la primaria (la afectada por el borrado lógico –softclipped-) (1504A) y calcular la posición cromosómica donde se produce la ruptura. Asignar un clúster de rupturas (break points) (1505A) para el *read* a partir de la posición de ruptura particular que este presenta y registrar en dicho clúster (para su posición de clúster) la secuencia afectada por el borrado lógico.
   Para determinar la posición de ruptura en los casos de softclipping se examina el campo CIGAR del *read*. Cuando la primera operación (ignorando las operaciones de *borrado fuerte* o por su término en inglés de *hard clipping*) del CIGAR sea un softclip de tamaño suficiente, la posición recortada es la posición de inicio de alineamiento del

read más el número de bases que abarca dicha operación. Cuando el bloque de *softclipping* sea la última operación (ignorando las de *hard clipping*) será la posición de inicio de alineamiento del *read* más el número de operaciones cigar match (M) o del (D). Si p es la posición de ruptura, la posición de clúster cp se calcula como: cp = floor((p+5)/10)*10, es decir la parte entera por 10 de la división entre 10 de la posición de ruptura más 5 unidades.

La secuencia de softclip, seqS, se toma de la secuencia del *read* extrayendo tantas bases como se indiquen en el CIGAR vinculadas al bloque de softclipping desde el principio o el final de la secuencia del *read* dependiendo de donde se encuentre el bloque de *softclipping* indicado en el CIGAR.

2) Recuperar del fichero de alineamiento de la muestra de estudio los *reads* que solapen el intervalo [0.fmin,O.fmax] (1502B), es decir que solapen el rango cromosómico que delimita el final del outlier constituido por las dos posiciones extremas como se ha comentado, la mínima coordenada que se vería afectada y la que como mucho podría llegar a estar afectada por una variante estructural subyacente, estos reads forman el conjunto $READS_{3'}$,. Para cada uno de los *reads* recuperados, procesar la información del alineamiento (1503B) para estudiar el CIGAR y como en el paso anterior cuando reúnan el requisito en bases, proceder extrayendo las secuencias *softclipped* (1504B) (y asignándolas a las posiciones de clúster que correspondan (1505B).

3) Eliminar del registro aquellos clústeres (posiciones clúster y sus datos asociados) que tengan un número de secuencias registradas inferior al determinado límite establecido para el sistema Cfg.BRKminsop (por defecto 20) (1506). La motivación de este filtrado es eliminar de la revisión las regiones donde no hay una apilación de softclipping o rupturas (es decir son operaciones que se sospechan motivadas por el ruido de secuenciación y no por eventos biológicos).

4) Para cada uno de los clústeres que permanecen en el registro tras el filtrado del paso 3, recuperar las secuencias que tienen vinculadas (1507) y alinear dichas secuencias para buscar los clúster pareja.

   a. Si el clúster que se está considerando tiene una posición de clúster igual o superior a la posición mínima calculada para el fin del outlier (O.fmin), se alinean sus secuencias (el alineamiento se permite en forward, reverse, complementario o reverse complementario) contra la secuencia de adn para el genoma de referencia comprendida entre las posiciones cromosómicas [O.imin,O.imax] (1508), es decir se busca la pareja en el rango que va desde la posición inicial máxima potencial determinada para

el outlier y la posición inferior extrema potencial. Cuando se produce un alineamiento, siendo ap1 y ap2 las posiciones de inicio y fin del alineamiento contra la secuencia objetivo, se calcula la posición de clúster pareja como pcp = floor((O.imin+ap2+5)/10)*10, esto es, la parte entera de la división entre 10 del cociente de la suma de la posición de final de alineamiento, más la posición cromosómica de inicio de la secuencia contra la que se alinea, más 5 entre 10. Cuando hay alineamiento se aumenta el contador o se abre uno y se establece a 1 (en caso de que no exista el registro previamente) para el par pcp-cp.

b. Cuando el clúster considerado tiene una posición de clúster igual o inferior a la posición calculada como máxima para el inicio del outlier (O.imin), se alinean sus secuencias contra la secuencia de adn para el genoma de referencia comprendida entre las posiciones cromosómicas [O.fmin,O.fmax] (1509), es decir se busca la pareja en el rango que va desde las posiciones extremas para el final del outlier. Cuando se produce un alineamiento, siendo ap1 y ap2 las posiciones de inicio y fin del alineamiento contra la secuencia objetivo, se calcula la posición de clúster pareja como pcp = floor((O.imin+ap1+5)/10)*10, esto es, la parte entera de la división entre 10 del cociente de la suma de la posición de inicial de alineamiento, más la posición cromosómica de inicio de la secuencia contra la que se alinea, más 5 entre 10. Cuando hay alineamiento se aumenta el contador o se abre uno y se establece a 1 (en caso de que no exista el registro previamente) para el par cp-pcp.

5) Registrar para el outlier (1510), para el CNV candidato, la posición de clúster de aquel clúster individual que tenga el mayor número de secuencias vinculadas (O.brk), así como dicho número de secuencias (O.cbrk) para tener constancia del número de *reads* que apoyan dicho clúster. Registrar también el par de posiciones de el par de clústeres que tiene un contador más elevado (O.pbrk), así como dicho contador (O.cpbrk) para tener constancia del número de *reads* que apoyan dicho par de clústeres.

Considerando la información registrada a partir de la revisión de los alineamientos también se establece un filtro. Un CNV se marca como filtrado por breakpoints cuando, siendo la región de inicio y fin del outlier la misma, para alguno de los dos intervalos [O.imin,Oimax],[O.fmin,O.fmax] ambos extremos pertenecen a la misma región y todas sus bases tienen una cobertura cruda para la muestra de estudio igual o superior a la exigida a una posición perteneciente a las *Puntos de trabajo* (cfg.WPmcb) y el número

de *reads* soporte registrado para el clúster de break points (O.cbrk) es inferior a cfg.FILTmbrk.

6) Finalizar (1511) el proceso.

Para cada CNV candidato se persiste un flujo de bytes en un espacio de almacenamiento, (que puede tener carácter diferente dependiendo del despliegue del sistema elegido, por ejemplo podría persistirse en un fichero temporal o de manera definitiva en una base de datos). La etapa de exportación de datos codifica todas las posiciones de las regiones del contexto del CNV (regiones abarcadas por el CNV más las dos anteriores y las dos siguientes siempre que estén disponibles en la iteración –en otro caso dentro de este rango las que sea posible-). Junto con las posiciones se codifica y persiste el valor de las señales procesadas para las distintas muestras y la señal de referencia. La manera en que se codifica dicha información, se estructura y se persiste en el contenedor de datos se especifica posteriormente al desarrollar el elemento contenedor de datos. Este volcado de información, ya sea en un almacenamiento temporal o con carácter definitivo permite, la liberación de recursos al término de una iteración, además su estructuración y codificación permiten un acceso y uso de la información eficiente posteriormente. Una vez identificado un outlier como candidato a CNV, la exportación de datos puede hacerse en cualquier momento a lo largo la iteración correspondiente.

La figura 16 presenta el cálculo y registro de una puntuación que refleja el grado de confianza en que efectivamente el comportamiento de las señales para el CNV candidato refleja una variante estructural subyacente. Dicha confianza se calcula considerando la distancia, el ratio muestra-referencia y su variabilidad así como la cobertura modelo en la región y el soporte (en tamaño o por evidencias de *split-reads*).

El primer paso tras el inicio (1601) es el cálculo (1602) del término de valoración de la distancia asociada al CNV candidato Ω_{dist}. Se calcula otorgando una puntuación de acuerdo con la aplicación de una función que toma valor 0 en 1 y crece de manera simétrica y exponencial al disminuir o aumentar el valor de la distancia hasta alcanzar sendas asíntotas en 1, matemáticamente:

$$\Omega_{dist} = \left|\frac{1 - x^{-\alpha}}{1 + x^{-\alpha}}\right| \; d.d. \; x = O.dist, \alpha = 1.7$$

El valor por defecto para la constante α se ha determinado en 1.7 aunque puede variarse dependiendo de la implementación particular. .

El término de valoración del ratio asociado al CNV Ω_{ratio} se calcula (1603) otorgando una puntuación de acuerdo con la aplicación de una función que es el resultado de 4 funciones gaussianas centradas x = O.ratio = [0 , 0.5 , 1.5 , 2], matemáticamente:

$$\Omega_{ratio} = 0.6 \sum_{i=0}^{4} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\left(\frac{x-\mu_i}{\sigma}\right)^2}})$$

$$d.d.\ x = O.ratio, \mu = \{0{,}0.5{,}1.5{,}2\}, \sigma = 0.1$$

También es importante valorar la fluctuación del ratio a lo largo del rango cromosómico asociado al CNV candidato, para este fin, en el cálculo (1604) del término $\Omega_{ratiovar}$ se aplica la función:

$$\Omega_{ratiovar} = 0.4\ (1 - \frac{1 - 10^{-7.5x}}{1 + 10^{-7.5x}})\ d.d.\ x = O.ratiovar$$

El factor de puntuación debido a la cobertura, Ω_{cob} , se valora (1605) de acuerdo a la siguiente función:

$$\Omega_{cob} = \frac{1 - 1.02^{-x}}{1 + 1.02^{-x}}\ d.d.\ x = O.cob$$

El soporte se valora como máximo en aquellos casos para los cuales se haya localizado un breakpoint con el suficiente soporte, esto es O.brk > cfg.BRKmin, en este caso el término Ω_{sop} toma (1606) valor 1. Cuando no se cumple la condición anterior el soporte se valora en función del tamaño real y potencial del CNV, siendo dicho tamaño O.ptam = 0.5.[(O.mfin - O.mini + 1)+(O.ini-O.fin + 1)]. La valoración del soporte se hace de acuerdo a la función:

$$\Omega_{sop} = \frac{1 - 1.005^{-\alpha}}{1 + 1.005^{-\alpha}}, \alpha = \frac{(O.mfin - O.mini + 1) + (O.ini - Ofin + 1)}{2}$$

donde:

$O.ini, O.fin$ posiciones mínimas detectada para el Outlier.

$O.mini, O.mfin$ posiciones inicial mínima y máxima final que podría llegar a abarcar el outlier.

Finalmente, la puntuación se calcula (1608) siguiendo la siguiente fórmula, tras lo cual finaliza (1609) el proceso:

$$O.score = ceil[\frac{5}{2}(\Omega_{dist} + \Omega_{ratio} + \Omega_{rationvar}) . (\Omega_{cob} + \Omega_{sop})]$$

Dicha puntuación toma valores entre 0 y 10. Nótese, no obstante, que otras formulaciones, rangos, pesos de promedio, etc, pueden utilizarse para evaluar el grado de confianza en sendas realizaciones particulares.

La figura 17 muestra el protocolo de fusión de CNVs candidatos en otros nuevos que los abarquen, una vez exploradas las señales de las regiones incluidas en una iteración de trabajo para un modelo, y caracterizados, filtrados y valorados todos los CNVs candidatos. El protocolo de fusión se inicia (1701) ordenando (1702) la lista de CNVs candidatos según cromosoma y posición de detección inicial. A continuación se siguen los siguientes pasos:

1) Inicializar (1703) la lista de fusión tomando el primer elemento de la lista general ordenada de CNVs candidatos que no ha sido previamente explorado.
2) Si no quedan CNVs por analizar en la lista general ordenada, o el grupo y tipo del último CNV candidato incluido en la lista de fusión y el del siguiente CNV candidato en la lista general no coinciden, efectuar el protocolo de fusión, esto es, ir a paso 4 (el tipo de un CNV candidato es de ganancia cuando el ratio característico es superior a uno o de pérdida cuando es inferior a la unidad). Cuando ninguna de las dos condiciones anteriores se da:
    a. Calcular (1704) un ratio de fusión, *rf*, como el promedio del ratio característico del último elemento incluido en la lista de fusión y el siguiente en la lista general a este. Calcular también la variación de fusión, *vf*, como el máximo de las variaciones de ratio características de estos dos mismos elementos.
    b. Si el valor absoluto de las diferencias de variación de ratio supera la variación de fusión, el último elemento tomado no puede incluirse en la lista de fusión actual, se continúa en el paso cuatro donde se resolverá el conjunto de fusión y este se usará como primer miembro de una nueva lista. Cuando la variación de fusión no se supere se continua en el paso 3 para la búsqueda de bloques intermedios no compatibles con CNVs.

3) Buscar (1705) bloques de posiciones, entre las afectadas por el último CNV candidato incluído en la lista de fusión y el siguiente elemento en la lista general, que serían incompatibles con la existencia de CNVs.

   a. Inicializar un contador (noCNV) a 0, y tomar como posición y región de inicio para la exploración la última región y posición abarcadas por el último CNV detectado de la lista de fusión (CNV_e).

   b. Valorar la posición actual:

      i. Si no es inferior a la primera posición abarcada por CNV_e: Cuando el valor del contador noCNV no supera el valor umbral especificado por el parámetro cfg.FUSmnoCNV (por defecto 10), incorporar CNV_e y continuar en 2 tomando como CNV_e el CNV candidato detectado de la lista general siguiente al que acaba de ser incluido en la lista de fusión. Cuando el contador noCNV es igual o superior a cfg.FUSnoCNV CNV_e no puede añadirse a la lista de fusión, se continúa en el paso cuatro donde se resuelve el conjunto de fusión y este elemento será el primer miembro de una lista de fusión nueva (salvo que sea el último de la lista general).

      ii. Cuando la posición actual es inferior a la primera abarcada por el CNV candidato detectado CNV_e:

         1. si la posición pertenece a los *Puntos de trabajo* y el ratio vinculado para la muestra de estudio en dicha posición no está dentro del intervalo definido por el ratio de fusión más/menos dos veces la variabilidad de fusión o no se cumple la siguiente desigualdad:

$$\left(\frac{NC_{m,p}}{RF_{s,p}} - 1\right).\left(\frac{2(rf-1)}{3}\right) \geq \left(\frac{2(rf-1)}{3}\right)^2$$

donde:

$RF_{s,p}$ es el factor de referencia para la posición p según el conjunto de control s (asignado para la muestra m).

$NC_{m,p}$ es la cobertura normalizada en la posición p para la muestra m.

rf es el ratio de de fusión.

incrementar el contador noCNV, si entonces es inferior al valor umbral cfg.FUSnoCNV evaluar la siguiente posición yendo al paso 3.b, en otro caso continuar en el paso 4 pues CNV_e no puede añadirse a la lista de fusión, sino que se tomará como el primer miembro de una lista de fusión nueva, resolviendo la fusión de los miembros actuales de la lista de fusión.

2. Cuando no se cumplen los tres criterios requeridos en el punto anterior (ii.1), entonces, se decrementa el contador noCNV en una unidad (salvo que ya esté a 0) y se evalúa la siguiente posición regresando al paso 3b.

4) Resolver un conjunto de fusión. Cuando la lista de fusión consta sólo de un miembro no se produce fusión y simplemente se vacía la lista de fusión, en otro caso se procesa un nuevo outlier que tiene por posiciones extremas de detección inicial: la posición inicial del primer CNV candidato incluido en la lista de fusión y la posición final detectada del último CNV candidato incluido en la lista de fusión. Cuando se produce la fusión de varios elementos, se recupera (1706) la credibilidad máxima. En caso de que la puntuación de credibilidad del nuevo CNV candidato fruto de la fusión sea inferior a la de alguno de los miembros fusionados se restablece a dicho valor más alto. Una vez terminado el proceso de resolución se desestiman (1707) los registros de los CNVs candidatos que han sido fusionados quedando un único registro para el nuevo CNV candidato fruto de la fusión. Si quedan elementos sin explorar en la lista general de CNVs candidatos se inicializa una nueva lista de fusión con el CNV candidato detectado siguiente en dicha lista al último incluido en la lista de fusión previa y se continúa en el paso 2, en otro caso se procesa la fusión (1708) y termina (1709) el proceso.

La información generada por el subsistema de detección (101) y anotación tras todos los procesos descritos se almacena finalmente en el subsistema contenedor de datos (102), de

modo que puede accederse a ella desde el subsistema de exploración (103). El subsistema contenedor de datos (102) dispone de un flujo de bytes para cada cnv candidato. El flujo de bytes codifica, para todo el intervalo de exploración asociado al cnv candidato, las posiciones incluidas según la resolución de exportación configurada en el subsistema de detección y anotación, así como la señal de referencia y de cada muestra para dichas posiciones. Para un intervalo de exploración, el flujo de bytes está formado por una sucesión de bloques de igual tamaño que depende del número de muestras. Cada bloque contiene información asociada a una posición cromosómica, preferentemente sucedidos es en orden creciente de coordenada cromosómica.

La figura 18A muestra un ejemplo no limitativo de codificación de un bloque, que el subsistema de detección (101) está configurado para seguir a la hora de guardar los resultados de su análisis en el subsistema contenedor de datos (102). En dicho ejemplo, aparecen codificados en binario y según la representación máquina de un entero de 2 bytes los siguientes parámetros:

- Los dígitos superiores (1801) de la coordenada cromosómica. Es decir, la parte entera del resultado de dividir dicha coordenada entre 10000.
- Los 5 dígitos menos significativos (1802) de dicha coordenada. Es decir, el resto de la división de la coordenada entre 10000.
- La señal de referencia (1803) según el modelo para dicha posición reescalada a la señal de la muestra de estudio para dicha posición y las señales reescaladas (1804, 1805, 1806) asociadas a dicha posición para cada una de las muestras consideradas en el estudio.

