(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 766 860**

(51) Int. CI.:

C12Q 1/6869 (2008.01)
G16B 30/00 (2009.01)
G16B 40/00 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **15.05.2013 PCT/CN2013/075622**

(87) Fecha y número de publicación internacional: **20.11.2014 WO14183270**

(96) Fecha de presentación y número de la solicitud europea: **15.05.2013 E 13884613 (4)**

(97) Fecha y número de publicación de la concesión europea: **04.12.2019 EP 2998407**

(54) Título: **Método para detectar anomalías estructurales cromosómicas y dispositivo para ello**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
15.06.2020

(73) Titular/es:

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

(72) Inventor/es:

YANG, CHUANCHUN

(74) Agente/Representante:

FÚSTER OLAGUIBEL, Gustavo Nicolás

DESCRIPCIÓN

Método para detectar anomalías estructurales cromosómicas y dispositivo para ello

Antecedentes

Campo técnico

La presente invención se refiere al campo técnico de las tecnologías de genómica y bioinformática, y más particularmente a un método y a un dispositivo para detectar anomalías estructurales cromosómicas.

Técnica relacionada

Actualmente, los métodos de examen de cromosomas usados frecuentemente incluyen los siguientes.

El análisis de cariotipo, por ejemplo, el análisis de cariotipo de banda G, determina anomalías estructurales cromosómicas basándose en la distribución de 400 a 600 bandas y, por tanto, generalmente puede detectar las anomalías sólo a nivel cromosómico. El método puede detectar la deleción o repetición de más de 5 Mpb en las mejores situaciones, pero no puede detectar fragmentos mucho más pequeños (< 5 M). Además, el método requiere el cultivo de células vivas, y se requiere que las células permanezcan vivas.

La hibridación *in situ* fluorescente (FISH) puede detectar la deleción, repetición y translocación equilibrada de fragmentos mucho más pequeños. Sin embargo, se requiere predeterminar el fragmento de cromosoma que va a detectarse para preparar la sonda correspondiente. Como tal, el método está limitado por el diseño de la sonda. FISH se utiliza con frecuencia para la validación de resultados de detección, debido a que falla en la detección de regiones desconocidas.

El método de microalineamientos incluye dos métodos de sonda. Uno está diseñado basándose en polimorfismos de un solo nucleótido (SNP), y el otro está diseñado basándose en CNV. Por tanto, el método adolece de limitaciones similares a FISH.

Con el desarrollo continuo de tecnologías de secuenciación del genoma completo, el coste de la secuenciación se reduce continuamente, de modo que puede ser posible la generalización de la secuenciación del genoma completo. Es necesario investigar medios para encontrar anomalías estructurales cromosómicas basándose en resultados de secuenciación del genoma completo. Por ejemplo, el documento de Talkowski-ME *et al*. The American Journal of Human Genetics, 88, 469-481 (2011), da a conocer estrategias de secuenciación de próxima generación para detectar reordenamientos cromosómicos equilibrados. Las lecturas de secuenciación se analizan y los pares de lecturas anómalos se agrupan y filtran para detectar puntos de ruptura de reordenamiento.

Sumario

Según un aspecto de la presente invención, que se define por las reivindicaciones adjuntas, se proporciona un método para detectar anomalías estructurales cromosómicas, que comprende: adquirir un resultado de secuenciación del genoma completo de un individuo objetivo, donde el resultado de secuenciación del genoma completo incluye múltiples pares de lecturas, consistiendo cada par de lecturas en dos secuencias de lectura ubicadas respectivamente en dos extremos de un fragmento de cromosoma determinado, y cada par de lecturas se deriva por separado de las cadenas positiva y negativa del fragmento de cromosoma correspondiente, o tanto de la cadena positiva como de la negativa del fragmento de cromosoma correspondiente; alinear el resultado de secuenciación con una secuencia de referencia, para obtener un conjunto de apareamiento anómalo, donde el conjunto de apareamiento anómalo incluye un primer tipo de pares de lecturas según la descripción en que dos secuencias de lectura en el primer tipo de par de lecturas se aparean respectivamente con cromosomas diferentes de la secuencia de referencia; agrupar las secuencias de lectura en el conjunto de apareamiento anómalo basándose en las posiciones apareadas con las mismas, donde cada agrupación contiene secuencias de lectura de un solo extremo procedentes de un grupo de pares de lecturas, y las secuencias de lectura del otro extremo correspondientes residen en otra agrupación; filtrar las agrupaciones resultantes, incluyendo calcular la compacidad de cada agrupación y eliminar por filtración las agrupaciones que tienen una compacidad que no cumple un requisito preestablecido R-va y las agrupaciones emparejadas con ellas, y obtener las agrupaciones de resultados filtradas que contienen el primer tipo de pares de lecturas, para determinar la aparición de anomalía estructural cromosómica de tipo translocación.

Según otro aspecto de la presente invención, se proporciona un dispositivo para detectar anomalías estructurales cromosómicas, que comprende una unidad de entrada de datos, configurada para introducir datos; una unidad de salida de datos, configurada para emitir datos; una unidad de almacenamiento que contiene un programa ejecutable en la misma y configurada para almacenar datos; y un procesador, en conexión de datos con la unidad de entrada de datos, la unidad de salida de datos y la unidad de almacenamiento, y configurado para ejecutar el programa ejecutable almacenado en la unidad de almacenamiento, donde la ejecución del programa incluye realizar el método

para detectar anomalías estructurales cromosómicas.

Según un aspecto adicional de la presente invención, se proporciona un medio de almacenamiento legible por ordenador, que está configurado para almacenar un programa ejecutable por un ordenador. Los expertos habituales en la técnica deben entender que cuando se ejecuta el programa, la totalidad o parte de las etapas del método para detectar anomalías estructurales cromosómicas pueden realizarse mediante hardware relevante según instrucciones. El medio de almacenamiento puede incluir una memoria de solo lectura, una memoria de acceso aleatorio, un disco magnético o un disco óptico.