La señal reescalada (1804, 1805, 1806) para la muestra M_i , siendo M_R la muestra objeto de estudio se calcula según la fórmula:

$$\dddot{Cob}_i = floor\left(\frac{cobAutosomal(M_R)}{cobAutosomal(M_i)}\frac{\mathrm{sexCor(O,M_R)}}{\mathrm{sexCor(O,M_i)}}C_{m,p}\right)$$

donde la función *floor* dado un número proporciona el número entero inmediatamente inferior o igual al mismo; *cobAutosom(M)* representa el total de cobertura en los cromosomas autosómicos (1 a 22) en la región de interés para la muestra M; *sexCor(O, M)* para un cnv candidato O y una muestra M toma valor 2 en caso de que el cromosoma asociado al cnv candidato sea el X o el Y y el sexo vinculado a la muestra M sea “Mujer”, en

otro caso vale 1; $C_{m,p}$ representa la cobertura cruda para la muestra M en la posición cromosómica p.

La señal de referencia (1803) según el modelo rescalado se calcula conforme:

$$\dddot{Cob}_{REF} = RF_{s,p}.cobAutosomal(M_R).sexCor(O, M_R)$$

donde $RF_{s,p}$representa el valor de referencia para la posición p según el conjunto de control asignado a la muestra objeto de estudio.

Los bloques incluidos en el flujo de datos son los vinculados a coordenadas que abarcan las regiones afectadas por el cnv candidato y las dos regiones de los márgenes (de estar disponibles). Dependiendo de la resolución configurada para el visionado, se toman datos para todas las coordenadas mencionadas, o bien se toman coordenadas equiespaciadas según la resolución configurada para cada región abarcada empezando por la coordenada inicial de cada región.

Cada flujo de datos asociado a un cnv candidato lleva asimismo asociados metadatos con su tamaño en bytes, el valor máximo de señal que contiene y el número de coordenadas para las cuales se ha codificado la información. Todos los metadatos de un flujo de datos asociados a un cnv candidato se proveen directamente al subsistema de exploración, sin necesidad de examinar todos los datos del contenedor. Localizado un flujo de datos, el sistema de exploración solicitará bloques de datos, asociados a coordenadas concretas reclamadas por el subsistema de exploración (103), que se proveen mediante un acceso aleatorio, es decir sin necesidad de acceder al resto de bloques del flujo.

Cuando el subsistema de detección (101) está configurado para generar un resultado que integra el subsistema contenedor de datos (102) y el subsistema de exploración (103) en un único fichero, o bien cuando el subsistema contenedor de datos (102) no es un sistema de gestión (por ejemplo, una base de datos), sino un fichero a procesar por un subsistema de exploración (103) externo, los flujos de datos para cada cnv candidato se organizan dentro del fichero secuencialmente, como se ilustra en la figura 18B. A la sucesión de flujos de

datos, contenidos en un cuerpo (1813), le precede una sección de metadatos o cabecera (1812) y le sigue una sección de información de localización o cola (1814).

La cabecera (1812) se estructura en dos partes: una parte de tamaño fijo o sección invariable (1807) de 22 bytes de tamaño, y una sección de tamaño variable (1808) situada a continuación. La sección invariable (1807) de la cabecera codifica once campos en binario y con representación de entero sin signo de 2 bytes, mostrados con mayor detalle en la figura 18C. En particular, la sección invariable (1807) comprende la división entera (1815) del tamaño de datos entre 10000, el módulo de dicha división (1816), versión del formato de datos (1817), número de muestras (1818), resolución máxima (1819), factor de escala (1820), delimitador de lista (1821), delimitador de contenido de lista (1822), delimitador de item (1823), delimitador de atributo (1824), y delimitador de valor (1825). La división entera (1815) y el módulo de dicha división (1816) codifican el tamaño total de los datos. La versión del formato de datos (1817) permite identificar la estructura de los datos en caso de que se modifiquen o amplíen en versiones posteriores del método. La resolución máxima (1819) establece la mínima separación entre dos coordenadas recogidas en el flujo de datos. El factor de escala (1820) indica el factor por el que hay que multiplicar los valores de las señales codificadas en los datos para obtener las reales (con una posible pérdida de precisión si es superior a 1), permitiendo codificar valores que superan el máximo natural codificable con 2 bytes a costa de reducir la precisión. Los campos delimitador de lista (1821), delimitador de contenido de lista (1822), delimitador de item (1823), delimitador de atributo (1824), y delimitador de valor (1825) codifican los códigos ascii de los caracteres que se emplean para delimitar la información de metadatos contenida en la sección variable de la cabecera.

La sección de tamaño variable (1808) de la cabecera (1812) está compuesta por un conjunto de listas con un nombre y una serie de ítems. Cada item tiene asociados una serie de atributos y sus respectivos valores. Cada carácter de las listas se codifica en binario empleando su código ASCII y la representación máquina para un valor entero sin signo de 16 bits. A modo de ejemplo, una posible implementación de la cabecera comprende entre otras, una primera lista con información para señalar los metadatos referentes a la muestra de estudio, una segunda lista con metadatos relativos a las muestras incluidas en el análisis y una tercera lista con metadatos sobre los cnvs de la muestra de estudio.

La primera lista contiene un atributo que indica la posición que ocupa el item correspondiente a esta muestra en dicha primera lista. La segunda lista comprende un item por cada muestra contemplada en el análisis. Cada item de la segunda lista tiene atributos con sus correspondientes valores para registrar: identificador de la muestra, placa de secuenciación, lane e índice dentro de la placa, sexo, y conjunto de muestras de control para dicha muestra. La tercera lista contiene un item por cada cnv candidato asignado a la muestra de estudio. Cada item de la tercera lista tiene atributos con sus correspondientes valores para registrar: identificador único del cnv; locus y rango cromosómico asignado al cnv; gen y región asignados; tamaño (atributo compuesto: tamaño en pares de bases y tamaño en bases codificantes); desviación de la señal de la muestra de estudio en la región asignada al cnv respecto del modelo (atributo compuesto: valor de referencia y variación a lo largo del intervalo); ratio entre la señal de la muestra de estudio y la señal referencia (atributo compuesto: valor de referencia y variación a lo largo del intervalo); información sobre break points (atributo compuesto: localización del breakpoint con más profundidad y localización del par); snps en el intervalo asignado al cnv para cada muestra indicando muestra, coordenada y calidad; coordenadas cromosómicas de inicio y fin para regiones codificantes, regiones exónicas e intrónicas abarcadas por el rango asignado al cnv así como sus nombres; coordenadas de inicio de los rangos cromosómicos entre regiones de interés abarcados por el cnv candidato, no muestreados y a omitir en la representación para el cnv candidato por el subsistema de exploración, valores de filtrado asociados al cnv candidato y puntuación de credibilidad asignada; número de posiciones cromosómicas incluidas en el flujo de bytes asociadas al cnv candidato; valor máximo de señal asociada al cnv candidato, offset en bytes del comienzo del flujo de datos asociado al cnv candidato desde el inicio el cuerpo del bloque de datos.

Finalmente, la cola (1814) está formada por un primer campo de cola (1810) y un segundo campo de cola (1811) codificados en 4 bytes. El primer campo de cola (1810) contiene el tamaño en bytes de la cabecera (1812), mientras que el segundo campo de cola (1811) repite los dos primeros bloques de datos incluidos en la sección invariable (1807) de la cabecera. Es decir, el segundo campo de cola (1811) repite la información de la división entera (1815) y el módulo de dicha división (1816), codificando así el tamaño total de datos. La redundancia de este campo permite comprobar que el bloque de datos está bien formado.

La figura 19 muestra una posible implementación del procedimiento para la generación del flujo de datos descrito. Tras su inicialización (1901) **:**

1. Tomar (1902) las regiones afectadas por el cnv así como la región previa al evento y siguiente si está disponible. Ordenar dichas regiones en función de la posición inicial asociada a cada una de ellas.

2. Establecer (1903) un contador de bytes, de cobertura máxima y de total de puntos a 0.

3. Mientras queden regiones por procesar, tomar las posiciones vinculadas a la que está en curso ordenadas de menor a mayor, y comenzar su procesado empezando por la más baja (1904). Para cada posición de esta lista a una distancia de la primera múltiplo de la resolución máxima de exploración establecida y en cualquier caso considerando también la posición final:

   3.1. Calcular (1905) la división entera de la posición por 1000 exportando como entero el resultado a un almacenamiento temporal empleando una representación binaria de 2 bytes. Exportar el resto de la posición entre 10000 a un almacenamiento temporal empleando también una representación binaria de 2 bytes.

   3.2. Recuperar (1906) el conjunto de muestras de control para la muestra objeto de estudio y el valor de referencia asociado en el modelo para dicho conjunto. Calcular la cobertura de referencia $\dddot{Cob}_{REF}$ según:

$$\dddot{Cob}_{REF} = RF_{s,p}.cobAutosomal(M_R).sexCor(O, M_R)$$

   donde $RF_{s,p}$representa el valor de referencia para la posición p según el conjunto de control asignado a la muestra objeto de estudio.

   3.3. Actualizar (1907) el registro de bytes añadiendo 6 bytes extra.

   3.4. Para cada una de las muestras de estudio:

      3.4.1. Calcular (1908) la cobertura para cada muestra según:

$$\dddot{Cob}_i = floor\left(\frac{cobAutosomal(M_R)}{cobAutosomal(M_i)}\frac{\mathrm{sexCor(O, M_R)}}{\mathrm{sexCor(O, M_i)}}C_{m,p}\right)$$

      donde la función *floor* dado un número proporciona el número entero inmediatamente inferior o igual al mismo; cobAutosomal*(M)* representa el total de cobertura en los cromosomas autosómicos (1 a 22) en la región de interés para la muestra M; *sexCor(O, M)* para un cnv candidato O y una muestra M toma valor 2 en caso de que el cromosoma asociado al cnv candidato sea el X

o el Y y el sexo vinculado a la muestra M sea “Mujer”, en otro caso vale 1; $C_{m,p}$ representa la cobertura cruda en la posición p para la muestra m.

Exportar dicho valor como entero el resultado a un almacenamiento temporal empleando una representación binaria de 2 bytes.

3.4.2. Actualizar (1909) el registro de bytes incrementándolo en 2 unidades y el registro de puntos incrementándolo en 1 unidad.

3.4.3. Actualizar (1910) el registro de cobertura máxima al valor $\dddot{Cob}_i$ cuando este sea superior al valor almacenado.

4 Previamente a finalizar el procedimiento (1911), asociar al cnv candidato los metadatos de puntos totales exportados, valor máximo de cobertura y número de bytes del flujo de datos.

Finalmente, el subsistema de exploración (103) se encarga de acceder al subsistema contenedor de datos (102), recuperando los datos que han de mostrarse a través de la interfaz de usuario en cada momento. Dicho interfaz de usuario puede ser por ejemplo un explorador web al que se proveen en forma de gráficos los resultados de un proceso de detección. El interfaz de usuario presenta preferentemente cuatro áreas cuando se reportan cvn candidatos: cabecera, tabla de cnvs candidatos y controles, área de mapa de cnv y área de detalle. El área de detalle se compone a su vez preferentemente de tres secciones: sección de señales respuesta, sección de anotaciones y sección de ratios.

La cabecera muestra información relativa al estudio y muestra en particular, como por ejemplo un identificador de la muestra y del experimento indicando placa de secuenciación; línea e índice asociado a la muestra; el sexo de la muestra asignado por el detector; y datos vinculados a la precisión del resultado, como el grado de correlación con el conjunto de control, el tamaño de fragmento, ratios de variación generales, etc.

La tabla de cnvs candidatos muestra la lista de cnvs candidatos asociados al resultado que se está explorando. En cada columna se disponen los valores para un atributo o un conjunto relacionado (en adelante denominado valores compuestos), por ejemplo una medida y su tasa de variación). Los valores en una determinada fila se corresponden con los asociados a un cnv candidato. Los atributos (o conjuntos de atributos, en adelante atributos

compuestos), disponibles para su visualización, están determinados por la información contenida en el subsistema contenedor de datos (102) a consecuencia de la configuración establecida para el subsistema de detección (101).

Nótese que los atributos (simples o compuestos) mostrados pueden variar dependiendo de cada implementación y/o configuración particular. En un ejemplo, no limitativo, la tabla puede mostrar los atributos “confianza”, “locus”, “región”, “tamaño”, “desviación”, “ratio”, “alineamientos” y “filtros”. El atributo “confianza” es un valor numérico que expresa el grado de credibilidad en que el cnv candidato sea una variante estructural real, no un artefacto. “Locus” es un atributo compuesto de la posición cromosómica detectable de inicio del cnv candidato y la posición cromosómica detectable de fin del cnv candidato. “Región” es un atributo compuesto del nombre de gen afectado y la región o regiones afectadas (indicándolas según las etiquetas establecidas en la región de interés en la configuración del subsistema de detección y anotación). La afectación parcial de una región es suficiente para su inclusión en este valor. “Tamaño” es un atributo compuesto del tamaño detectable que abarca la variante en pares de bases (desde la posición inicial indicada en locus hasta la final) y el tamaño de región codificante abarcada (únicamente contabilizando en dicho intervalo las posiciones codificantes). “Desviación” indica el grado de divergencia de los valores de la señal respuesta del área que abarca el cnv candidato para la muestra de estudio frente al valor de referencia calculado a partir de las muestras control, y se compone de un valor de referencia para el intervalo y otro valor que indica variación a lo largo del mismo. “Ratio” indica la proporción de pérdida o ganancia en de la señal o en consecuencia del número de copias de la muestra de estudio para el locus asignado al cnv candidato respecto a los valores de referencia establecidos por el conjunto control, y se compone de un valor de referencia para todo el intervalo afectado y otro de variación a lo largo del mismo. “Alineamiento” es un atributo compuesto que informa el máximo número de break points encontrados las regiones cromosómicas próximas a los extremos del intervalo afectado por el cnv candidato y su localización cromosómica –para él cromosoma vinculado al cnv candidato- y el número máximo de break points con secuencias laterales compatibles y su localización cromosómica. “Filtros” indica el estado de una serie de filtros o comprobaciones.

El área de mapa muestra, para el cnv candidato seleccionado, la señal respuesta (por ejemplo, la cobertura) para la muestra de estudio. La señal respuesta se muestra para las

regiones afectadas por el cnv candidato y las dos regiones laterales, siempre que se disponga de dichos datos disponibles (la primera región de cada iteración del algoritmo de detección no tendrá otra previa, ni tiene una posterior tampoco la última de cada iteración). La señal se muestra en una gráfica bidimensional, el valor para el eje Y es el de la señal respuesta para la posición cromosómica con valor en el eje X. Los ejes muestran etiquetas (valores de referencia). El eje X se presenta discontinuo, omitiendo los rangos entre regiones de interés. De manera análoga, se presentan las señales respuesta normalizada para el resto de muestras procesadas junto con la de estudio, incluyendo las muestras asociadas a su conjunto de muestras de control. Las señales normalizadas de las muestras distintas de la de estudio se ajustan a la escala de la muestra de estudio. La muestra objeto de estudio está destacada visualmente del resto. Se muestra también destacada la señal de referencia a lo largo de las regiones presentadas, también ajustada a la muestra de estudio.

En el área de mapa se presentan además, esquemáticamente, las anotaciones vinculadas a las regiones visualizadas. Se sitúan en la correspondiente localización sobre el eje X, respetando las discontinuidades que corresponden. Asimismo, se resalta el área cromosómica afectada por el cnv candidato. De esta manera, el mapa del cnv candidato permite una visualización completa de su contexto. Sobre el mapa puede seleccionarse un intervalo cromosómico para restringir la visualización en el área de detalle a dicho intervalo, funcionando a modo de zoom.

La sección de señales repuesta presenta las señales respuesta de igual modo que la sección de mapa (incluida la de referencia) pero restringiendo el dominio del eje X al intervalo seleccionado en el mapa. En caso de no haberse realizado selección, se muestra en su totalidad. Para cada una de las muestras, se muestran (cuando dicha información está disponible en el contenedor de datos) las variantes genéticas presentadas. Para cada variante se refleja gráficamente su localización y la frecuencia alélica. Las variantes identificadas como de baja calidad también se destacan gráficamente. Se sitúan en la coordenada cromosómica más cercana a su localización, representada en la gráfica en el eje X (abscisas) y con valor en el eje Y (ordenadas) igual al que corresponda a la señal respuesta de la muestra para dicho punto, de esta manera las marcas de variantes se sitúan a lo largo de la señal respuesta de las muestras que las presentan quedando determinada su pertenencia.

En la sección de anotaciones se presenta gráficamente toda la información contextual y anotaciones aplicables al intervalo representado en el área de detalle. Esta información puede estar en el subsistema contenedor de datos (102), o bien puede cargarse a partir de un recurso externo.

En la sección de ratios se presentan los ratios de las señales respuesta normalizada de cada muestra frente a la señal de referencia. En esta sección también se destaca visualmente la muestra objeto de estudio.

El área de detalle muestra información emergente al pasar por encima de los distintos elementos.

Además de los visores descritos, el interfaz de usuario puede comprender algunos de los siguientes botones, siempre entendidos como ejemplos no limitativos:

- Botón toogle para ocultar la cabecera y mostrar únicamente en la tabla los atributos de puntuación y locus.
- Botón toogle para oculta todos los cnvs candidatos que no han superado algún filtro.
- Botón toogle para presentar las señales de todas las muestras estudiadas.
- Selector de panel para filtrar por genes según prioridades establecidas en un fichero de paneles.

El subsistema de exploración (103) puede estar embebido en el resultado, o bien estar integrado en un servidor. Si el subsistema de exploración (103) actúa como proveedor de datos además de subsistema de exploración (103), la codificación de los datos descrita para el subsistema contenedor de datos (102) soporta un acceso aleatorio tanto de los metadatos como las señales asociadas a cualquier coordenada cromosómica. Es decir, se puede acceder directamente al dato requerido sin necesidad de leer un conjunto más amplio.

Cuando el subsistema de exploración (103) carga un resultado y hay cnvs candidatos, se solicita el conjunto de metadatos asociados a dicho resultado. En primer lugar se lee la cabecera (1812), accediendo directamente a la localización. Si los datos están embebidos en un fichero resultado junto con el subsistema de exploración (103), la localización de la

cabecera (1812) se encuentra codificada en la cola (1814). Para un tamaño de fichero T_{total}, si el tamaño de la cabecera es $T_{cabecera}$ y el tamaño de datos codificados es T_{datos}, el proveedor de datos accede y devuelve los bytes en el rango [T_{total} -T_{datos} , T_{total} – T_{datos} + $T_{cabecera}$]. Una vez obtenidos los metadatos, se solicita la información a graficar en las áreas de mapa y de detalle, valores asociados a un número de puntos sobre el eje X que depende de la resolución gráfica que se haya configurado para el subsistema de exploración. La información correspondiente a cada punto se encuentra codificada en un bloque del flujo de datos correspondiente al cnv candidato seleccionado en la tabla de candidatos. Al ser el número de puntos a solicitar constante para cada cnv candidato, el tamaño de la información relativa a cada punto también es constante. Por lo tanto, el acceso es aleatorio y el tiempo requerido para graficar la información no depende ni del tamaño de los datos ni de los puntos concretos seleccionados.

Inicialmente, los puntos graficados en el área de detalle se corresponden con los graficados en el área de mapa por lo que no es necesario realizar una petición adicional para obtenerlos. Una vez obtenida la información, el subsistema de exploración (103) grafica la información correspondiente en el área de detalle y mapa (ya sea secuencialmente en orden indistinto o de forma paralela). Para un cnv candidato, la lista *P* de puntos a solicitar para visualizar un intervalo gráfico [p_{ini} , p_{fin}] donde p_{ini} y p_{fin} se corresponden con los bloques correspondientes a dos posiciones en el flujo de datos se calcula según el procedimiento:

1. Incluir p_{ini} en P
2. Calcular intervalo de muestreo de puntos como *int* = (p_{fin} – p_{ini} + 1)/Cfg.res donde Cfg.res es el valor de configuración del número de puntos a graficar o resolución.
3. Incluir los puntos p_{i*n} i=0..(Cfg.res-1)
4. Incluir p_{fin} en caso de no haya sido incluido previamente.

Una vez solicitados por el subsistema de exploración (103) los puntos incluidos en P = {p_1 .. p_n}, el proveedor de datos accederá a los bytes correspondientes según el procedimiento:

1. Leer o calcular el tamaño del bloque de datos asociado a un bloque de datos en bytes. *block* = 6 + (número de muestras * 2), donde el número de muestras está disponible en la cabecera del bloque de datos. Dependiendo de la arquitectura de despliegue, habrá sido procesado durante la carga de metadatos o será un valor conocido.

2. Añadir al buffer de bytes solicitados los bytes (o los valores que codifican) en los rangos de bytes del flujo de datos correspondiente: [i*block,(i+1)*block] para i = 0…(|P|-1)

3. Devolver los bytes o los valores que codifican incluidos en el buffer al subsistema de exploración (103) según sea el despliegue escogido.

Para la solicitud de puntos de mapa, p_{ini} y p_{fin} se corresponden con 1 y el número de puntos asociado al flujo de datos (disponible en los metadatos ya leídos) respectivamente, es decir el primer y último punto o bloque del flujo de datos.

Cuando el usuario seleccione un nuevo cnv en la tabla, el subsistema de exploración (103) solicita directamente los puntos de mapa, puesto que ya ha leído los metadatos previamente durante la carga del resultado. Una vez obtenida la información asociada a los puntos, procede a repintar las áreas de mapa y de detalle.

Cuando en el área de mapa el usuario selecciona una región para hacer zoom, el subsistema de exploración (103) solicita los puntos a graficar (calculados según el procedimiento ya descrito) para el intervalo [p_{zoom_ini} , p_{zoom_fin}]. p_{zoom_ini} será el más punto (punto representado en el área de mapa, dominio del eje X) de mapa más próximo al inicio del intervalo seleccionado y p_{zoom_fin} el punto de mapa más próximo al final del intervalo seleccionado. Una vez obtenida la información para los puntos solicitados el subsistema de exploración (103), se procede a repintar el área de detalle, mostrando una marca en el mapa de la región sobre la que se está haciendo el zoom.

En despliegues que no integran el subsistema de exploración (103) y los propios datos en un mismo fichero (por ejemplo, cuando el proveedor de datos sea un sistema de gestión de bases de datos), el acceso aleatorio a los metadatos y los flujos de bytes será gestionado por dicho proveedor de datos. Igualmente la estructura del flujo de bytes del subsistema contenedor de datos (102) permite un acceso aleatorio a los bytes requeridos según los puntos del flujo demandados por el subsistema de exploración (103). En los despliegues donde el subsistema de exploración (103) y los datos están integrados en un único fichero, el paquete de datos se sitúa como último elemento del fichero, mientras que a lógica del subsistema de exploración (103) se sitúa al inicio.

Una posible implementación del despliegue en el que se integran el subsistema de exploración (103) y los propios datos en un mismo fichero consiste en que el subsistema de detección (101) genera un fichero que implementa en HTML/javascript el subsistema de exploración (103). Para que el resultado pueda ser cargado rápidamente, y los datos no se carguen sino bajo demanda en el navegador web, se codifica toda la lógica para proveer gráficos y datos como una función asociada a un timeout, y se establece antes del bloque de datos una instrucción que aborta la carga e interpretación del resto del fichero. Esta solución permite que toda la lógica del subsistema de exploración (103) y del subsistema contenedor de datos (102) siga activa en el navegador, accediéndose, de manera aleatoria, a los datos requeridos del total del fichero, en función de la interacción del usuario. Esta estrategia minimiza el consumo de memoria y aumenta la eficiencia del proceso, pudiendo trabajar con ficheros pesados con millones de puntos de manera ágil. Por ejemplo, el acceso aleatorio a un rango de bytes es soportado en HTML5 mediante el uso de la función *slide*.

La representación de rangos discontinuos de coordenadas en los los gráficos presentados para las áreas de detalle y zoom es nativa y transparente al subsistema de exploración (103). Obtenidos los bloques a representar, el subsistema de exploración (103) asigna en el eje X valores naturales correlativos, empezando en 1, para cada uno de los bloques según su orden relativo en la secuencia del flujo de datos. El dominio del eje X siempre es [1..n] donde n será el número de puntos pedidos según el algoritmo descrito para un rango de selección. Las etiquetas en el eje X no serán los valores en X sino las coordenadas cromosómicas asociadas. Para una posición del eje X la coordenada cromosómica asociada se resuelve interpretando el valor de los dos primeros bytes del bloque de datos asociado al punto correspondiente, pues estos contienen codificada la coordenada cromosómica. Los valores en el eje y para cada señal se encuentran codificados en el byte correspondiente del bloque de datos asociado a un punto.

Para dividir las gráficas en secciones cromosómicas continuas, se pintan separadores sobre dichas secciones. Entre los metadatos asociados a cada cnv candidato se encuentran las localizaciones en coordenadas cromosómicas de distintos elementos incluidos en el rango cromosómico de visualización vinculado al cnv candidato. Como se ha indicado, se establece de forma transparente un mapeo entre rangos discontinuos de coordenadas cromosómicas y valores continuos en el eje X de las áreas gráficas, basado en la relación de orden de los bloques de datos vinculados a las posiciones cromosómicas mostradas. A

cada coordenada cromosómica se le asigna un valor en el eje X siguiendo el siguiente procedimiento:

1. Tomar la lista bloques de datos correspondientes a los puntos solicitados (ordenados según su localización en el flujo de datos original) P = {P_1...P_n}, o las listas de valores ya interpretados correspondientes (los que contiene codifica cada bloque).

2. Establecer como coordenada cromosómica de inicio de un rango de traducción la codificada en el bloque de datos del primer punto de la lista P_1 y como coordenada final la codificada para P_2.

3. Si la coordenada cromosómica a traducir se encuentra en el rango de traducción (es mayor o igual que la coordenada de inicial e inferior a la final) devolver 1 como valor correspondiente en X para la coordenada cromosómica.

4. En caso de que la coordenada cromosómica no esté incluida en el rango de traducción y el punto usado para establecer el final del rango de exploración no sea el último de la lista P, se repiten los pasos 2 y 3, tomando como punto inicial el punto final usado en la repetición previa y como punto final el siguiente de la lista.

5. Si la coordenada cromosómica se corresponde con el valor de coordenada cromosómica codificada para P_n el valor en X es n para dicha coordenada.

6. Si la coordenada cromosómica a traducir supera el valor cromosómico codificado en el bloque de datos vinculado a P_n , o es inferior a la coordenada cromosómica codificada para P_1 se sale del dominio de X por lo que se generará una excepción a procesar. En caso de que la coordenada se esté mapeando como parte de un rango cromosómico donde una de las coordenadas está dentro del dominio en X, se asigna el valor en X del punto que posea codificada la coordenada más próxima a la que se desea traducir, 1 cuando la coordenada cromosómica es inferior a la codificada para P_1 , y n cuando la coordenada cromosómica es superior a la codificada para P_n.

Empleando el procedimiento descrito se mapean las coordenadas en eje X que corresponden a las coordenadas cromosómicas de las variantes vinculadas a cada flujo de datos de cnv candidato de cara a su graficación. De la misma manera se procede para situar y dibujar la localización de exones, intrones y otras regiones en el área de mapa y de

anotaciones. Para el cálculo de los valores en Y de la sección de ratios, los valores de cada señal se dividen por el valor de referencia, contenido para cada bloque de datos en el 4 y 5 byte.

A la vista de esta descripción y figuras, el experto en la materia podrá entender que la invención ha sido descrita según algunas realizaciones preferentes de la misma, pero que múltiples variaciones pueden ser introducidas en dichas realizaciones preferentes, sin salir del objeto de la invención tal y como ha sido reivindicada.

REIVINDICACIONES

1. Método de detección de variantes genéticas estructurales a partir de datos de secuenciación de una pluralidad de muestras caracterizada por que comprende, para cada muestra:
   - caracterizar (230) cada muestra determinando un género en función de, al menos, una cobertura cromosómica diferencial del cromosoma X;
   - calcular al menos una matriz de correlaciones asociada a la covariabilidad experimental de la pluralidad de muestras, a partir de los datos de secuenciación de dichas muestras;
   - seleccionar (406) al menos una estructura de control mediante división iterativa en clústeres (404) de la pluralidad de muestras;
   - establecer (213) unos puntos de trabajo en función de unas variaciones respecto a un valor de referencia de las estructuras de control; y
   - detectar (240) variantes genéticas estructurales en los puntos de trabajo determinados.
2. Método de acuerdo con la reivindicación 1 caracterizado porque los datos de secuenciación comprenden lecturas de nuevas tecnologías de secuenciación.
3. Método de acuerdo con cualquiera de las reivindicaciones anteriores caracterizado por que el paso de determinar el género de la muestra comprende evaluar una medida relativa de cobertura entre el cromosoma X y los cromosomas autosómicos.
4. Método de acuerdo con cualquiera de las reivindicaciones anteriores caracterizado por que el paso de calcular al menos una matriz de correlaciones comprende calcular (403) una primera matriz de correlaciones de las muestras en base a al menos una variables seleccionada de entre el perfil de cobertura, un rango de posiciones, un número de lecturas mapeadas y una métrica derivada de una región cromosómica.
5. Método de acuerdo con cualquiera de las reivindicaciones anteriores caracterizado por que el paso de calcular al menos una matriz de correlaciones comprende calcular (404) una segunda matriz de correlaciones de las muestras en base a al menos una variables seleccionada de entre el tamaño de los fragmentos secuenciados incluyendo adaptadores, el tamaño de los fragmentos secuenciados sin incluir adaptadores, y una separación de lecturas en lecturas pareja-final.
6. Método de acuerdo con cualquiera de las reivindicaciones anteriores caracterizado por que la clusterización aplica un algoritmo kmeans.

7. Método de acuerdo con cualquiera de las reivindicaciones anteriores caracterizado por que comprende normalizar (703) los datos de la pluralidad de muestras en función del género determinado y eliminar sesgos de enriquecimiento.

8. Método de acuerdo con cualquiera de las reivindicaciones 1 a 7 caracterizado por que comprende normalizar (703) los datos de la pluralidad de muestras en función de un valor de lecturas alineadas, siendo dicho valor asignado a un rango de posiciones.

9. Método de acuerdo con cualquiera de las reivindicaciones anteriores caracterizado por que el valor de referencia del paso de establecer (213) los puntos de trabajo se mide sobre una variable seleccionada de entre cobertura, contaje de lecturas y métricas derivadas.

10. Método de acuerdo con cualquiera de las reivindicaciones anteriores caracterizado por que el valor de referencia del paso de establecer (213) los puntos de trabajo se calcula mediante una métrica seleccionada de entre la media, la mediana y otras medidas de tendencia central; y la variación se calcula mediante una métrica seleccionada de entre el rango intercuartílico, la varianza, la desviación típica y otras medidas de variación.

11. Método de acuerdo con cualquiera de las reivindicaciones anteriores caracterizado por que comprende aplicar un cribado en función de una retención de ratios, información de variantes, zonas codificantes y paneles.

12. Método de acuerdo con cualquiera de las reivindicaciones anteriores caracterizado por que el paso de detectar (240) variantes genéticas estructurales comprende asignar (1007) un valor indicativo de un grado confianza a cada variante estructural detectada, en función de, al menos, la desviación respecto a la estructura de control y la región cromosómica donde se sitúa variantes genéticas estructurales.

13. Método de acuerdo con cualquiera de las reivindicaciones anteriores caracterizado por que comprende agrupar múltiples variantes estructurales detectadas en una misma variante en función de, al menos, la desviación respecto a la estructura de control y la región cromosómica donde se sitúa variantes genéticas estructurales.

14. Método de acuerdo con cualquiera de las reivindicaciones anteriores caracterizado por que comprende además, almacenar los resultados de la detección de variantes genéticas estructurales en unos medios de almacenamiento de datos (102) de acuerdo con una codificación que comprende:

    - una cabecera (1812) con una sección de tamaño invariable (1807) que comprende información del tamaño de datos y una sección de tamaño variable (1808) que comprende información de señalización de metadatos;

- un cuerpo (1813) que comprende bloques de igual tamaño con información de coordenada cromosómica, señal de referencia (1803) reescalada y señales reescaladas (1804, 1805, 1806) asociadas a cada muestra; y
- una cola (1814) que comprende información de localización.

15. Método de acuerdo con la reivindicación 14 caracterizado por que comprende además, acceder desde unos medios de exploración (103) a los resultados almacenados en los medios de almacenamiento de datos (102) mediante un acceso aleatorio que comprende:
    - acceder a la cabecera (1812);
    - obtener metadatos de los resultados; y
    - obtener información asociada a cada punto a representar accediendo al cuerpo (1813) y recuperando bloques de información de tamaño constante.
16. Sistema de detección de variantes genéticas estructurales a partir de datos de secuenciación de una pluralidad de muestras que comprende, medios de detección (101), medios de almacenamiento de datos (102) y medios de exploración (103) de los datos almacenados, caracterizado por que los medios de detección están configurados para implementar los pasos del método de acuerdo con cualquiera de las reivindicaciones 1 a 15.
17. Programa de ordenador que comprende medios de código de programa de ordenador adaptados para realizar las etapas del método de cualquiera de las reivindicaciones 1 a 15, cuando el mencionado programa se ejecuta en un procesador digital de la señal, un circuito integrado específico de la aplicación, un microprocesador, un microcontrolador o cualquier otra forma de hardware programable.