Según el método de la presente invención, la anomalía estructural cromosómica de tipo translocación puede examinarse obteniendo pares de lecturas apareadas con cromosomas diferentes a través de la alineación del resultado de secuenciación del genoma completo con la secuencia de referencia, y la eficacia y la fiabilidad del resultado obtenido se mejoran adicionalmente mediante agrupación y filtración, obteniendo de ese modo resultados de importancia analítica.

Breve descripción de los dibujos

Los aspectos y ventajas anteriores y/u otros adicionales de la presente invención se hacen evidentes y comprensibles a partir de la descripción de realizaciones detalladas en relación con los dibujos adjuntos, en los que:

la figura 1 es una vista esquemática de un par de lecturas obtenidas mediante secuenciación de extremos emparejados según una realización de la presente invención;

la figura 2 es una vista esquemática de un primer tipo de pares de lecturas de apareamiento anómalo según una realización de la presente invención;

la figura 3 es una vista esquemática de un segundo tipo de pares de lecturas de apareamiento anómalo según una realización de la presente invención;

la figura 4 es una vista esquemática de un tercer tipo de pares de lecturas de apareamiento anómalo según una realización de la presente invención;

la figura 5 es una vista esquemática de un par de agrupaciones ubicadas en cromosomas diferentes según una realización de la presente invención;

la figura 6 es una vista esquemática de RPK para “FA” en el ejemplo experimental 1 según una realización de la presente invención; y

la figura 7 es una vista esquemática de RPK para “SON” en el ejemplo experimental 1 según una realización de la presente invención.

Descripción detallada

Según una realización de la presente invención, se proporciona un método para detectar anomalías estructurales cromosómicas, que incluye las etapas siguientes.

Etapa 1. Obtener un resultado de secuenciación del genoma completo de un individuo objetivo

El resultado de secuenciación incluye lecturas emparejadas, consistiendo cada par de lecturas en dos secuencias de lectura ubicadas respectivamente en dos extremos de un fragmento de cromosoma determinado, y cada par de lecturas se deriva por separado de las cadenas positiva y negativa del fragmento de cromosoma correspondiente, o tanto de la cadena positiva como de la negativa del fragmento de cromosoma correspondiente.

El fragmento de cromosoma determinado se obtiene generalmente segmentando la muestra de cromosoma de un individuo objetivo, y luego se usa para la preparación de una biblioteca correspondiente según el método de secuenciación seleccionado. El método de secuenciación puede seleccionarse dependiendo de la plataforma de secuenciación incluyendo, pero sin limitarse a, Complete Genomics (CG), Illumina/Solexa, ABI/SOLiD y Roche 454, y la preparación de una biblioteca de secuenciación de un solo extremo o de extremos emparejados se realiza según la plataforma de secuencia seleccionada. Según una realización específica de la presente invención, puede realizarse secuenciación de extremos emparejados. Las dos secuencias de lectura Read1 y Read2 en cada par de lecturas obtenidas se derivan por separado de la cadena positiva Sp y la cadena negativa Sm del fragmento de cromosoma correspondiente, tal como se muestra en la figura 1. La longitud L-r1 de Read1 puede ser igual o diferente de la longitud L-r2 de Read2. Definitivamente, cuando el método de secuenciación de un solo extremo usado puede adquirir completamente la secuencia del fragmento de cromosoma completo, puede ser factible segmentar una longitud de secuencia adecuada, respectivamente, de los dos extremos de la secuencia totalmente adquirida, para formar un par de lecturas. En este caso, las dos secuencias de lectura en cada par de lecturas se

derivan ambas de la cadena positiva o negativa del fragmento de cromosoma correspondiente. En esta realización, no hay limitaciones en el método de secuenciación específico seleccionado.

En la presente invención, el tamaño de la biblioteca usada para la secuenciación se designa como L-lib, y en general, una biblioteca con una L-lib de 100 a 1000 pb se denomina biblioteca de fragmentos pequeños, y una biblioteca con una L-lib de 2 K, 5 K-6 K, 10 K, 20 K y 40 Kpb se denomina biblioteca de fragmentos grandes. En la presente invención, el tamaño de L-lib no está limitado. Sin embargo, con la calidad de construcción de la biblioteca garantizada, una biblioteca más larga generalmente contribuye a la adquisición de un resultado válido. Por tanto, se prefiere que L-lib $\geq$ 300 pb. Generalmente, puede usarse una biblioteca de fragmentos grandes de por ejemplo 5 Kpb o una biblioteca de fragmentos pequeños de por ejemplo 500 pb. Para permitir que el resultado de secuenciación tenga una buena abundancia, la profundidad de secuenciación puede seleccionarse para que sea superior a 2X para la biblioteca de fragmentos grandes y superior a 5X para la biblioteca de fragmentos pequeños. Para evitar el desperdicio de datos, la profundidad de secuenciación es preferiblemente 2X para la biblioteca de fragmentos grandes y 5X para la biblioteca de fragmentos pequeños. Debe indicarse que debido a que la mayoría de los datos específicos implicados en la presente invención son estadísticamente significativos, cualquier valor numérico que se expresa con precisión representa un intervalo, es decir, un rango dentro de $\pm$ el 10% del valor numérico, a menos que se establezca particularmente otra cosa. La descripción no se repite a continuación en el presente documento.

L-r1 y L-r2 son preferiblemente mayores de o iguales a 25 pb, porque cuando L-r1 y L-r2 son menores de 25 pb, la tasa de alineación única disminuye, de manera que la complejidad aumenta en la adquisición posterior de los resultados de la alineación. No es necesario que L-r1 y L-r2 sean demasiado grandes para evitar el desperdicio de datos. Por tanto, L-r1 y L-r2 son preferiblemente de 50 pb. L-r1 y L-r2 no tienen limitaciones de valor máximo y pueden variar con el desarrollo de la tecnología de secuenciación. Por ejemplo, L-r1 y L-r2 generalmente no son mayores de 150 pb basándose en las tecnologías de secuenciación existentes.