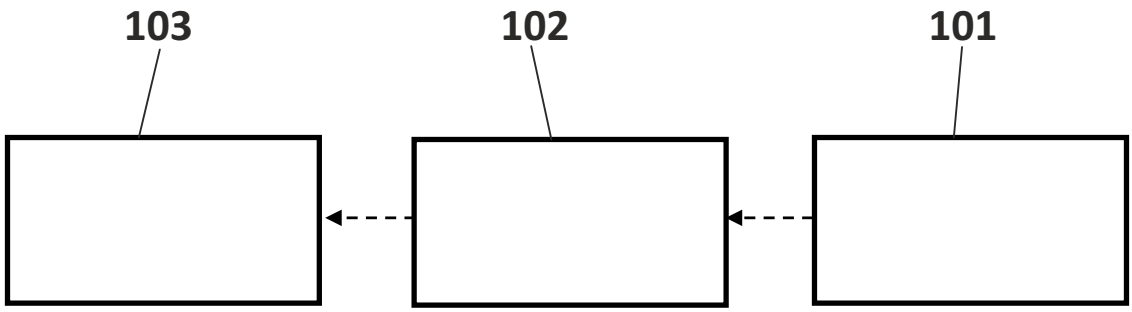

FIG. 1A

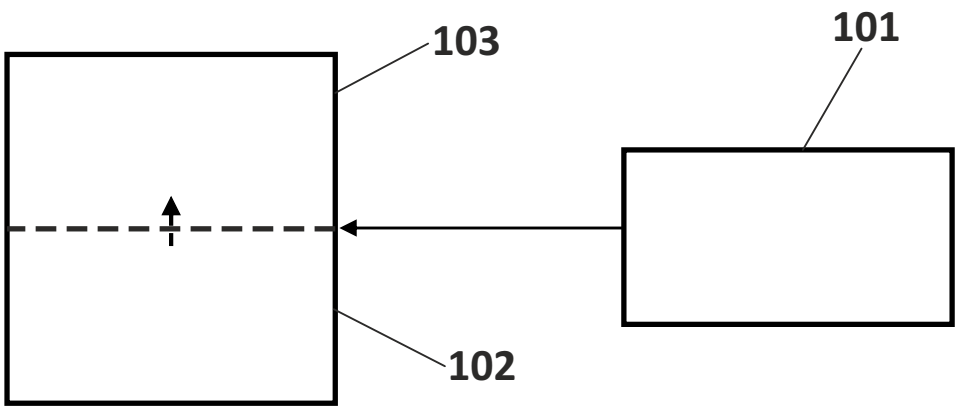

FIG. 1B

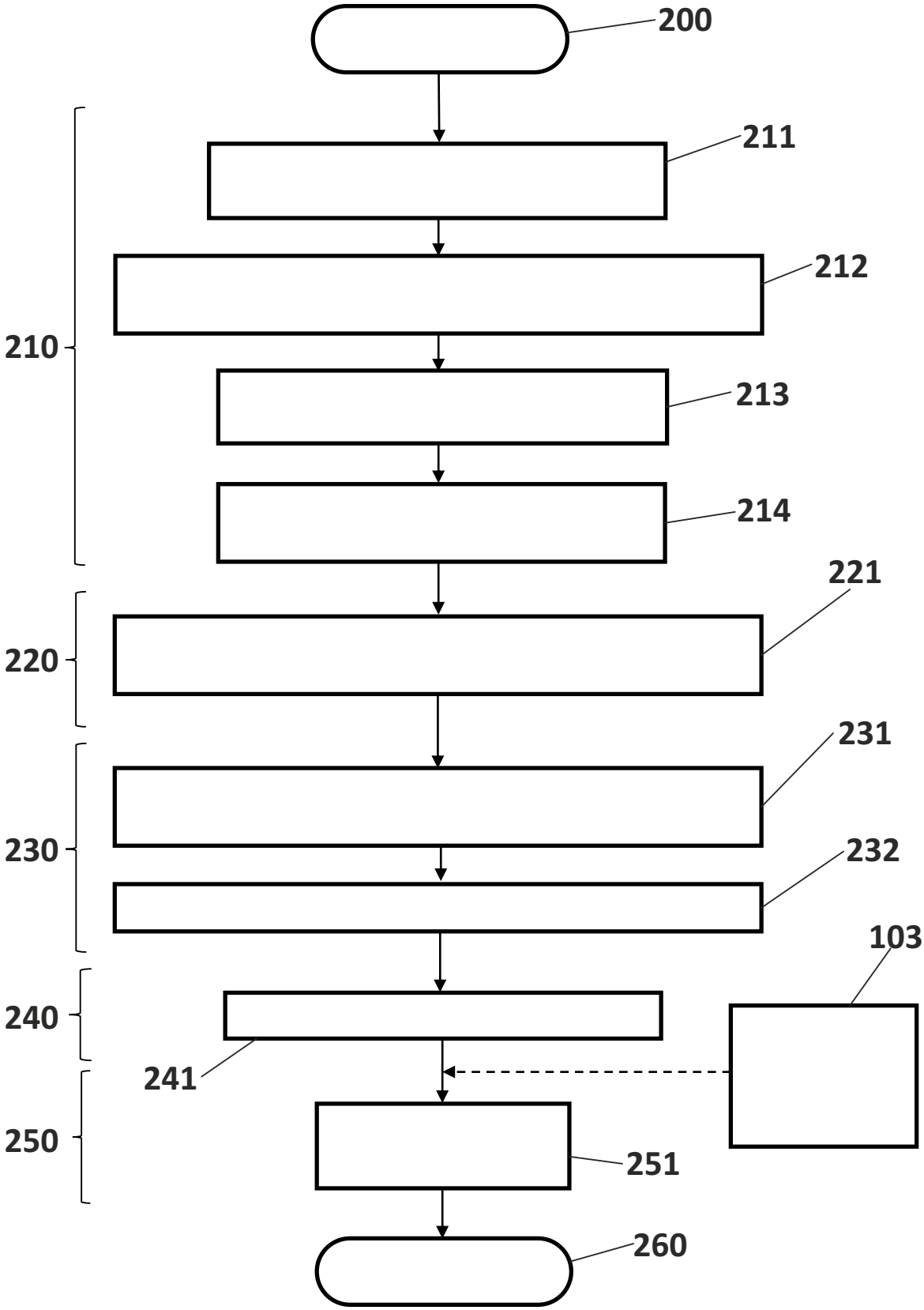
200
211
212
210
213
214
221
220
231
230
232
103
240
241
250
251
260

FIG. 2

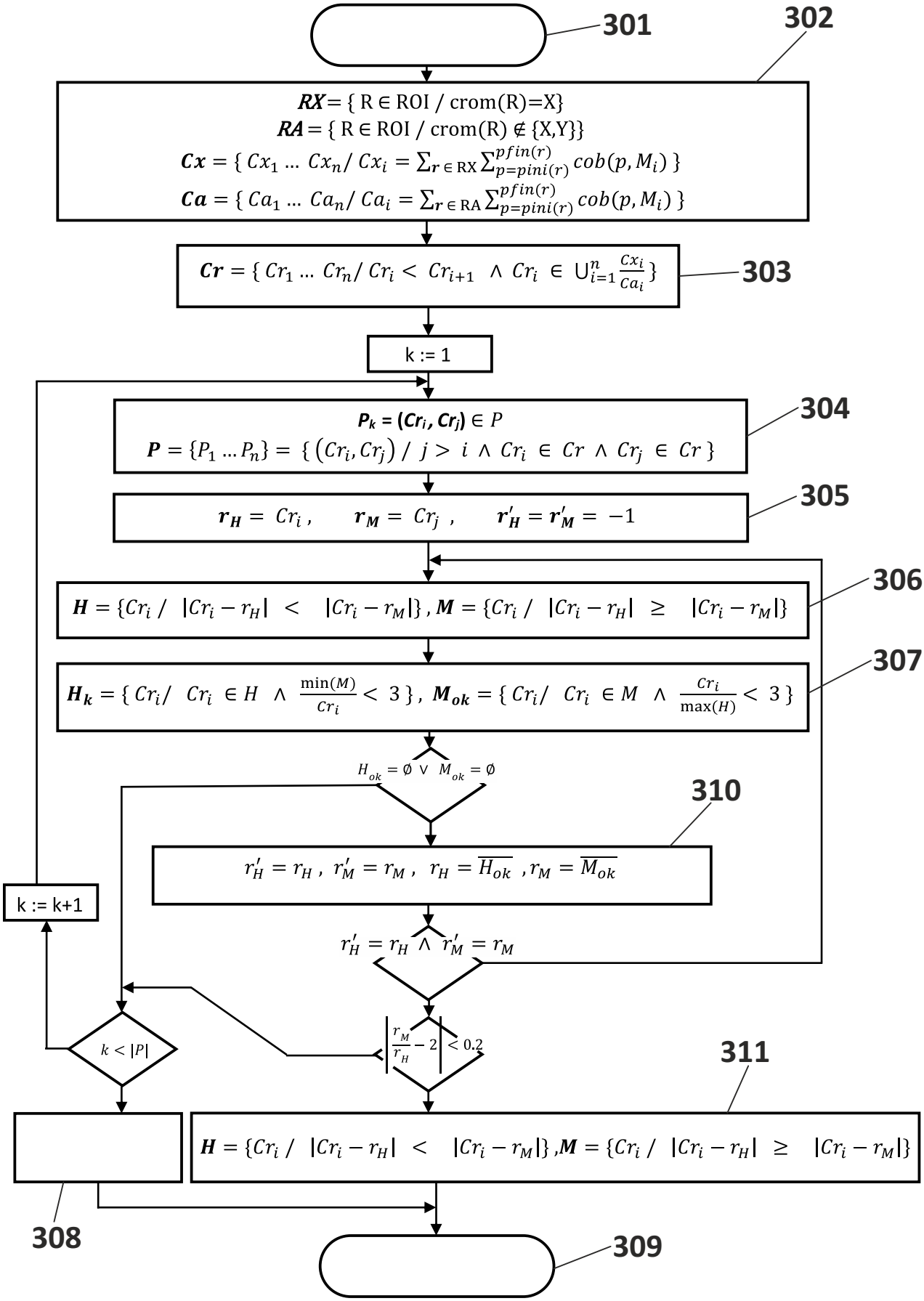
301
302
$\boldsymbol{RX}$= { R ∈ ROI / crom(R)=X}
$\boldsymbol{RA}$ = { R ∈ ROI / crom(R) ∉ {X,Y}}
$\boldsymbol{Cx} = \{ Cx_1 \dots Cx_n / Cx_i = \sum_{\boldsymbol{r} \in RX} \sum_{p=pini(r)}^{pfin(r)} cob(p, M_i) \}$
$\boldsymbol{Ca} = \{ Ca_1 \dots Ca_n / Ca_i = \sum_{\boldsymbol{r} \in RA} \sum_{p=pini(r)}^{pfin(r)} cob(p, M_i) \}$
$\boldsymbol{Cr} = \{ Cr_1 \dots Cr_n / Cr_i < Cr_{i+1} \wedge Cr_i \in \cup_{i=1}^{n} \frac{Cx_i}{Ca_i} \}$
303
k := 1
$\boldsymbol{P_k} = (\boldsymbol{Cr_i}, \boldsymbol{Cr_j}) \in P$
$\boldsymbol{P} = \{P_1 \dots P_n\} = \{ (Cr_i, Cr_j) / j > i \wedge Cr_i \in Cr \wedge Cr_j \in Cr \}$
304
$\boldsymbol{r_H} = Cr_i$, $\boldsymbol{r_M} = Cr_j$, $\boldsymbol{r'_H} = \boldsymbol{r'_M} = -1$
305
$\boldsymbol{H} = \{Cr_i / |Cr_i - r_H| < |Cr_i - r_M|\}$, $\boldsymbol{M} = \{Cr_i / |Cr_i - r_H| \geq |Cr_i - r_M|\}$
306
$\boldsymbol{H_k} = \{ Cr_i / Cr_i \in H \wedge \frac{\min(M)}{Cr_i} < 3 \}$, $\boldsymbol{M_{ok}} = \{ Cr_i / Cr_i \in M \wedge \frac{Cr_i}{\max(H)} < 3 \}$
307
$H_{ok} = \emptyset \vee M_{ok} = \emptyset$
310
$r'_H = r_H$, $r'_M = r_M$, $r_H = \overline{H_{ok}}$, $r_M = \overline{M_{ok}}$
k := k+1
$r'_H = r_H \wedge r'_M = r_M$
$k < |P|$
$\left| \frac{r_M}{r_H} - 2 \right| < 0.2$
311
$\boldsymbol{H} = \{Cr_i / |Cr_i - r_H| < |Cr_i - r_M|\}$,$\boldsymbol{M} = \{Cr_i / |Cr_i - r_H| \geq |Cr_i - r_M|\}$
308
309

FIG. 3

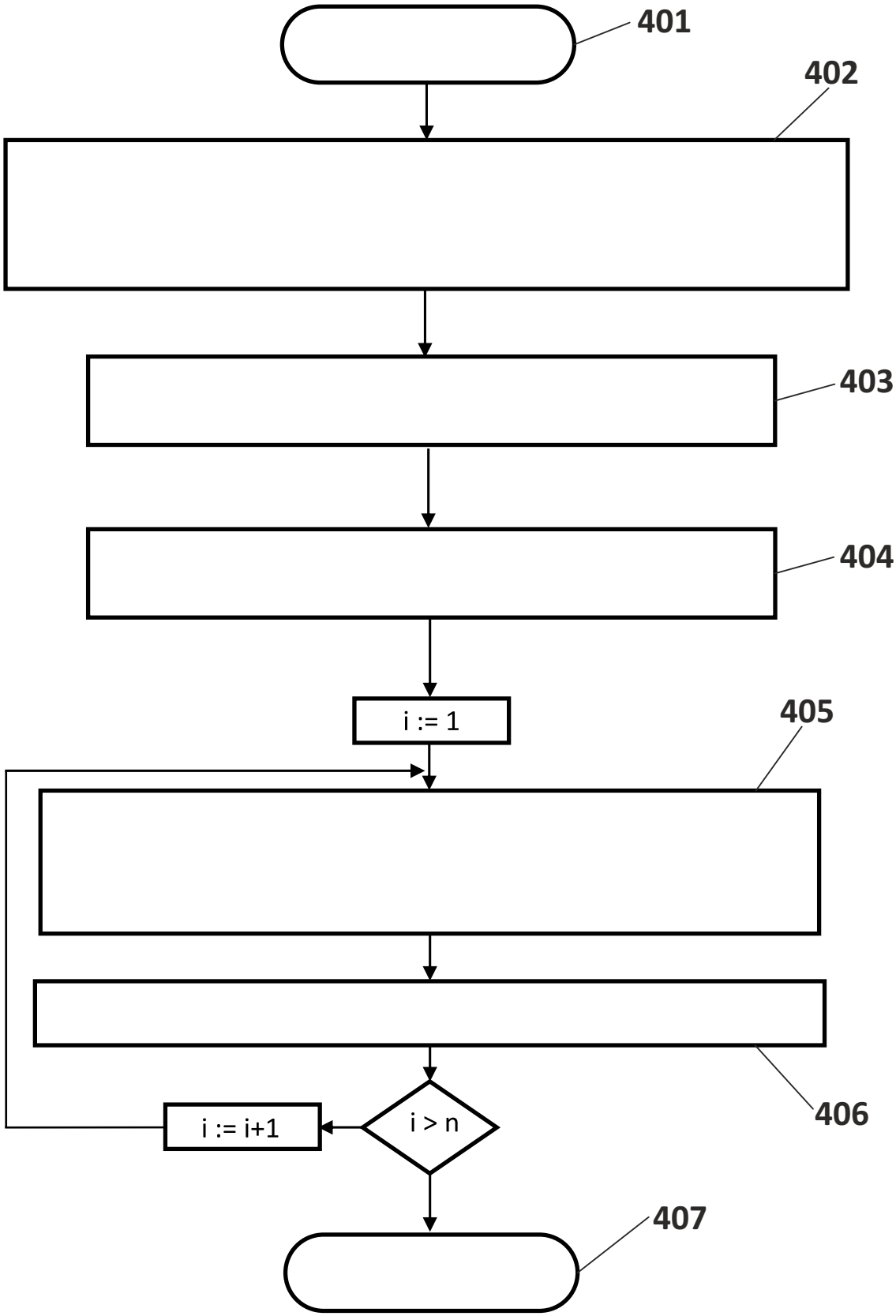

FIG. 4

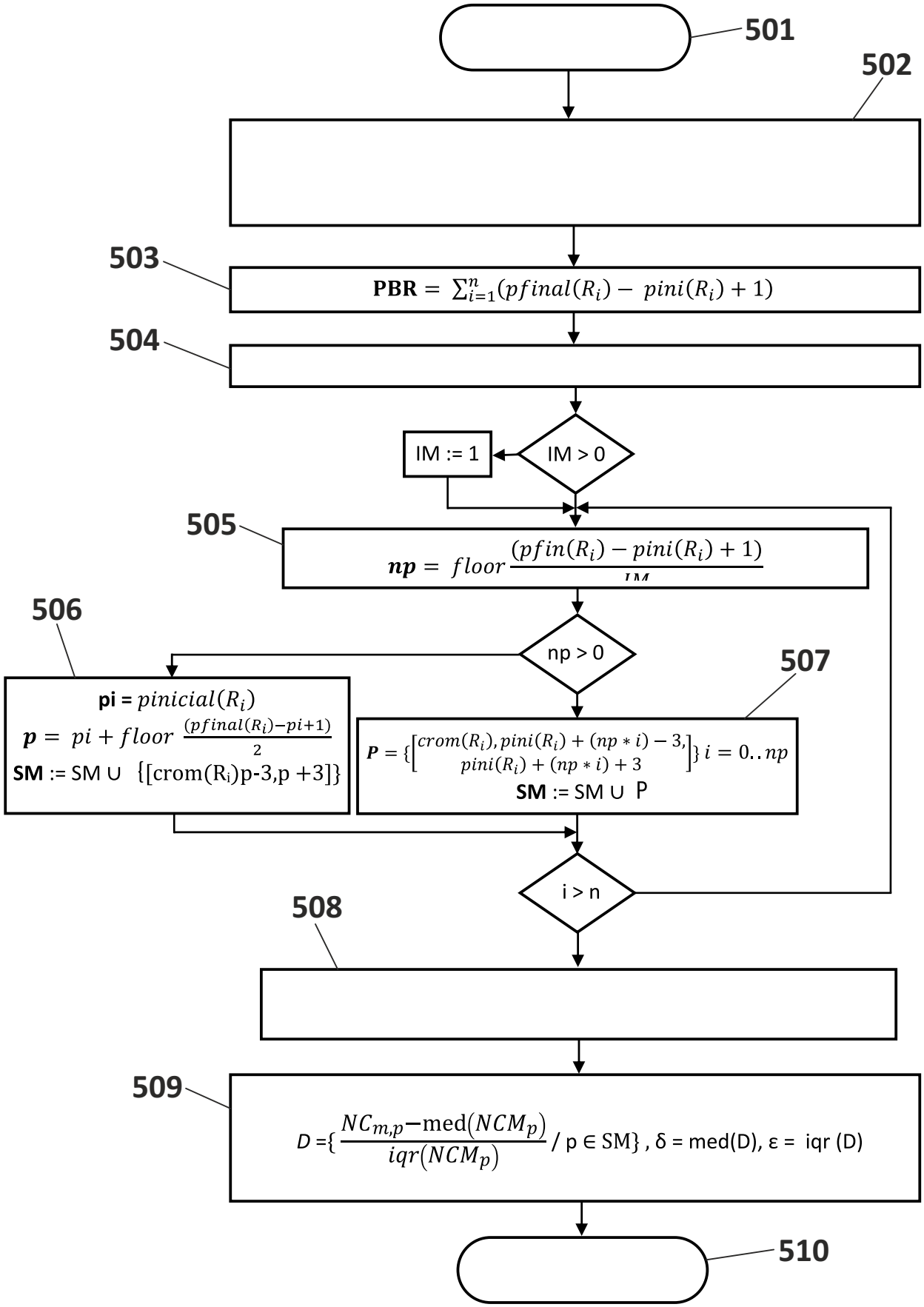
501
502
503
PBR = $\sum_{i=1}^{n}(pfinal(R_i) - pini(R_i) + 1)$
504
IM := 1
IM > 0
505
$np = floor \frac{(pfin(R_i) - pini(R_i) + 1)}{IM}$
506
np > 0
507
pi = $pinicial(R_i)$
$p = pi + floor \frac{(pfinal(R_i) - pi + 1)}{2}$
SM := SM ∪ {[crom(Ri)p-3,p +3]}
$P = \{\begin{bmatrix} crom(R_i), pini(R_i) + (np * i) - 3, \\ pini(R_i) + (np * i) + 3 \end{bmatrix}\} i = 0..np$
SM := SM ∪ P
i > n
508
509
$D = \{\frac{NC_{m,p} - \text{med}(NCM_p)}{iqr(NCM_p)}$ / p ∈ SM} , δ = med(D), ε = iqr (D)
510

FIG. 5

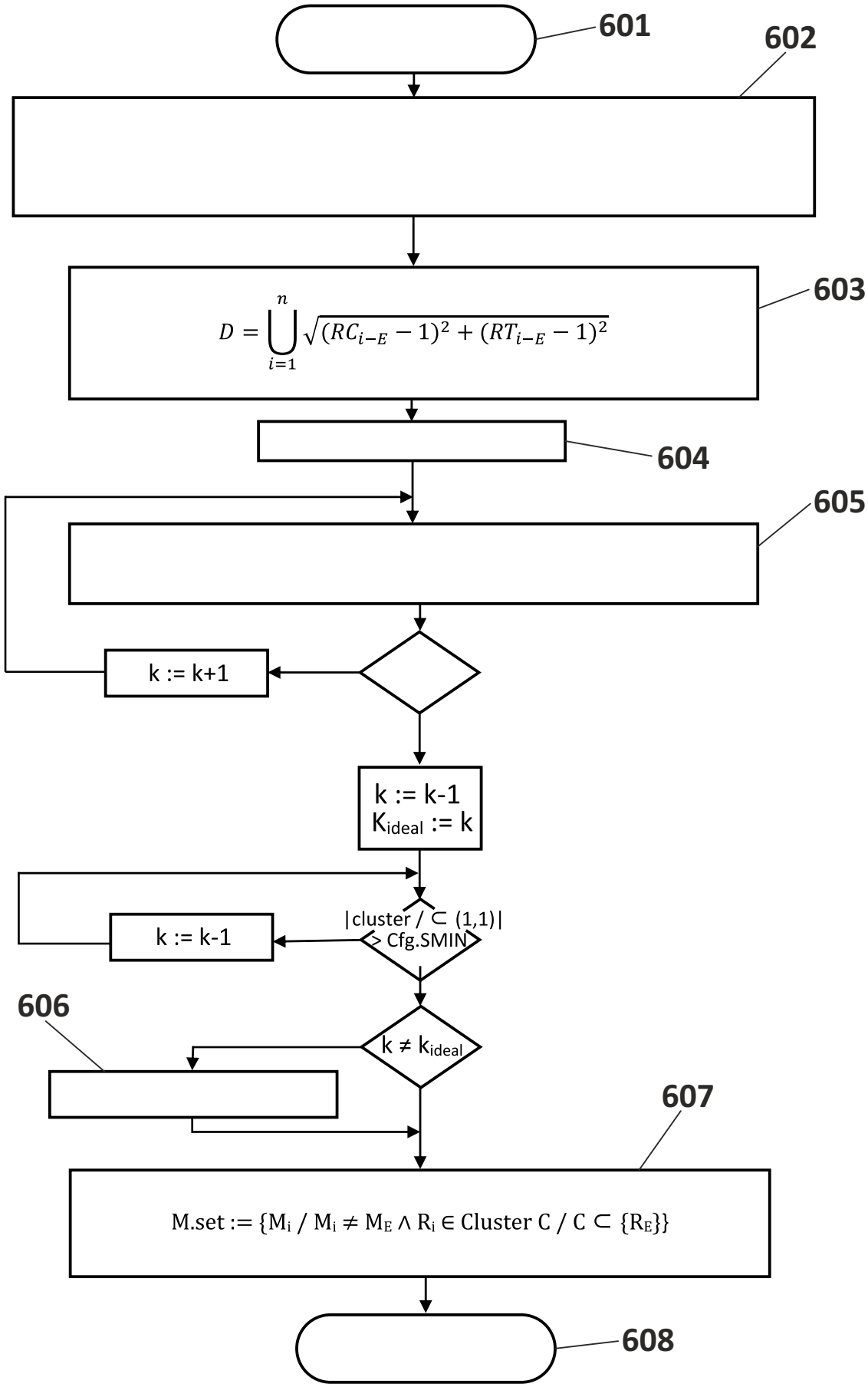
601
602
$D = \bigcup_{i=1}^{n} \sqrt{(RC_{i-E} - 1)^2 + (RT_{i-E} - 1)^2}$
603
604
605
k := k+1
k := k-1
K_ideal := k
|cluster / ⊂ (1,1)|
> Cfg.SMIN
k := k-1
606
k ≠ k_ideal
607
M.set := {M_i / M_i ≠ M_E ∧ R_i ∈ Cluster C / C ⊂ {R_E}}
608

FIG. 6

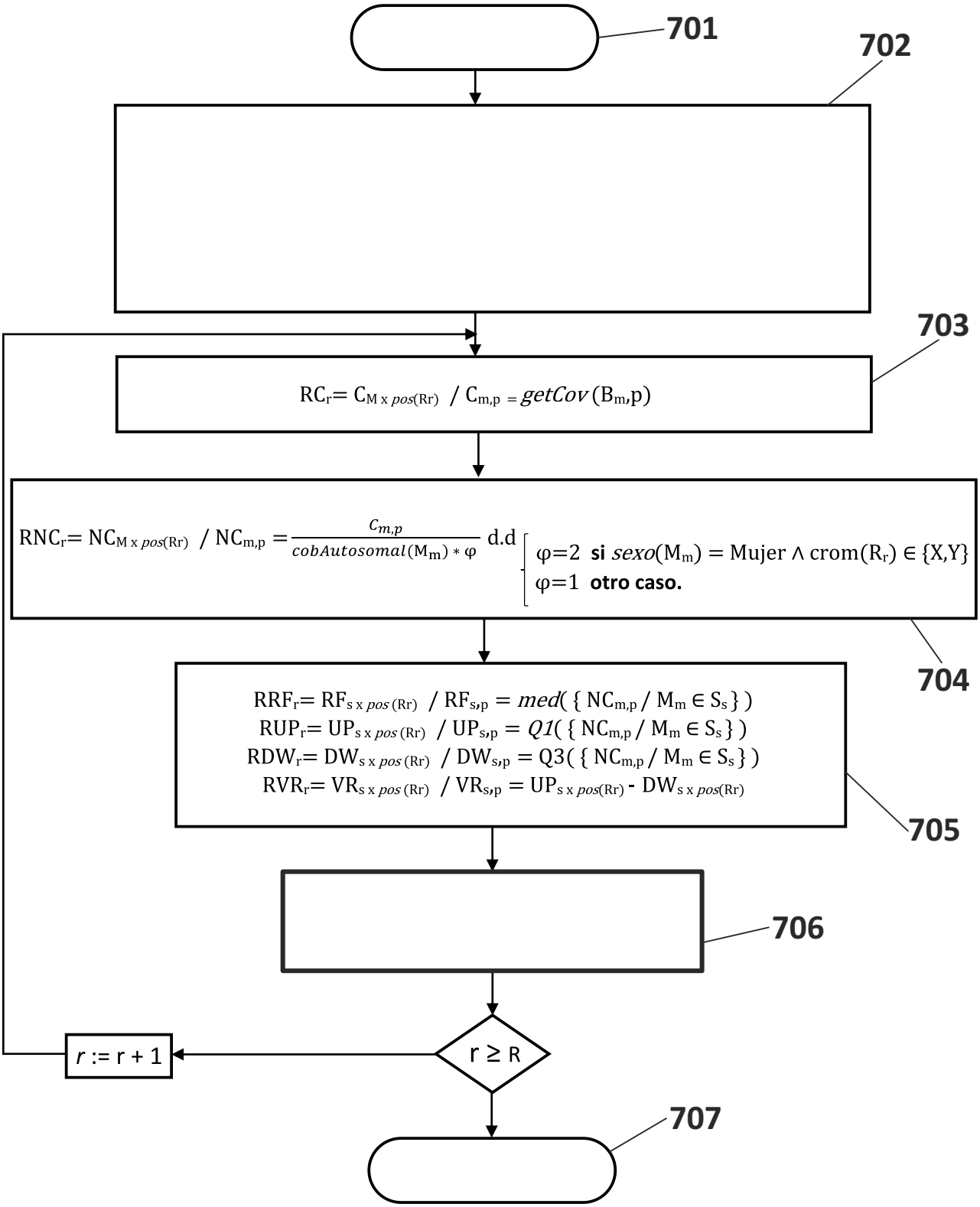
701
702
703
$RC_r = C_{M \times pos(Rr)} / C_{m,p} = getCov(B_m, p)$
$RNC_r = NC_{M \times pos(Rr)} / NC_{m,p} = \frac{C_{m,p}}{cobAutosomal(M_m) * \varphi}$ d.d
$\varphi=2$ si $sexo(M_m) = Mujer \wedge crom(R_r) \in \{X,Y\}$
$\varphi=1$ otro caso.
704
$RRF_r = RF_{s \times pos(Rr)} / RF_{s,p} = med(\{ NC_{m,p} / M_m \in S_s \})$
$RUP_r = UP_{s \times pos(Rr)} / UP_{s,p} = Q1(\{ NC_{m,p} / M_m \in S_s \})$
$RDW_r = DW_{s \times pos(Rr)} / DW_{s,p} = Q3(\{ NC_{m,p} / M_m \in S_s \})$
$RVR_r = VR_{s \times pos(Rr)} / VR_{s,p} = UP_{s \times pos(Rr)} - DW_{s \times pos(Rr)}$
705
706
r ≥ R
r := r + 1
707

FIG. 7

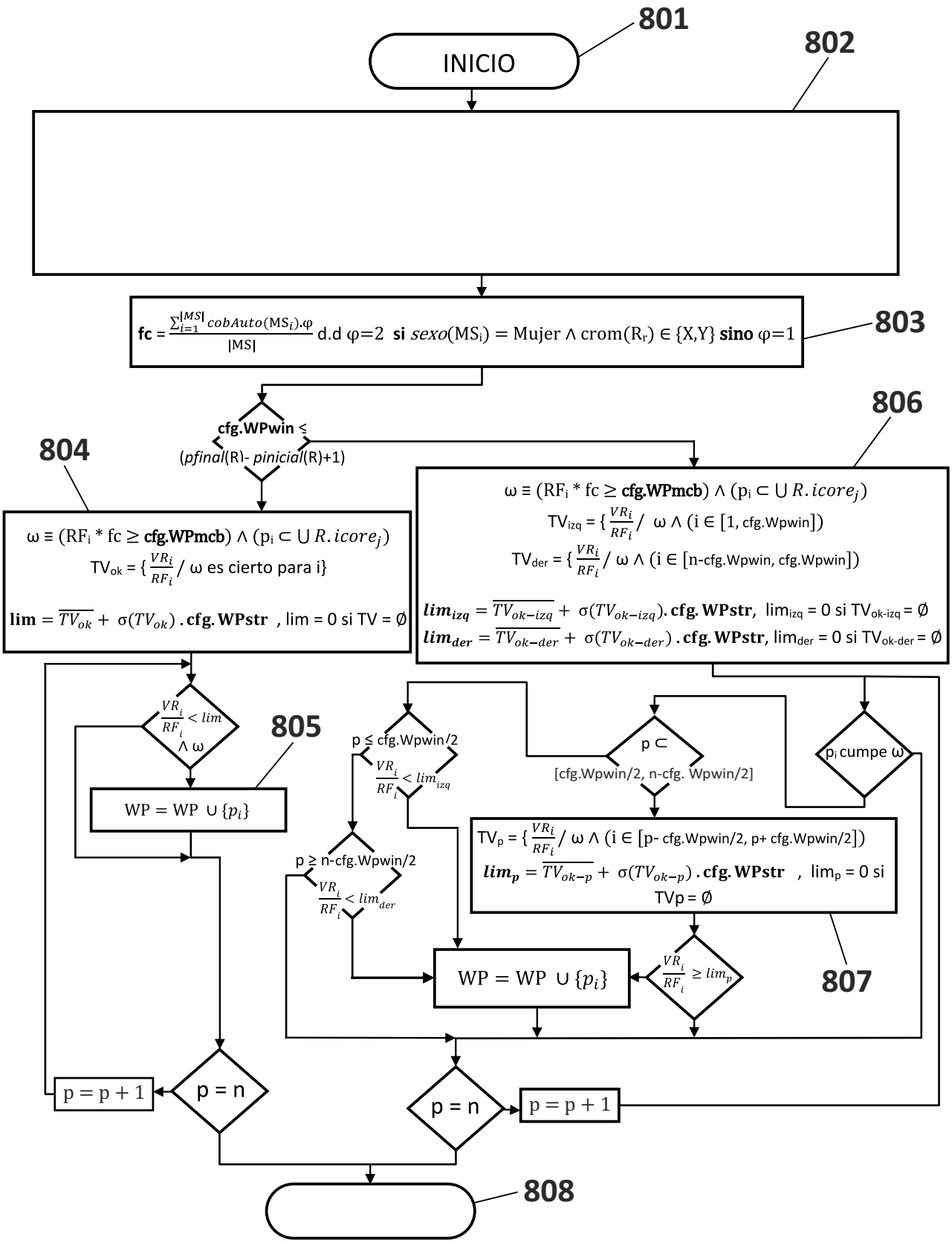
801
INICIO
802
fc = $\frac{\sum_{i=1}^{|MS|} cobAuto(MS_i).\varphi}{|MS|}$ d.d φ=2 si $sexo(MS_i) = Mujer \wedge crom(R_r) \in \{X,Y\}$ sino φ=1
803
cfg.WPwin ≤
(pfinal(R)- pinicial(R)+1)
804
ω ≡ (RFi * fc ≥ cfg.WPmcb) ∧ (pi ⊂ ∪ R.icorej)
TVok = { VRi/RFi / ω es cierto para i}
lim = $\overline{TV_{ok}}$ + σ(TVok) . cfg. WPstr , lim = 0 si TV = ∅
806
ω ≡ (RFi * fc ≥ cfg.WPmcb) ∧ (pi ⊂ ∪ R.icorej)
TVizq = { VRi/RFi / ω ∧ (i ∈ [1, cfg.Wpwin])
TVder = { VRi/RFi / ω ∧ (i ∈ [n-cfg.Wpwin, cfg.Wpwin])
$lim_{izq} = \overline{TV_{ok-izq}} + \sigma(TV_{ok-izq})$. cfg. WPstr, limizq = 0 si TVok-izq = ∅
$lim_{der} = \overline{TV_{ok-der}} + \sigma(TV_{ok-der})$. cfg. WPstr, limder = 0 si TVok-der = ∅
$\frac{VR_i}{RF_i} < lim$ ∧ ω
805
WP = WP ∪ {pi}
p ≤ cfg.Wpwin/2
$\frac{VR_i}{RF_i} < lim_{izq}$
p ≥ n-cfg.Wpwin/2
$\frac{VR_i}{RF_i} < lim_{der}$
p ⊂ [cfg.Wpwin/2, n-cfg. Wpwin/2]
pi cumpe ω
TVp = { VRi/RFi / ω ∧ (i ∈ [p- cfg.Wpwin/2, p+ cfg.Wpwin/2])
$lim_p = \overline{TV_{ok-p}} + \sigma(TV_{ok-p})$. cfg. WPstr , limp = 0 si TVp = ∅
807
WP = WP ∪ {pi}
$\frac{VR_i}{RF_i} \geq lim_p$
p = n
p = p + 1
p = n
p = p + 1
808