Etapa 2. Alinear el resultado de secuenciación con una secuencia de referencia

La secuencia de referencia usada es una secuencia conocida, y puede ser cualquier molde de referencia obtenido previamente de la categoría a la que pertenece el individuo objetivo. Por ejemplo, si el individuo objetivo es humano, la secuencia de referencia puede ser HG19 proporcionada por el Centro Nacional de Información Biotecnológica (NCBI). Además, puede configurarse previamente un repositorio de recursos que comprende más secuencias de referencia, y se selecciona una secuencia de referencia más próxima dependiendo del sexo, la raza, la región geográfica y otros factores del individuo objetivo antes de la alineación de secuencia, para promover la adquisición de un resultado de detección más preciso. Durante la alineación, según los parámetros de alineación establecidos, se permite que un par de lecturas tenga n apareamientos erróneos como máximo, donde n es preferiblemente 1 ó 2. Si se producen más de n apareamientos erróneos en el par de lecturas, se considera que el par de lecturas no puede alinearse con la secuencia de referencia, o si las n bases con apareamiento erróneo están todas ubicadas en uno del par de lecturas, se considera que la secuencia de lectura en el par de lecturas no puede alinearse con la secuencia de referencia. Específicamente, pueden usarse diversos software de alineación, por ejemplo, Short Oligonucleotide Analysis Package (SOAP), bwa y samtools, etc., durante la alineación. Esto no está limitado en esta realización.

Dependiendo de la alineación de los pares de lecturas, pueden obtenerse las siguientes categorías.

(I) Conjunto de apareamiento normal *.pair, incluye pares de lecturas según la descripción en que dos secuencias de lectura Read1 y Read2 en el par de lecturas se aparean con el mismo cromosoma de la secuencia de referencia, la relación de las cadenas positiva y negativa de las posiciones apareadas con las mismas es conforme con la de en el par de lecturas, y la longitud L-pr del fragmento de cromosoma calculada según las posiciones apareadas con las mismas se desvía de L-lib en un valor menor que un umbral preestablecido V-lib. V-lib es preferiblemente desde el 5% x L-lib hasta el 15% x L-lib, y más preferiblemente del 10% x L-lib. El umbral se establece empíricamente basándose en la desviación estándar del tamaño de biblioteca. Empíricamente, la desviación estándar de una biblioteca de fragmentos pequeños es de aproximadamente 15 pb, y la desviación estándar de una biblioteca de fragmentos grandes es de aproximadamente 50 pb. Se considera adecuada una desviación de L-pr de L-lib en el intervalo de deviaciones estándar de 3 veces, por ejemplo, se considera que L-pr es adecuada en el intervalo de 455-545 pb para una biblioteca de 500 pb.

Basándose en *.pair, puede obtenerse la distribución cuantitativa de los pares de lecturas según la posición apareada con las mismas, por ejemplo, el número de lecturas por longitud unitaria (RPU) puede calcularse estadísticamente. La longitud unitaria correspondiente puede establecerse según L-lib, por ejemplo, puede establecerse en 1,5-4 veces L-lib. Si L-lib es de 500 pb, la longitud unitaria puede establecerse en 1 Kpb, y en este caso, RPU puede designarse como RPK. La variación de RPU en comparación con el valor promedio, por ejemplo, si la variación está por encima de un umbral preestablecido V-rm, puede ayudar en la determinación de la aparición de anomalías estructurales, aumentando de ese modo la precisión del análisis del resultado. Preferiblemente, V-rm es del 10 al 30% y más preferiblemente del 20%. Además, las RPU promedio pueden obtenerse a través del cálculo estadístico o de la estimación. Por ejemplo, las RPU promedio pueden estimarse mediante una fórmula: profundidad

de secuenciación x (longitud unitaria/L-lib). Si las RPU no son necesarias, no es necesario obtener * .pair.

(II) Conjunto de apareamiento anómalo *.sin, incluye tres tipos de lecturas según las descripciones a continuación.

En un primer tipo, dos secuencias de lectura en un par de lecturas se aparean respectivamente con cromosomas diferentes de la secuencia de referencia. Este tipo de pares de lecturas se correlacionan con anomalía estructural de tipo translocación, por ejemplo, translocación equilibrada y desequilibrada. La figura 2 muestra una situación de translocación equilibrada, en la que Read1 en un par de lecturas se aparea con el cromosoma chra, y Read2 se aparea con el cromosoma chrb, y el caso es el opuesto para otro par de lecturas. En la figura 2, la línea discontinua que conecta Read1 y Read2 indica la relación de posición de cabeza con cola de Read1 y Read2 en el fragmento de cromosoma (el mismo a continuación), y pa y pb indican posiciones potenciales de puntos de ruptura respectivamente, donde “punto de ruptura” significa punto de límite de aparición de anomalías estructurales en un cromosoma.

En un segundo tipo, dos secuencias de lectura en un par de lecturas se aparean con el mismo cromosoma de la secuencia de referencia, pero L-pr es negativa. Este tipo de pares de lecturas se correlacionan con anomalía estructural de tipo de repetición en tándem. Tal como se muestra en la figura 3, Read1 y Read2 en un par de lecturas se aparean ambas con el cromosoma chra; sin embargo, la relación de posición de cabeza con cola de las posiciones apareadas con las mismas es opuesta a la de Read1 y Read2 en el fragmento de cromosoma. En la figura 3, pa1 y pa2 indican las posiciones de inicio y terminación del fragmento de repetición potencial respectivamente, L-sv indica la longitud del fragmento de repetición, y la línea discontinua en la parte media de chra indica longitud omitida (la misma a continuación).

En un tercer tipo, dos secuencias de lectura en un par de lecturas se aparean con el mismo cromosoma de la secuencia de referencia; sin embargo, L-pr es mayor que L-lib, y la desviación está por encima del umbral preestablecido V-lib. Este tipo de pares de lecturas se correlacionan con anomalía estructural de tipo deleción. Tal como se muestra en la figura 4, Read1 y Read2 en un par de lecturas se aparean ambas con el cromosoma chra, y la relación de posición de cabeza con cola de las posiciones apareadas con las mismas es igual que la de Read1 y Read2 en el fragmento de cromosoma, pero la distancia va más allá de un intervalo adecuado. En la figura 4, pa1 y pa2 indican las posiciones de inicio y terminación del fragmento potencialmente delecionado, respectivamente, y L-sv indica la longitud del fragmento delecionado.

Puesto que diferentes tipos de pares de lecturas en el conjunto de apareamiento anómalo representan diferentes tipos de anomalías estructurales cromosómicas que pueden producirse respectivamente, los pares de lecturas de apareamiento anómalo de los tipos anteriores pueden no tener que adquirirse totalmente, según se desee mediante detección. Por ejemplo, cuando sólo es necesario detectar la anomalía estructural del tipo de translocación, el primer tipo de pares de lecturas puede adquirirse únicamente del resultado de alineación. Además, el conjunto de apareamiento anómalo no incluye exclusivamente los tres tipos anteriores de pares de lecturas, y los pares de lecturas o una secuencia de lectura en los pares de lecturas que no pertenecen al conjunto de apareamiento normal y que pueden aparearse con la secuencia de referencia se incluyen en el conjunto de apareamiento anómalo. Las manifestaciones de diferentes tipos de apareamientos anómalos pueden correlacionarse con anomalías estructurales cromosómicas correspondientes que pueden ocurrírseles a los expertos en la materia. Además, teniendo en cuenta la influencia del ruido potencial y otras interferencias, puede no discriminarse el apareamiento o apareamiento erróneo de las cadenas positiva y negativa en el conjunto de apareamiento anómalo.

(III) Conjunto sin apareamiento *.unmap, incluye secuencias de lectura que no pueden aparearse con la secuencia de referencia, y que pueden emparejarse (ninguna puede aparearse con la secuencia de referencia) o ser secuencias de lectura de un solo extremo (las secuencias de lectura del otro extremo pueden aparearse con la secuencia de referencia).

Las secuencias de lectura de un solo extremo existentes en *.unmap se usan adicionalmente para el ensamblado de puntos de ruptura tras adquirir las agrupaciones de resultados, para obtener un intervalo de puntos de ruptura más preciso. Si no es necesario el ensamblado de puntos de ruptura, no es necesario obtener *.unmap.

Etapa 3. Agrupar las secuencias de lectura en *.sin basándose en las posiciones apareadas con las mismas

El agrupamiento puede lograrse mediante el uso de varios algoritmos de agrupamiento, que no están limitados en esta realización. Por ejemplo, un enfoque sencillo es dividir las agrupaciones basándose en una distancia mínima establecida entre agrupaciones V-cl. Es decir, se buscan secuencias de lectura clasificadas según las posiciones. A partir de una primera secuencia de lectura, si la distancia entre una segunda secuencia de lectura y la primera secuencia de lectura es menor que V-cl, se dividen en la misma agrupación. Entonces, la búsqueda continúa desde la segunda secuencia de lectura, hasta que la distancia entre una enésima secuencia de lectura y una (n-1) secuencia de lectura es mayor que V-cl. Entonces puede dividirse una segunda agrupación a partir de la enésima secuencia de lectura, y el procedimiento se realiza repetidamente a través de todas las secuencias de lectura. Durante el agrupamiento, esto puede realizarse según la posición de lectura apareada en el cromosoma, independientemente de la cadena positiva o negativa.

Tras agrupar, cada agrupación contiene secuencias de lectura de un solo extremo procedentes de un grupo de pares de lecturas, y de manera correspondiente las secuencias de lectura del otro extremo están ubicadas en otra agrupación. Por tanto, las dos agrupaciones se denominan un par de agrupaciones. La figura 5 es una vista esquemática de un par de agrupaciones formadas por cluster1 y cluster2 ubicadas en cromosomas diferentes respectivamente. En definitiva, pueden ubicarse agrupaciones emparejadas en el mismo cromosoma. Para que el análisis tras el agrupamiento sea significativo, cada agrupación incluye preferiblemente más de dos lecturas. En caso de que la distancia entre una lectura individual y una lectura anterior y siguiente sea mayor que V-cl, dicho valor atípico puede descartarse.

V-cl es al menos L-lib. Si el valor establecido es demasiado bajo, el número de agrupaciones candidatas es demasiado grande y el número de secuencias de lectura en la agrupación es demasiado pequeño, lo que resulta inconveniente para el examen y la filtración posteriores, y también puede conducir a un aumento de resultados falsos positivos. Si el valor establecido es demasiado alto, la determinación de los puntos de ruptura puede resultar inconveniente y se amplía el intervalo de puntos de ruptura. Por tanto, V-cl es preferiblemente de 10 Kpb. Dependiendo de los diferentes algoritmos de agrupamiento utilizados, V-cl puede tener diferentes significados específicos. Por ejemplo, V-cl puede ser la distancia entre los centros de gravedad de dos agrupaciones adyacentes, o la distancia entre dos lecturas más próximas en la posición en dos agrupaciones adyacentes.