FIG. 8

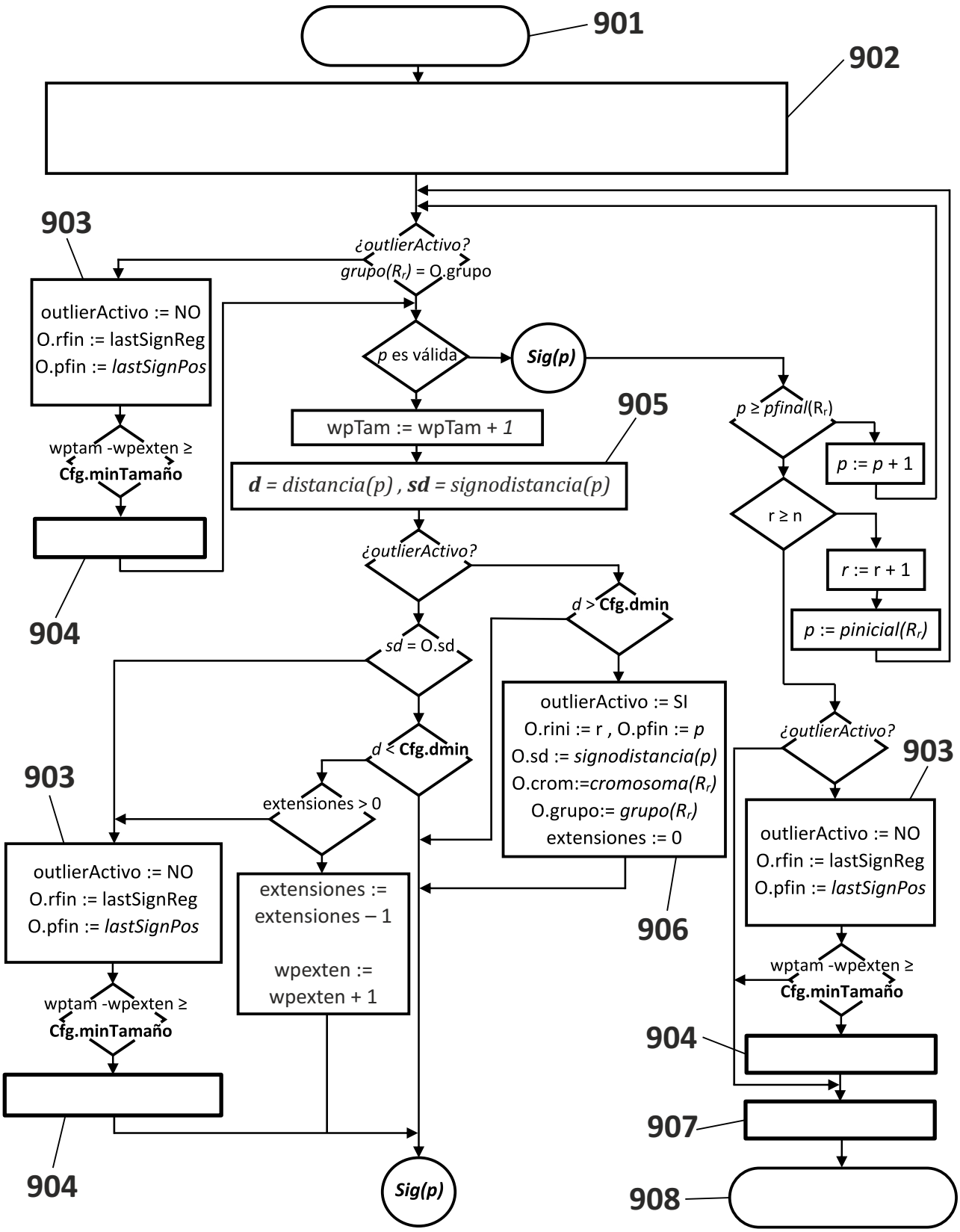
901
902
903
¿outlierActivo?
grupo(Rr) = O.grupo
outlierActivo := NO
O.rfin := lastSignReg
O.pfin := lastSignPos
p es válida
Sig(p)
wptam -wpexten ≥
Cfg.minTamaño
904
wpTam := wpTam + 1
905
d = distancia(p) , sd = signodistancia(p)
p ≥ pfinal(Rr)
p := p + 1
r ≥ n
r := r + 1
p := pinicial(Rr)
¿outlierActivo?
d > Cfg.dmin
sd = O.sd
outlierActivo := SI
O.rini := r , O.pfin := p
O.sd := signodistancia(p)
O.crom:=cromosoma(Rr)
O.grupo:= grupo(Rr)
extensiones := 0
906
d < Cfg.dmin
extensiones > 0
extensiones := extensiones – 1
wpexten := wpexten + 1
907
908

FIG. 9

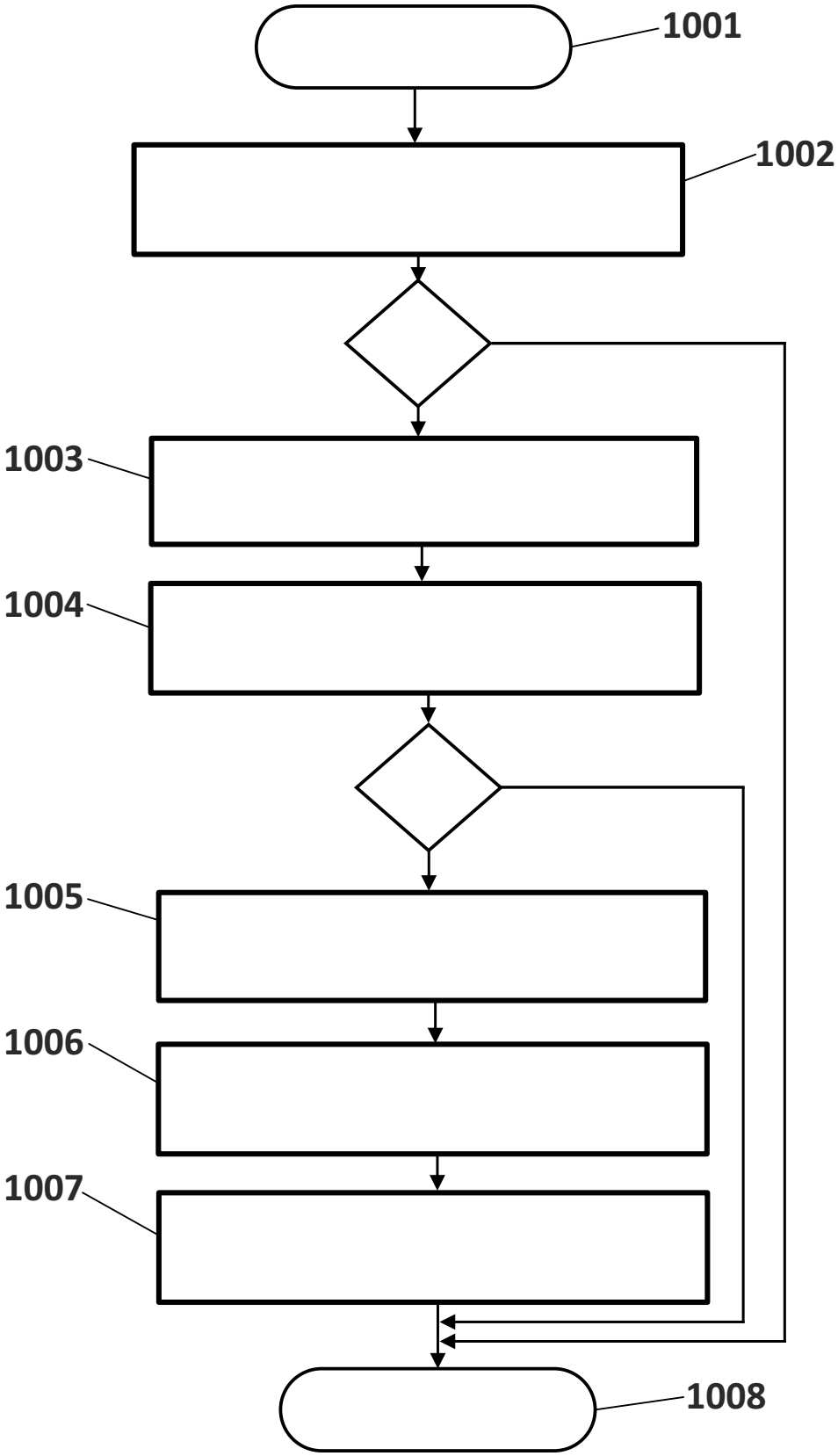

FIG. 10

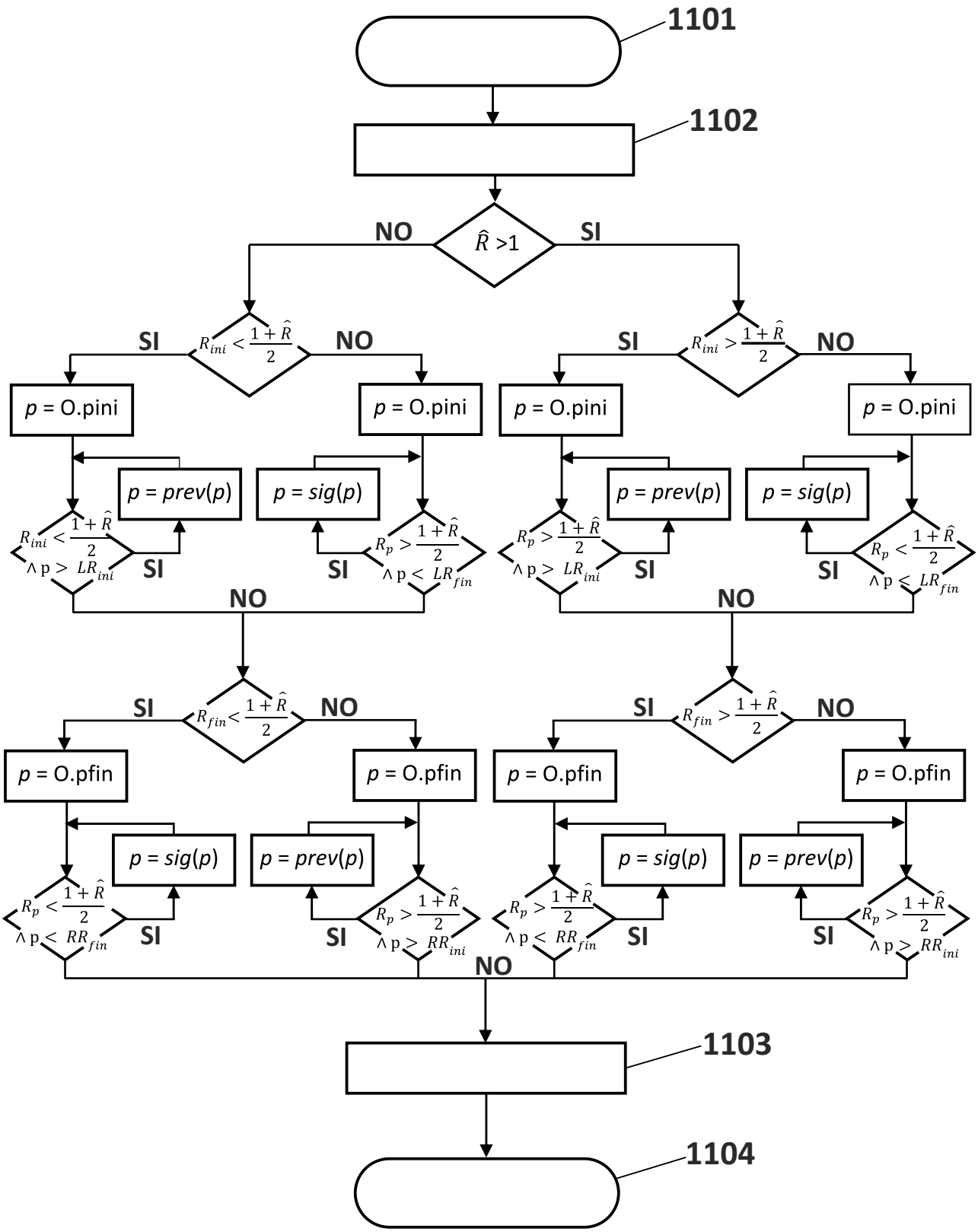
1101
1102
NO
$\hat{R}$ >1
SI
SI
$R_{ini} < \frac{1+\hat{R}}{2}$
NO
SI
$R_{ini} > \frac{1+\hat{R}}{2}$
NO
p = O.pini
p = O.pini
p = O.pini
p = O.pini
p = prev(p)
p = sig(p)
p = prev(p)
p = sig(p)
$R_{ini} < \frac{1+\hat{R}}{2}$ ∧ p > LR_{ini}
SI
SI
$R_p > \frac{1+\hat{R}}{2}$ ∧ p < LR_{fin}
$R_p > \frac{1+\hat{R}}{2}$ ∧ p > LR_{ini}
SI
SI
$R_p < \frac{1+\hat{R}}{2}$ ∧ p < LR_{fin}
NO
NO
SI
$R_{fin} < \frac{1+\hat{R}}{2}$
NO
SI
$R_{fin} > \frac{1+\hat{R}}{2}$
NO
p = O.pfin
p = O.pfin
p = O.pfin
p = O.pfin
p = sig(p)
p = prev(p)
p = sig(p)
p = prev(p)
$R_p < \frac{1+\hat{R}}{2}$ ∧ p < RR_{fin}
SI
SI
$R_p > \frac{1+\hat{R}}{2}$ ∧ p > RR_{ini}
$R_p > \frac{1+\hat{R}}{2}$ ∧ p < RR_{fin}
SI
SI
$R_p > \frac{1+\hat{R}}{2}$ ∧ p > RR_{ini}
NO
1103
1104

FIG. 11

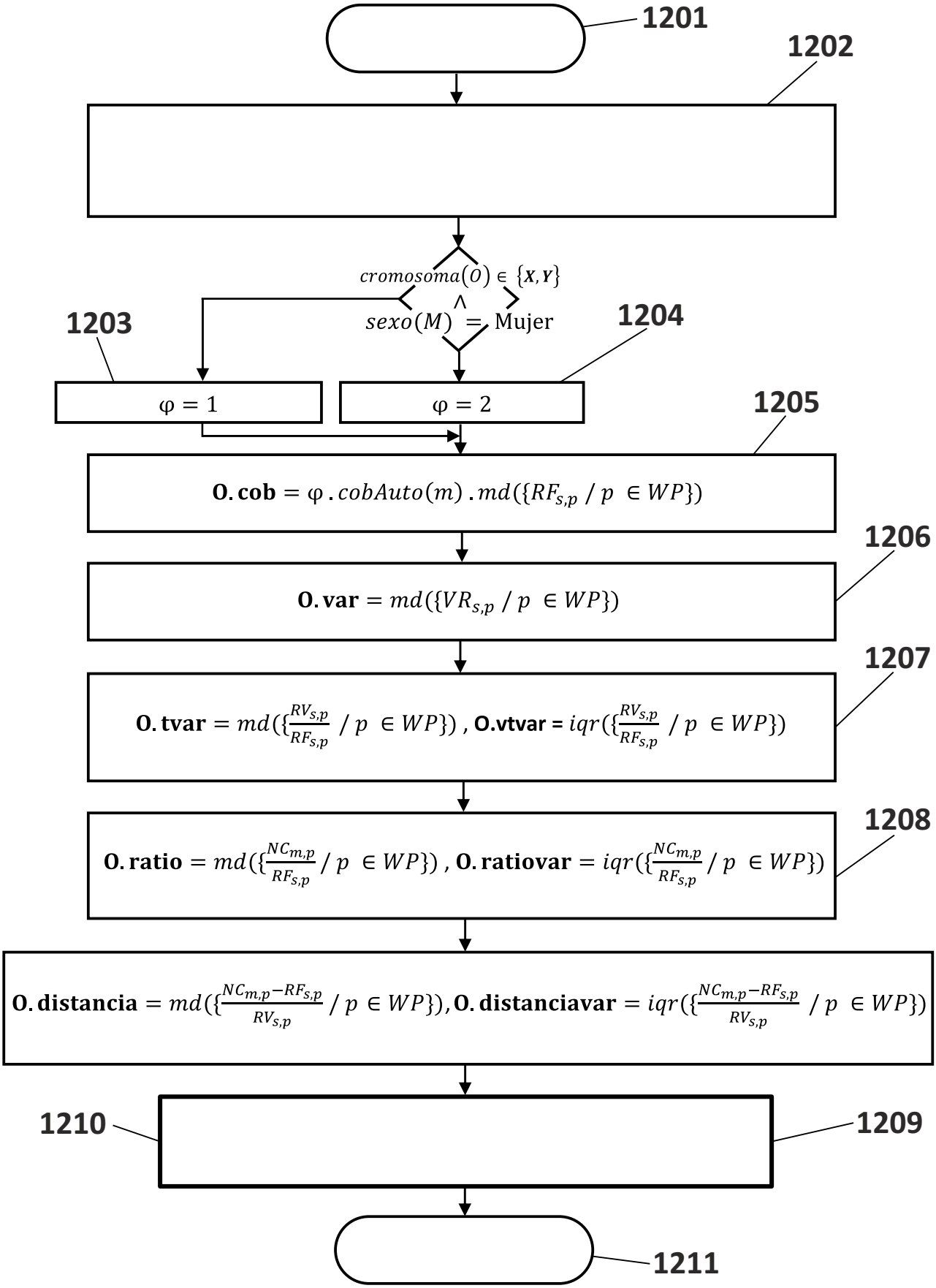
1201
1202
cromosoma(O) ∈ {X, Y}
∧
sexo(M) = Mujer
1203
1204
φ = 1
φ = 2
1205
O. cob = φ . cobAuto(m) . md({RF_s,p / p ∈ WP})
1206
O. var = md({VR_s,p / p ∈ WP})
1207
O. tvar = md({RV_s,p / RF_s,p / p ∈ WP}) , O.vtvar = iqr({RV_s,p / RF_s,p / p ∈ WP})
1208
O. ratio = md({NC_m,p / RF_s,p / p ∈ WP}) , O. ratiovar = iqr({NC_m,p / RF_s,p / p ∈ WP})
O. distancia = md({NC_m,p − RF_s,p / RV_s,p / p ∈ WP}), O. distanciavar = iqr({NC_m,p − RF_s,p / RV_s,p / p ∈ WP})
1210
1209
1211