Etapa 4. Filtrar las agrupaciones resultantes

La filtración tiene como objetivo eliminar diversas interferencias potenciales en la mayor medida, por ejemplo, contaminación de la muestra, errores de secuenciación, errores de alineación y ruido, de manera que el resultado pueda reflejar anomalías estructurales cromosómicas reales en el mayor grado. Por tanto, las condiciones de filtración pueden establecerse dependiendo de los requisitos reales y los tipos de posibles interferencias. En esta realización, se proporcionan preferiblemente los siguientes medios de filtración, que pueden usarse solos o en combinación de varios de ellos en aplicación práctica.

(I) En virtud de la compacidad de las agrupaciones - Se calcula la compacidad de cada agrupación, y se eliminan por filtración las agrupaciones que tienen una compacidad que no cumple un requisito preestablecido R-va y las agrupaciones emparejadas con ellas. La compacidad de cada agrupación puede calcularse usando diversos métodos matemáticos disponibles. Por ejemplo, la compacidad puede indicarse mediante varianza. Se calcula la varianza de la posición de cada lectura en una agrupación con respecto al centro o al centro de gravedad de la agrupación, y cuanto menor es la varianza, mayor será la compacidad. Preferiblemente, durante el cálculo de la compacidad de cada agrupación, pueden excluirse secuencias de lectura ubicadas en dos extremos en el intervalo del 5 al 25% y preferiblemente del 20% en longitud de la agrupación, para reducir la influencia de los datos periféricos en el resultado del cálculo. Preferiblemente, R-va puede establecerse en un umbral fijo, por ejemplo, se requiere que la varianza sea menor que el umbral fijo. O de otro modo, R-va se establece para que sea una tasa de eliminación. Por ejemplo, se requiere que los rangos de las varianzas en todas las agrupaciones se encuentren dentro del intervalo más bajo preestablecido. Por ejemplo, R-va se establece de manera que los rangos de las varianzas en todas las agrupaciones se encuentran dentro del intervalo más bajo del 2%-10% y preferiblemente del 5%.

La compacidad de la agrupación refleja la estabilidad de la distribución de lectura e indica si las secuencias de lectura se concentran en un intervalo pequeño. Generalmente, una variación estructural real puede absorberse en una gran cantidad de “ruidos ambientales”. Sin embargo, la influencia de los “ruidos ambientales” en el genoma completo es sustancialmente uniforme y, por tanto, parece tener una distribución esencialmente uniforme en toda la secuencia (que, por supuesto, puede verse afectada, por ejemplo, por el contenido de GC (guanina y citosina) y otros). En un sitio donde tiene lugar una variación estructural real, las secuencias de lectura en la agrupación parecen tener una distribución sustancialmente normal. Por tanto, la compacidad, por ejemplo, la varianza, puede reflejar bien la diferencia entre agrupaciones.

(II) En virtud de la correlación lineal de agrupaciones emparejadas - Se calcula la correlación lineal de dos agrupaciones emparejadas, y se eliminan por filtración las agrupaciones emparejadas que tienen una correlación lineal que no cumple un requisito preestablecido R-li. La correlación lineal de un par de agrupaciones puede calcularse usando diversos métodos matemáticos disponibles. Por ejemplo, se calcula el coeficiente de correlación de dos agrupaciones, y cuanto mayor es el coeficiente de correlación, mayor será la correlación lineal. Preferiblemente, R-li puede establecerse como un umbral fijo, por ejemplo, se requiere que el coeficiente de correlación sea mayor que el umbral fijo. O de otro modo, R-li se establece para que sea una tasa de eliminación. Por ejemplo, se requiere que los rangos de los coeficientes de correlación en todas las agrupaciones se encuentren dentro del intervalo más alto preestablecido. Por ejemplo, R-li se establece de manera que los rangos de los coeficientes de correlación en todas las agrupaciones se encuentran dentro del intervalo más alto del 2%-10% y preferiblemente del 5%.

La correlación lineal hace mucho hincapié en la consistencia en la distribución de secuencias de lectura en las agrupaciones emparejadas, es decir, indica si las secuencias de lectura de dos extremos en lecturas emparejadas

tienen la distribución sustancialmente consistente. Por tanto, la correlación lineal puede reflejar mejor la distribución en agrupaciones emparejadas.

En una realización preferida, se logra un buen resultado filtrando las agrupaciones candidatas usando la compacidad, por ejemplo, varianza de las agrupaciones, y la correlación lineal de las agrupaciones en combinación.

(III) En virtud de un conjunto de control de muestras normales - Las agrupaciones emparejadas se alinean con un conjunto de control que comprende una pluralidad de muestras normales, y se eliminan por filtración las agrupaciones emparejadas que tienen un recuento de aciertos de muestras normales que alcanzan un umbral preestablecido V-con. Las muestras normales se refieren a un conjunto de agrupaciones de resultados obtenidas sometiendo otros individuos normales que pertenecen a la misma categoría que el individuo objetivo a los procedimientos de “alineación-agrupamiento-filtración” tal como se describió anteriormente. Para facilidad de alineación, todas las secuencias de lectura en la agrupación pueden fusionarse en una sola, y por tanto se genera un par de valores numéricos fusionados a partir de agrupaciones emparejadas (que se asemeja a un par de lecturas). La alineación se realiza con los pares de valores numéricos fusionados. Al recopilar un conjunto de control que comprende un gran número de muestras normales, puede obtenerse la frecuencia de las agrupaciones de resultados producidas en individuos normales. Si una determinada agrupación de resultados tiene una frecuencia de aparición alta, puede sugerirse que la agrupación de resultados puede resultar de la calidad de la muestra, el procedimiento experimental, el procedimiento de secuenciación o el ruido ambiental, y no indica que la muestra experimente realmente una variación estructural de este tipo en sí misma. Una agrupación de resultados de este tipo es un resultado falso positivo obtenido analizando diferentes muestras con el mismo método, y debe eliminarse. Por tanto, puede reducirse adicionalmente la probabilidad de acontecimientos falsos positivos al filtrar las agrupaciones de resultados usando un conjunto de control, lo que contribuye a la adquisición del resultado de análisis verdadero de la variación estructural. V-con puede determinarse dependiendo de los modos de construcción y de las características de muestras normales. Por ejemplo, la proporción de V-con con respecto al número de muestras normales en el conjunto de control puede ser del 3-10%, y preferiblemente del 5-6%. Por ejemplo, si el conjunto de control contiene 90 muestras normales, se considera que 5 aciertos alcanzan el umbral.

(IV) En virtud de otros parámetros auxiliares - Los parámetros auxiliares incluyen varios parámetros útiles para una confirmación y distinción adicionales de los tipos de anomalías estructurales o para comprender los detalles de las anomalías estructurales, por ejemplo, el número de apareamientos erróneos generados durante la alineación, el número de pares de lecturas que soportan las agrupaciones, el valor de RPU de la región relevante obtenido basándose en * .pair, si las agrupaciones están ubicadas en una región N y otros. Los parámetros auxiliares pueden usarse de las dos formas siguientes. 1. Los parámetros auxiliares se usan como condiciones de filtración. Los requisitos de filtración asociados con los parámetros auxiliares se establecen para eliminar por filtración las agrupaciones que no cumplen los requisitos directamente. 2. Los parámetros auxiliares se usan como base de referencia para ayudar en la determinación. Los parámetros auxiliares se proporcionan con las agrupaciones de resultados, y luego se realiza la determinación mediante análisis manual. Por tanto, el contenido de esta sección puede usarse en la etapa 4 (para filtrar) o en la etapa 5 (para ayudar en el análisis manual). En esta realización, las formas específicas de usar los parámetros auxiliares no están limitadas. A continuación se facilitan a modo de ejemplo algunos parámetros auxiliares y su relación con el análisis de resultados. En la aplicación práctica, los parámetros auxiliares pueden establecerse como condiciones de filtración siguiendo las descripciones a continuación, o como base para ayudar en la determinación a través del análisis manual. Pueden usarse diferentes parámetros auxiliares en combinación o solos.

(1) El número de apareamientos erróneos. - Los apareamientos erróneos promedio de los pares de lectura en las agrupaciones emparejadas generalmente no son más de 1 ó 2, es decir, se permite que cada par de lecturas tenga 1 ó 2 apareamientos erróneos, y preferiblemente no más de 1 apareamiento erróneo. No es necesario tener en cuenta el parámetro si el requisito de apareamiento en la alineación se establece basándose de esto. Si el establecimiento en la alineación es menos estricto, por ejemplo, se establece que se permite que se produzcan 2 apareamientos erróneos, pueden realizarse adicionalmente filtración o determinación usando el parámetro durante la adquisición de la agrupación de resultados, por ejemplo, se establece que se permite que se produzca 1 apareamiento erróneo como promedio.

(2) El número de pares de lectura que soportan las agrupaciones, es decir, el número de pares de lectura contenidos en agrupaciones emparejadas - En principio, será mejor si el parámetro es más alto. En general, puede establecerse que la base de juicio concuerde con o sea ligeramente menor (por ejemplo, tomar un valor integral) que el valor normalizado de la profundidad de secuenciación, donde el valor normalizado de la profundidad de secuenciación = profundidad de secuenciación x (intervalo de influencia de L-lib en puntos de ruptura/L-lib) x (extensiones promedio en dos extremos de agrupaciones emparejadas/L-lib). El “intervalo de influencia de L-lib en los puntos de ruptura” generalmente es más alto que “la suma de las extensiones en dos extremos de agrupaciones emparejadas”, y generalmente fluctúa a un promedio que es 2 veces L-lib, por ejemplo, fluctúa entre 1 -4 veces L-lib. Cuando se establece específicamente, el intervalo del parámetro puede ampliarse o reducirse adecuadamente según se desee por la situación práctica.

(3) El valor de RPU de la región relevante obtenido basándose en * .pair - Los diferentes tipos de anomalías

estructurales generalmente tienen una influencia diferente sobre las RPU. Por ejemplo, en el caso de la translocación equilibrada, las RPU en dos lados laterales de los puntos de ruptura no varían significativamente; sin embargo, en el caso de anomalías estructurales de tipo de deleción o repetición, las RPU de la región entre los puntos de ruptura disminuyen o aumentan considerablemente. Por tanto, el valor de RPU de una región relevante puede usarse más para confirmar o ayudar en la determinación de la aparición de anomalías estructurales cromosómicas.

Por ejemplo, para las agrupaciones que contienen el primer tipo de pares de lecturas, si la translocación equilibrada se determina según la relación entre los pares de lecturas en las agrupaciones (véase la sección I de la etapa 5 a continuación para obtener detalles), la variación de RPU en dos lados laterales de los puntos de ruptura de un promedio no es más alta que V-rm; y si la translocación desequilibrada se determina según la relación entre los pares de lecturas en las agrupaciones (véase la sección I de la etapa 5 a continuación para obtener detalles), las RPU en el lado de los puntos de ruptura que se apartan de las agrupaciones de resultados están por debajo del promedio, y la variación es mayor que V-rm.

Para las agrupaciones que contienen el segundo tipo de pares de lecturas, la RPU de la región entre los puntos de ruptura está por encima del promedio, y la variación es mayor que V-rm.

Para las agrupaciones que contienen el tercer tipo de pares de lecturas, las RPU de la región entre los puntos de ruptura están por debajo del promedio, y la variación es mayor que V-rm.

Cuando las RPU se usan como base para ayudar en la determinación a través del análisis manual, las RPU de una región relevante pueden presentarse en forma de gráfico, tabla u otra forma fácilmente identificable. Alternativamente, la variación de RPU en toda la región se presenta en forma de un gráfico, una tabla o similar, para promover la comprensión de las condiciones generales por parte del operario.