FIG. 12

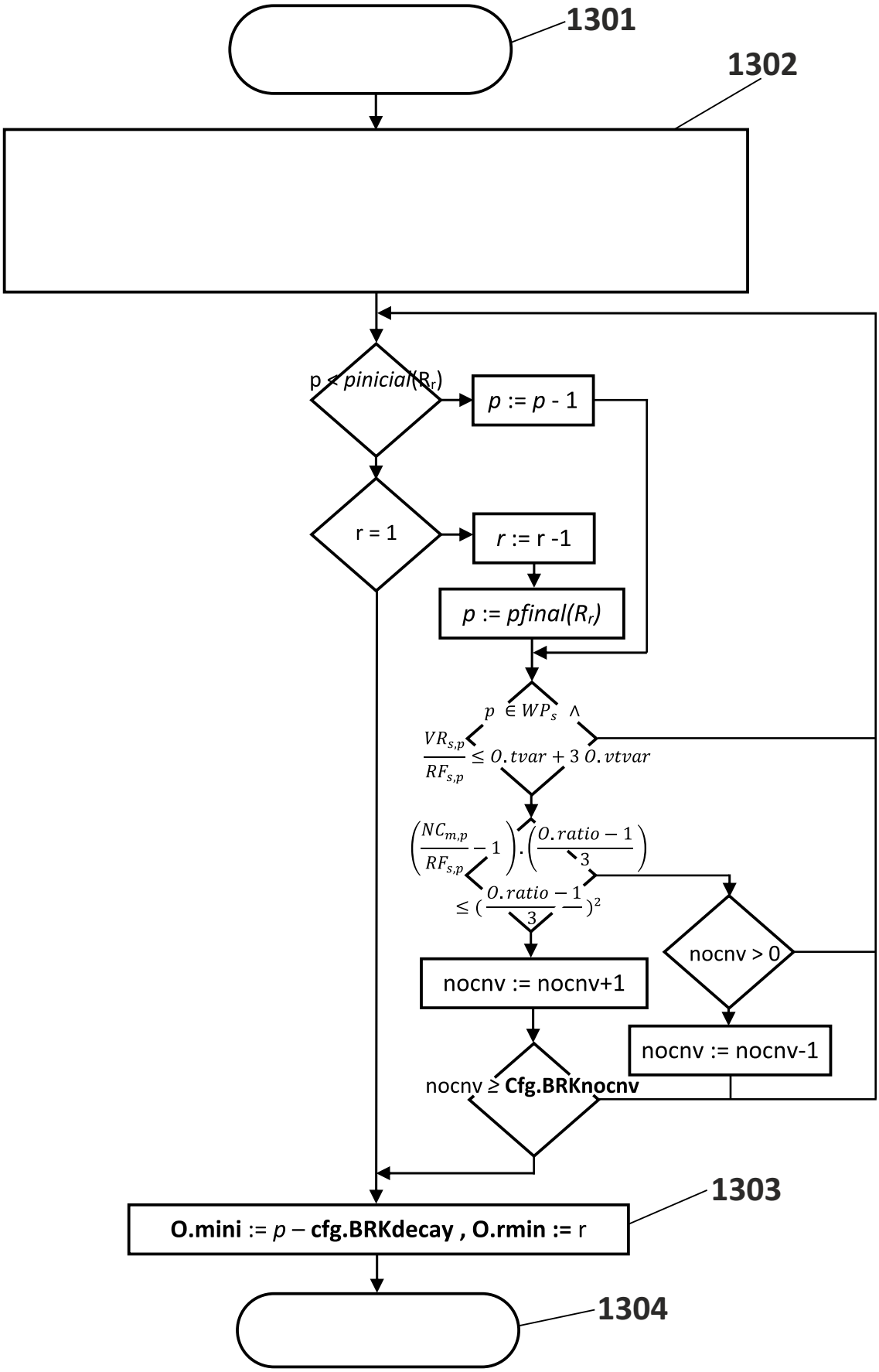
1301
1302
p < pinicial(R_r)
p := p - 1
r = 1
r := r -1
p := pfinal(R_r)
p ∈ WP_s ∧
VR_s,p / RF_s,p ≤ O.tvar + 3 O.vtvar
(NC_m,p / RF_s,p − 1) · ((O.ratio − 1)/3)
≤ ((O.ratio − 1)/3)²
nocnv := nocnv+1
nocnv > 0
nocnv := nocnv-1
nocnv ≥ Cfg.BRKnocnv
1303
O.mini := p – cfg.BRKdecay , O.rmin := r
1304

FIG. 13

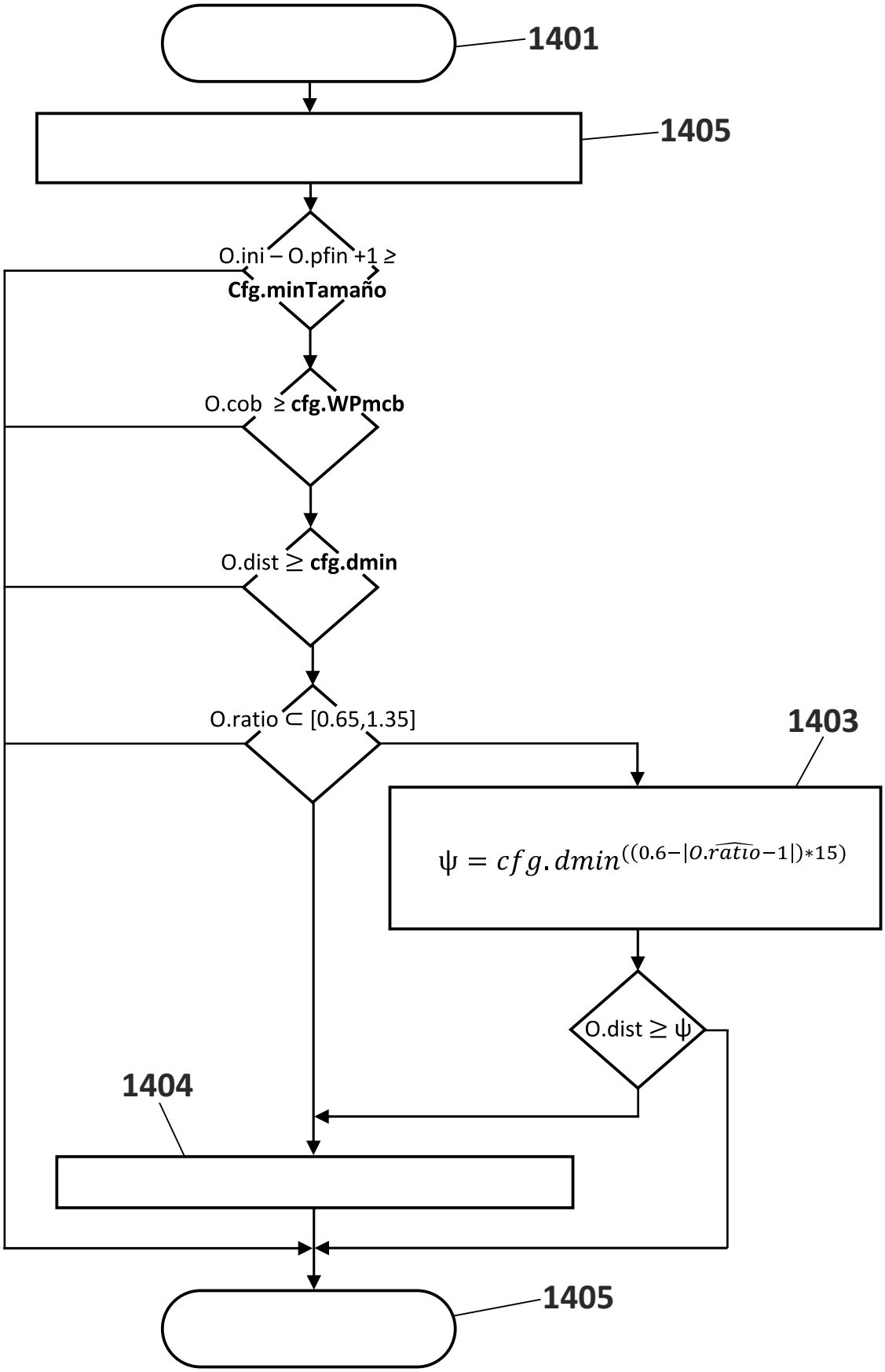
1401
1405
O.ini – O.pfin +1 ≥
Cfg.minTamaño
O.cob ≥ cfg.WPmcb
O.dist ≥ cfg.dmin
O.ratio ⊂ [0.65,1.35]
1403
$\psi = cfg.dmin^{((0.6-|O.\widehat{ratio}-1|)*15)}$
O.dist ≥ ψ
1404
1405

FIG. 14

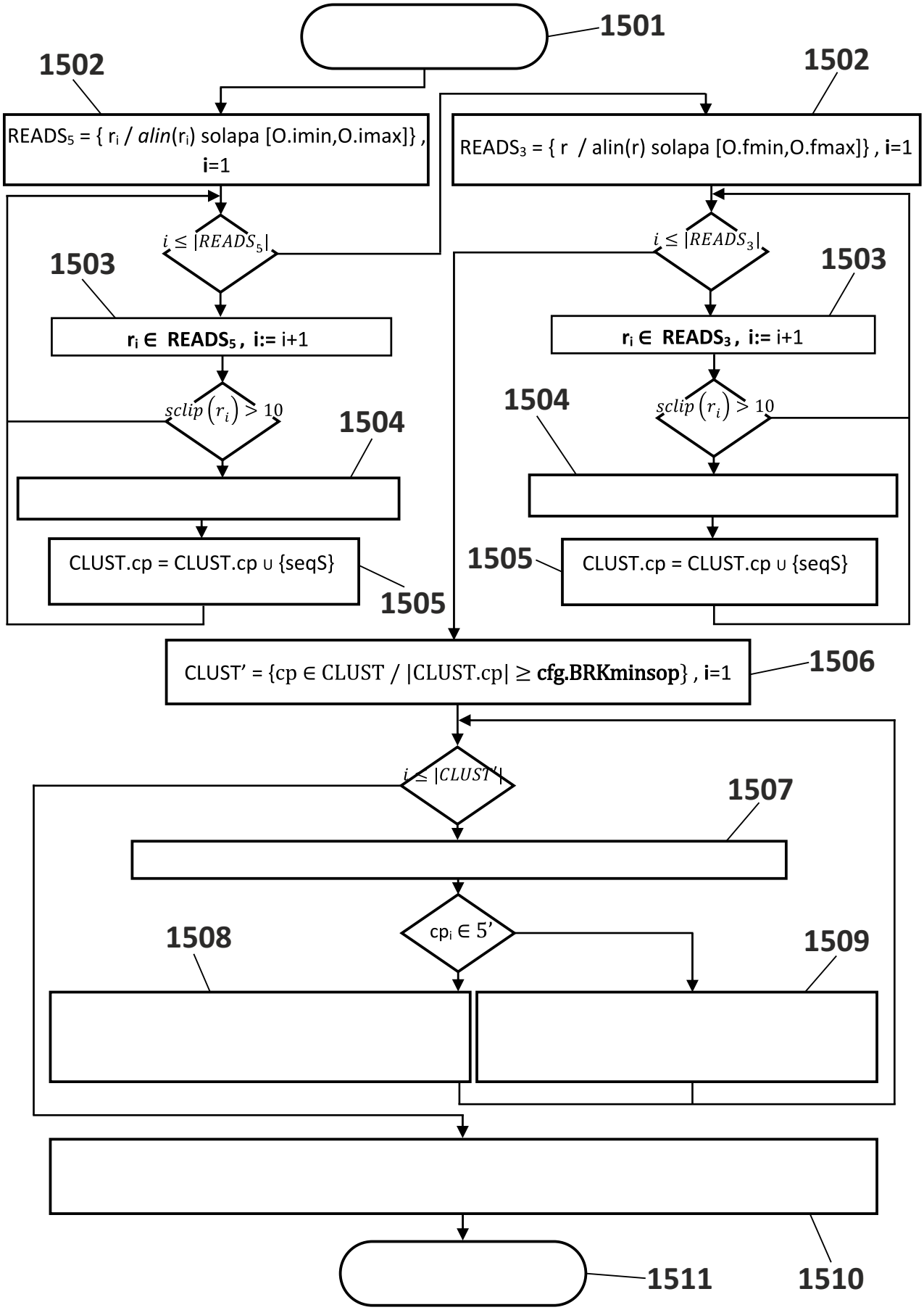
1501
1502
READS5 = { ri / alin(ri) solapa [O.imin,O.imax]} , i=1
1502
READS3 = { r / alin(r) solapa [O.fmin,O.fmax]} , i=1
i ≤ |READS5|
i ≤ |READS3|
1503
ri ∈ READS5 , i:= i+1
1503
ri ∈ READS3 , i:= i+1
sclip (ri) ≥ 10
sclip (ri) ≥ 10
1504
1504
CLUST.cp = CLUST.cp ∪ {seqS}
1505
1505
CLUST.cp = CLUST.cp ∪ {seqS}
CLUST' = {cp ∈ CLUST / |CLUST.cp| ≥ cfg.BRKminsop} , i=1
1506
i ≤ |CLUST'|
1507
cpi ∈ 5'
1508
1509
1511
1510

FIG. 15

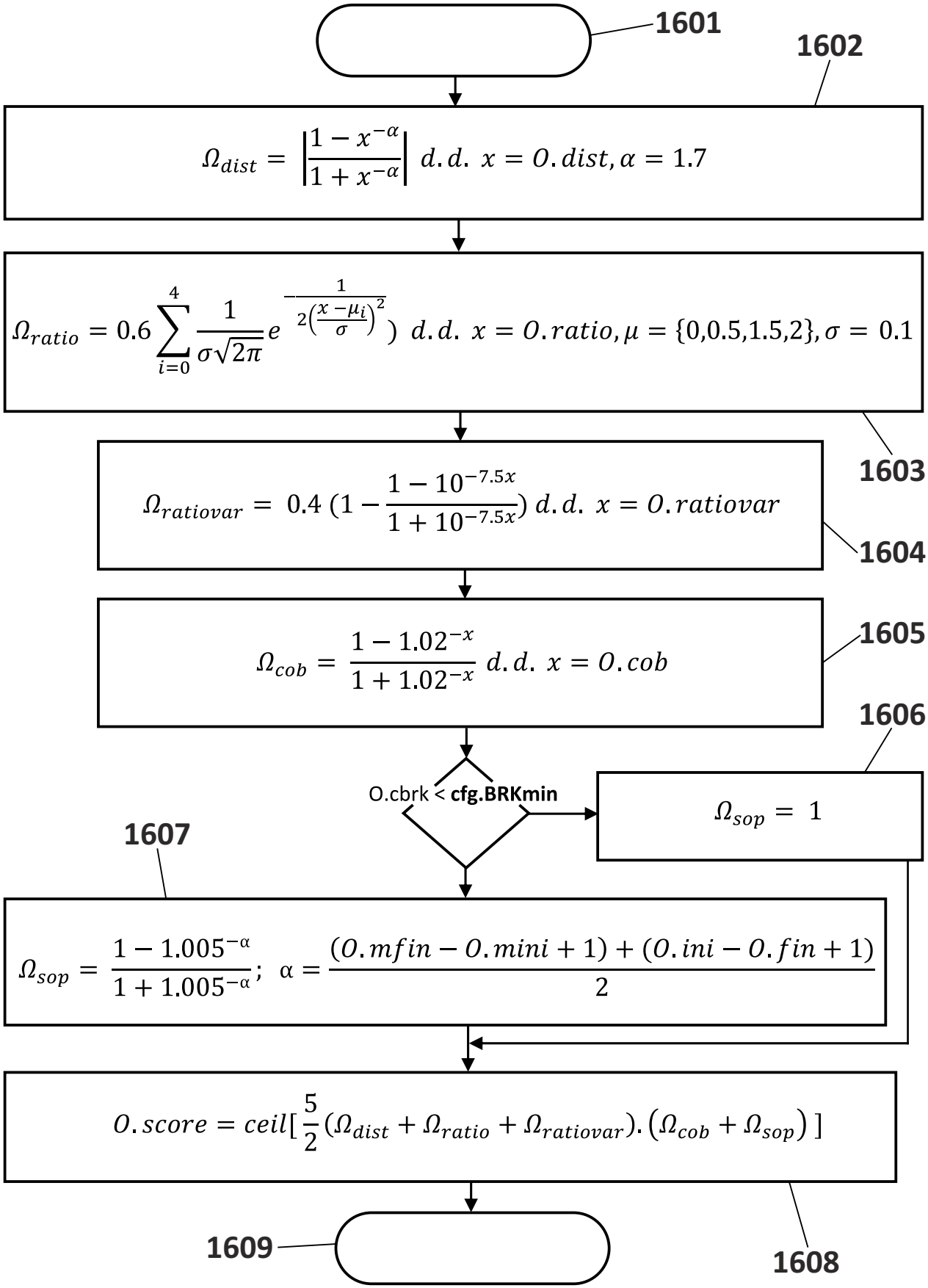
1601
1602
$\Omega_{dist} = \left|\frac{1-x^{-\alpha}}{1+x^{-\alpha}}\right| \; d.d.\; x = O.dist, \alpha = 1.7$
$\Omega_{ratio} = 0.6 \sum_{i=0}^{4} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma}\right)^2} \; d.d.\; x = O.ratio, \mu = \{0,0.5,1.5,2\}, \sigma = 0.1$
1603
$\Omega_{ratiovar} = 0.4\,(1 - \frac{1-10^{-7.5x}}{1+10^{-7.5x}})\; d.d.\; x = O.ratiovar$
1604
$\Omega_{cob} = \frac{1-1.02^{-x}}{1+1.02^{-x}}\; d.d.\; x = O.cob$
1605
1606
O.cbrk < cfg.BRKmin
$\Omega_{sop} = 1$
1607
$\Omega_{sop} = \frac{1-1.005^{-\alpha}}{1+1.005^{-\alpha}};\; \alpha = \frac{(O.mfin - O.mini + 1) + (O.ini - O.fin + 1)}{2}$
$O.score = ceil[\frac{5}{2}(\Omega_{dist} + \Omega_{ratio} + \Omega_{ratiovar}).(\Omega_{cob} + \Omega_{sop})]$
1608
1609