(4) Si la agrupación está ubicada en una región N - Empíricamente, la alineación de lecturas en las proximidades de la región N (que comprende áreas de centrómero y telómero) es más compleja que otras regiones. Si las agrupaciones obtenidas no están ubicadas en la región N, se considera que la determinación puede realizarse según la información adquirida. Si las agrupaciones obtenidas están ubicadas en la región N, puede requerirse una validación más cuidadosa. Por ejemplo, la determinación final se realiza mediante el uso combinado de condiciones de filtración y parámetros auxiliares, o conjuntamente con otros datos externos, por ejemplo, el fenotipo del individuo objetivo y/o el resultado de una secuenciación más precisa (por ejemplo, secuenciación de Sanger) de los puntos de ruptura.

Etapa 5. Análisis de datos de las agrupaciones de resultados filtradas

La presencia de las agrupaciones de resultados obtenidas después de la filtración refleja la aparición potencial de tipos correspondientes de anomalías estructurales cromosómicas. Por tanto, esta etapa no es necesaria cuando sólo se requiere encontrar anomalías estructurales potenciales. Con el propósito de obtener información más detallada con respecto a anomalías estructurales, las agrupaciones de resultados obtenidas pueden someterse adicionalmente a un análisis de datos. Dependiendo de diferentes tipos de agrupaciones de resultados, pueden adoptarse los siguientes modos de análisis.

(I) Anomalía estructural cromosómica de tipo translocación (el primer tipo de lecturas)

Se buscan las agrupaciones de resultados que contienen el primer tipo de pares de lecturas, y si dos secuencias de lectura adyacentes tienen posiciones opuestas en pares de lecturas respectivos, el intervalo entre posiciones en el que las dos secuencias de lectura se aparean se toma como un intervalo de puntos de ruptura. Esta situación se correlaciona generalmente con una translocación equilibrada, en la que las secuencias de lectura en la misma agrupación se distribuyen en dos lados laterales de los puntos de ruptura.

Si no existen tales secuencias de lectura, se adquiere la posición de la secuencia de lectura más interior, y el intervalo obtenido extendiéndose hacia el interior desde la posición en una longitud preestablecida se toma como un intervalo de puntos de ruptura. La secuencia de lectura más interior significa que cuando la agrupación incluye exclusivamente secuencias de lectura de extremo izquierdo, la secuencia de lectura más a la derecha es la secuencia de lectura más interior; y cuando la agrupación incluye exclusivamente secuencias de lectura de extremo derecho, la secuencia de lectura más a la izquierda es la secuencia de lectura más interior. Esta situación se correlaciona generalmente con una translocación desequilibrada, en la que las secuencias de lectura en la misma agrupación se distribuyen en un lado lateral de los puntos de ruptura. La extensión del intervalo de puntos de ruptura que se extiende desde la secuencia de lectura más interior puede determinarse según L-lib, L-r1/L-r2, la profundidad de secuenciación, y otros, y puede ser por ejemplo 0,5-2 veces L-lib y generalmente no más de 2 veces L-lib.

La figura 2 muestra una situación de translocación equilibrada. Un par obtenido de agrupaciones de resultados (sólo se ilustran dos secuencias de lectura en cada agrupación, y se considera que las demás se omiten) tiene una distribución tal como se muestra en la figura 2, una agrupación de resultados está ubicada en las proximidades de la

posición pa en el cromosoma chra, y la agrupación de resultados emparejada con la misma está ubicada en las proximidades de la posición pb en el cromosoma chrb. Dado que en la agrupación en chra, Read1 es una secuencia de lectura de extremo izquierdo del fragmento de cromosoma del que se deriva, y la Read2 adyacente es una secuencia de lectura de extremo derecho del fragmento de cromosoma del que se deriva, se considera que el punto de ruptura pa de chra está ubicado entre Read1 y Read2, y el mismo análisis se aplica a chrb.

Basándose en el análisis de datos anterior, la salida de datos de resultado para una potencial anomalía estructural de tipo translocación puede incluir numeraciones de dos cromosomas (en los que están ubicadas respectivamente las agrupaciones de resultados) que tienen potencialmente una anomalía estructural de tipo translocación, intervalos de posición de dos extremos de agrupaciones emparejadas de resultados (intervalos de posición de límites de dos extremos de las agrupaciones en los dos cromosomas, a partir de los que pueden obtenerse de manera correspondiente las extensiones de los dos extremos de las agrupaciones), intervalo de puntos de ruptura obtenidos después del análisis, y otros. Los parámetros relevantes generados durante la filtración y otros parámetros auxiliares también pueden emitirse junto a, por ejemplo, la compacidad respectiva de un par de agrupaciones de resultados, el grado de correlación lineal del par de agrupaciones de resultados, el número de pares de lecturas que soportan el par de agrupaciones de resultados, y un gráfico y una tabla que presentan la variación en RPU en dos lados laterales de los puntos de ruptura.

(II) Anomalía estructural cromosómica de tipo de repetición en tándem (el segundo tipo de pares de lecturas)

Se buscan las agrupaciones de resultados que contienen el segundo tipo de pares de lecturas, el intervalo entre dos posiciones apareadas con las mismas que son las más alejadas en distancia en las agrupaciones emparejadas se toma como un intervalo de aparición de repetición, y el intervalo obtenido extendiéndose hacia el exterior desde las dos posiciones respectivamente en una longitud preestablecida que es por ejemplo 0,5-2 veces L-lib se toma como un intervalo de puntos de ruptura (los puntos de inicio y terminación del fragmento de repetición).

La figura 3 muestra una situación de repetición en tándem. Dos extremos de agrupaciones emparejadas de resultados (sólo se ilustra una secuencia de lectura en cada agrupación, y se considera que las demás se omiten) se encuentran ambas dentro del intervalo entre los puntos de inicio y terminación del fragmento de repetición, y por tanto se considera que los puntos de inicio y terminación del fragmento de repetición están ubicados en un intervalo que se extiende hacia el exterior dese las secuencias de lectura (en el que las dos secuencias de lectura no pertenecen necesariamente a un par de lecturas) en la posición más exterior de los dos extremos de las agrupaciones.