FIG. 16

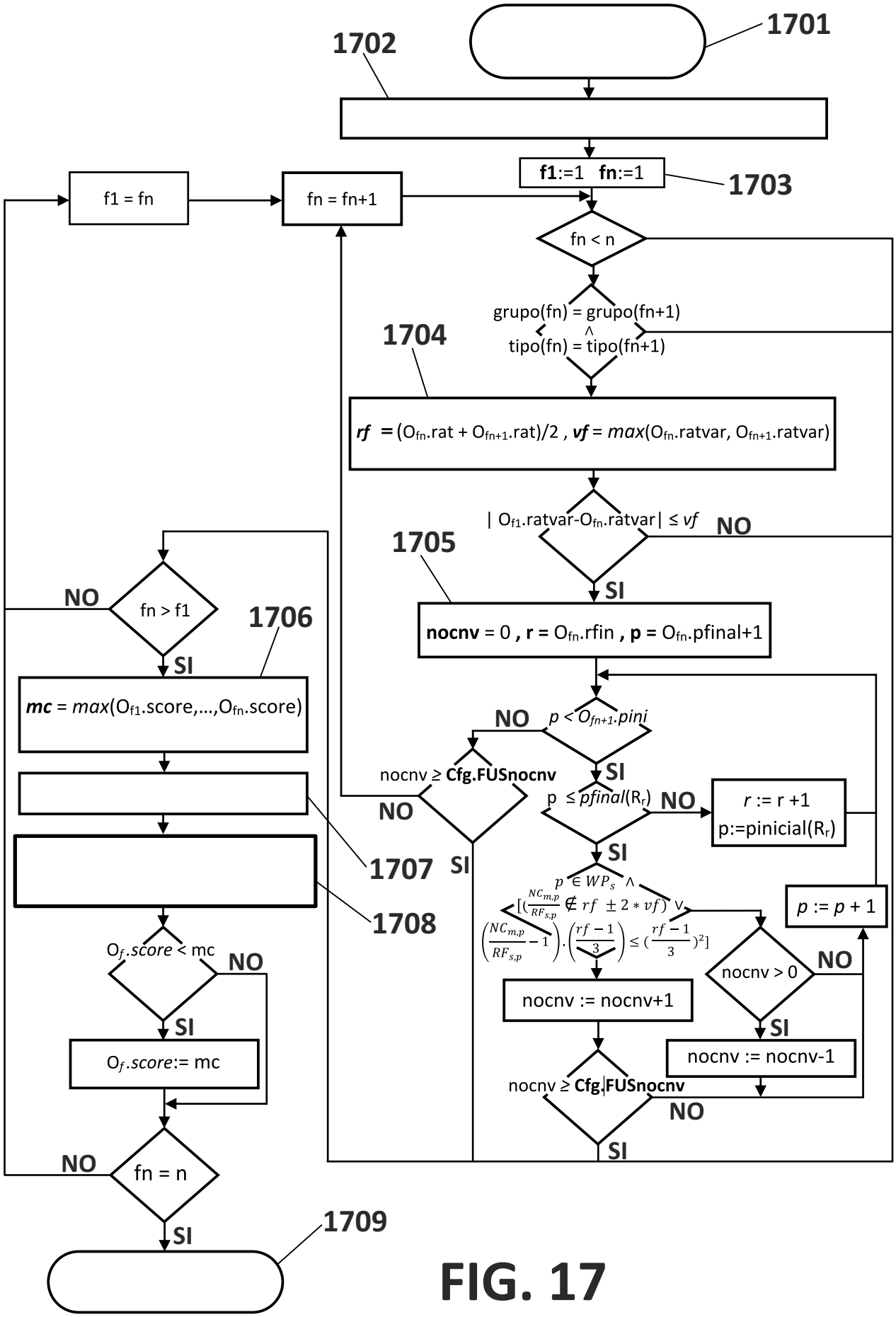
1701
1702
f1:=1 fn:=1
1703
f1 = fn
fn = fn+1
fn < n
grupo(fn) = grupo(fn+1)
∧
tipo(fn) = tipo(fn+1)
1704
rf = (Ofn.rat + Ofn+1.rat)/2 , vf = max(Ofn.ratvar, Ofn+1.ratvar)
| Of1.ratvar-Ofn.ratvar| ≤ vf
NO
SI
1705
nocnv = 0 , r = Ofn.rfin , p = Ofn.pfinal+1
NO
fn > f1
SI
1706
mc = max(Of1.score,...,Ofn.score)
p < Ofn+1.pini
SI
nocnv ≥ Cfg.FUSnocnv
NO
p ≤ pfinal(Rr)
NO
r := r +1
p:=pinicial(Rr)
SI
1707
SI
1708
p ∈ WPs ∧
[(NCm,p/RFs,p ∉ rf ± 2 * vf) ∨
(NCm,p/RFs,p − 1) . (rf − 1/3) ≤ (rf − 1/3)²]
p := p + 1
Of.score < mc
NO
SI
nocnv := nocnv+1
nocnv > 0
NO
SI
Of.score:= mc
nocnv := nocnv-1
nocnv ≥ Cfg.FUSnocnv
NO
SI
NO
fn = n
SI
1709

FIG. 17

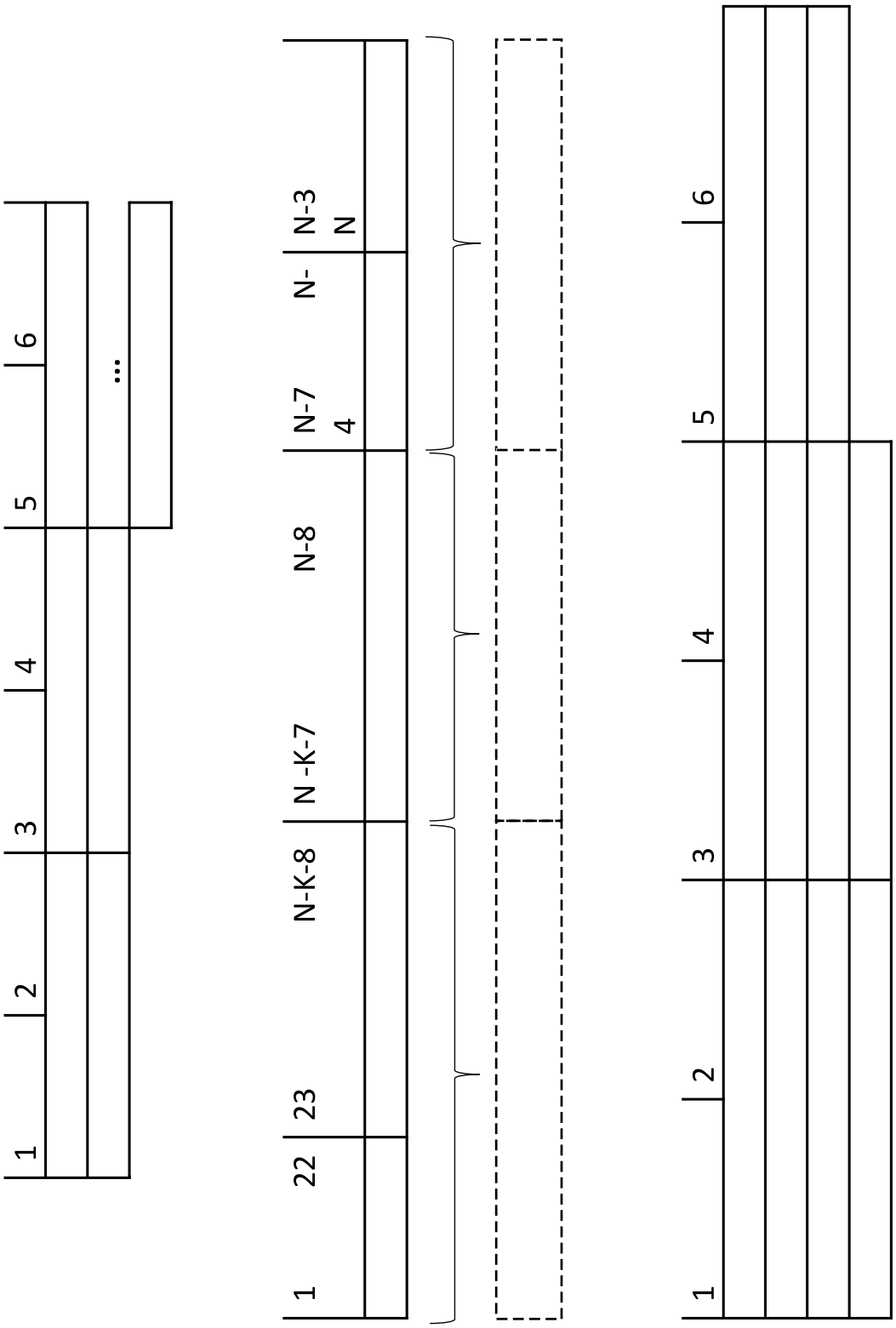

FIG. 18

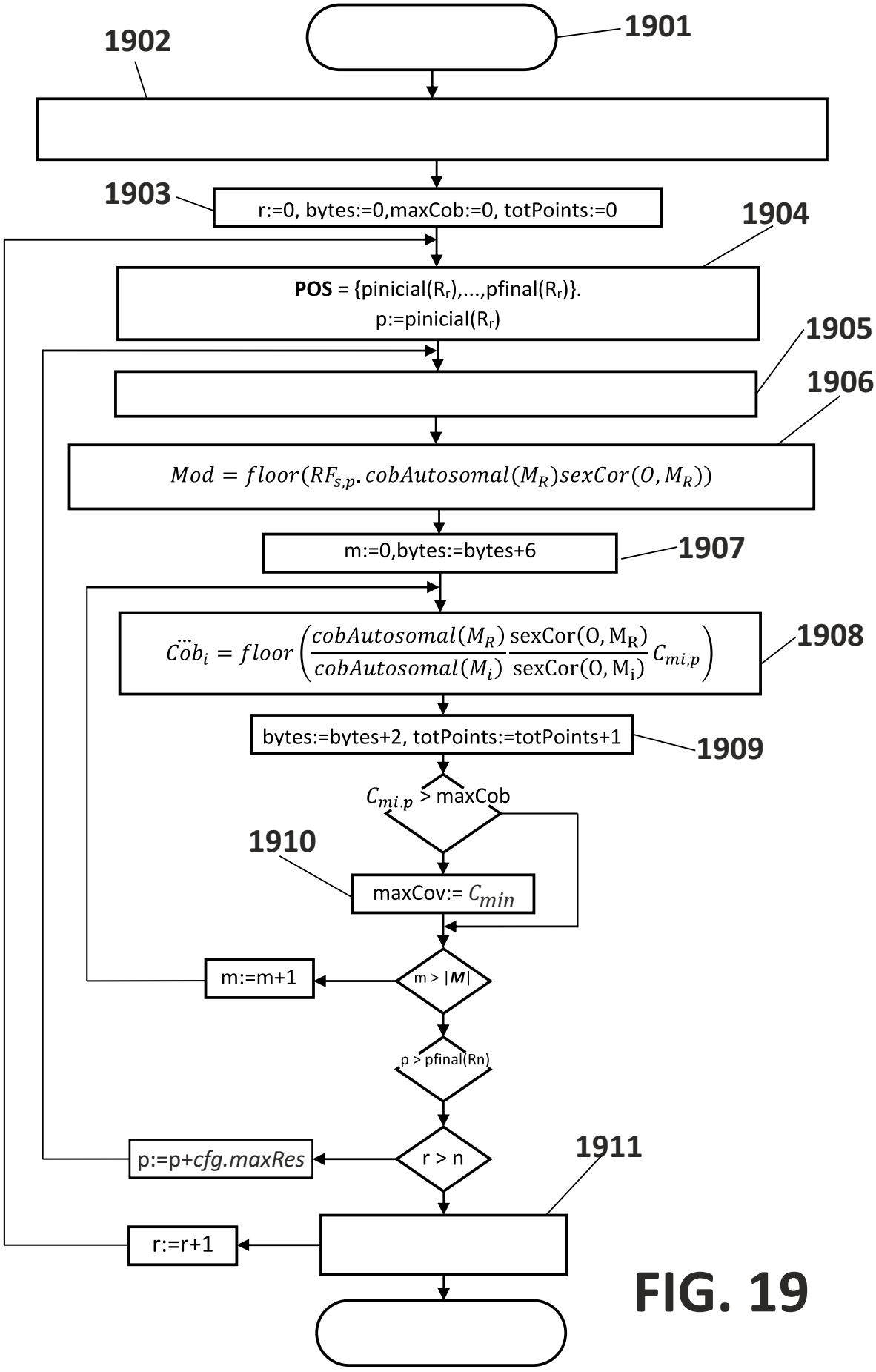
1901
1902
1903
r:=0, bytes:=0,maxCob:=0, totPoints:=0
1904
POS = {pinicial(Rr),...,pfinal(Rr)}.
p:=pinicial(Rr)
1905
1906
$Mod = floor(RF_{s,p}.cobAutosomal(M_R)sexCor(O,M_R))$
m:=0,bytes:=bytes+6
1907
$\dddot{C}ob_i = floor\left(\frac{cobAutosomal(M_R)}{cobAutosomal(M_i)}\frac{\text{sexCor}(\text{O},\text{M}_\text{R})}{\text{sexCor}(\text{O},\text{M}_\text{i})}C_{mi,p}\right)$
1908
bytes:=bytes+2, totPoints:=totPoints+1
1909
$C_{mi.p}$ > maxCob
1910
maxCov:= C_{min}
m:=m+1
m > |M|
p > pfinal(Rn)
p:=p+cfg.maxRes
r > n
1911
r:=r+1

FIG. 19

OFICINA ESPAÑOLA
DE PATENTES Y MARCAS

ESPAÑA

(21) N.º solicitud: 201731242

(22) Fecha de presentación de la solicitud: 23.10.2017

(32) Fecha de prioridad:

INFORME SOBRE EL ESTADO DE LA TECNICA

(51) **Int. Cl. :** ***G06F19/18*** (2011.01)

DOCUMENTOS RELEVANTES

Categoría	(56) **Documentos citados**	**Reivindicaciones afectadas**
X	CA 2739457 A1 (ABBOTT LAB) 06/05/2010, Todo el documento	1-17
X	US 8600718 B1 (STEPANIANTS SERGUEI et al.) 03/12/2013, Todo el documento	1-17
X	WANG, H et al. Copy number variation detection using next generation sequencing read counts. BMC Bioinformatics. Abril 2014, Vol. 15, Nº artículo 109. ISSN 1471-2105, <DOI:10.1186/1471-2105-15-109>	1-17
X	DUAN, J. et al. CNV-TV: A robust method to discover copy number variation from short sequencing reads. BMC Bioinformatics. Mayo 2013, Vol. 14, Nº artículo 150. ISSN 1471-2105, <DOI:10.1186/1471-2105-14-150>	1-17
X	YOON, S. et al. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Research. Septiembre 2009, Vol. 19, Nº 9, páginas 1586 – 1592. ISSN 1088-9051, <DOI:10.1101/gr.092981.109>	1-17
X	SINHA, S. et al. CNV-CH: A convex hull based segmentation approach to detect copy number variations (CNV) using next-generation sequencing data. PLoS ONE. Agosto 2015, Vol. 10, Nº 8, Nº articulo e0135895 <DOI: doi:10.1371/journal.pone.0135895>	1-17 1

Categoría de los documentos citados

X: de particular relevancia
Y: de particular relevancia combinado con otro/s de la misma categoría
A: refleja el estado de la técnica
O: referido a divulgación no escrita
P: publicado entre la fecha de prioridad y la de presentación de la solicitud
E: documento anterior, pero publicado después de la fecha de presentación de la solicitud

El presente informe ha sido realizado

☒ para todas las reivindicaciones ☐ para las reivindicaciones nº:

Fecha de realización del informe	**Examinador**	**Página**
17.09.2018	E. Relaño Reyes	1/3

OFICINA ESPAÑOLA
DE PATENTES Y MARCAS

ESPAÑA

(21) N.º solicitud: 201731242

(22) Fecha de presentación de la solicitud: 23.10.2017

(32) Fecha de prioridad:

INFORME SOBRE EL ESTADO DE LA TECNICA

(51) **Int. Cl. :** ***G06F19/18*** (2011.01)

DOCUMENTOS RELEVANTES

Categoría	(56) **Documentos citados**	**Reivindicaciones afectadas**
X	WO 2016/187051 A1 (REGENERON PHARMACEUTICALS) 24/11/2016, Todo el documento.	1-17
X	US 2009/098547 A1 (GHOSH SRINKA) 16/04/2009, Todo el documento.	1-17
A	PABINGER, S. et al. A survey of tools for variant analysis of next-generation genome sequencing data. Briefings in Bioinformatics. Marzo 2014, Vol. 15, Nº 2, páginas 256-278. ISSN 1477-4054 (electrónico), <DOI: 10.1093/bib/bbs086>	1-17

Categoría de los documentos citados

X: de particular relevancia
Y: de particular relevancia combinado con otro/s de la misma categoría
A: refleja el estado de la técnica
O: referido a divulgación no escrita
P: publicado entre la fecha de prioridad y la de presentación de la solicitud
E: documento anterior, pero publicado después de la fecha de presentación de la solicitud

El presente informe ha sido realizado

☒ para todas las reivindicaciones ☐ para las reivindicaciones nº:

Fecha de realización del informe	**Examinador**	**Página**
17.09.2018	E. Relaño Reyes	2/3

Documentación mínima buscada (sistema de clasificación seguido de los símbolos de clasificación)

G06F

Bases de datos electrónicas consultadas durante la búsqueda (nombre de la base de datos y, si es posible, términos de búsqueda utilizados)

INVENES, EPODOC, PATENW, BIOSIS, EMBASE, MEDLINE, NPL, INSPEC, COMPX, XPESP, XPOAC