En comparación con la anomalía estructural de tipo translocación, la anomalía estructural de tipo de repetición tiene sustancialmente los mismos tipos de salida de datos de resultado, excepto porque las numeraciones de los cromosomas en dos extremos de las agrupaciones son las mismas, y también pueden emitirse datos que indican la longitud estimada del fragmento de repetición.

(III) Anomalía estructural cromosómica de tipo deleción (el tercer tipo de pares de lecturas)

Se buscan las agrupaciones de resultados que contienen el tercer tipo de pares de lecturas, el intervalo entre dos posiciones apareadas con las mismas que son las más próximas en distancia en las agrupaciones emparejadas se toma como un intervalo de aparición de deleción, y el intervalo obtenido extendiéndose hacia el interior desde las dos posiciones respectivamente en una longitud preestablecida que es por ejemplo 0,5-2 veces L-lib se toma como un intervalo de puntos de ruptura (los puntos de inicio y terminación del fragmento delecionado).

La figura 4 muestra una situación de deleción de fragmento. Dos extremos de agrupaciones emparejadas de resultados (sólo se ilustra una secuencia de lectura en cada agrupación, y se considera que las demás se omiten) se encuentran ambas fuera del intervalo entre los puntos de inicio y terminación del fragmento delecionado, y por tanto se considera que los puntos de inicio y terminación del fragmento delecionado están ubicados en un intervalo que se extiende hacia el interior desde las secuencias de lectura (en el que las dos secuencias de lectura no pertenecen necesariamente a un par de lecturas) que son las más próximas en los dos extremos de las agrupaciones.

En comparación con la anomalía estructural de tipo de repetición, la anomalía estructural de tipo deleción tiene sustancialmente los mismos tipos de salida de datos de resultado, excepto porque los datos de salida que indican la longitud estimada del fragmento entre los puntos de ruptura representan la longitud del fragmento delecionado.

Etapa 6. Ensamblado de puntos de ruptura

Para reducir adicionalmente el intervalo de puntos de ruptura, puede realizarse un ensamblado de puntos de ruptura usando los datos de *.unmap. Por ejemplo, se obtienen secuencias de lectura de un solo extremo (que pueden aparearse con un solo extremo a la secuencia de referencia, y pueden designarse como *.sin durante la alineación) en un intervalo establecido periféricamente con respecto al intervalo de puntos de ruptura determinado (por ejemplo 0,5-2 veces L-lib), y secuencias de lectura emparejadas con las mismas se recuperan de *.unmap como secuencias de corrección. Todas las secuencias de corrección se truncan para dar N secciones, y N es preferiblemente 2.

Después, las subsecuencias obtenidas tras truncar las secuencias de corrección se alinean de nuevo con la secuencia de referencia. La región de puntos de ruptura se ensambla según el resultado de apareamiento normal.

En el uso práctico, el valor de N puede establecerse de manera racional según la longitud de Lr1/Lr2. Cuando la secuencia longitud es menor de 25 pb, se hace que la tasa de alineación única disminuya considerablemente. Por consiguiente, cuando se establece el valor de N, puede tenerse en cuenta que la longitud de la subsecuencia truncada no es, o no de manera evidente, menor de 25 pb.

Después del ensamblado de puntos de ruptura, el intervalo de puntos de ruptura puede reducirse eficazmente. Basándose en esto, puede prepararse adicionalmente una sonda según el intervalo de posiciones en el que residen los puntos de ruptura, y las posiciones exactas de los puntos de ruptura pueden obtenerse finalmente por medio de otra secuenciación exacta, por ejemplo una secuenciación de Sanger, para llevar a cabo adicionalmente el estudio sobre puntos de ruptura. Si no es necesario reducir el intervalo de puntos de ruptura, esta etapa puede omitirse.

Los expertos habituales en la técnica pueden entender que la totalidad o parte de las etapas de los métodos proporcionados en las realizaciones anteriores pueden realizarse mediante hardware relevante según instrucciones de un programa que puede almacenarse en un medio de almacenamiento legible por ordenador, incluyendo una memoria de sólo lectura, una memoria de acceso aleatorio, un disco magnético o un disco óptico.

Según otro aspecto de la presente invención, se proporciona adicionalmente un dispositivo para detectar anomalías estructurales cromosómicas, que incluye una unidad de entrada de datos, configurada para introducir datos; una unidad de salida de datos, configurada para emitir datos; una unidad de almacenamiento, configurada para almacenar datos y que contiene un programa ejecutable en la misma; y un procesador, en conexión de datos con la unidad de entrada de datos, la unidad de salida de datos, y la unidad de almacenamiento, y configurado para ejecutar el programa ejecutable almacenado en la unidad de almacenamiento, en el que la ejecución del programa incluye realizar la totalidad o parte de las etapas de los métodos proporcionados en las realizaciones anteriores.

A continuación en el presente documento, el resultado operativo de un método de detección específico según la presente invención se describe con detalle en relación con un individuo objetivo específico. En el procedimiento de detección, los parámetros específicos usados se establecen de la siguiente manera.

1. L-lib es 500 pb, y se emplea una secuenciación de PE50 (secuenciación de extremo de par, en la que L-r1 y L-r2 son aproximadamente 50 pb).

2. Como secuencia de referencia se usa HG19 a partir de NCBI, y el resultado de secuenciación se alinea mediante software SOAP.

3. V-lib es ± 45 pb, V-rm de RPK es el 20%, V-cl es 10 Kpb (la distancia entre agrupaciones se define como la distancia entre dos secuencias de lectura más cercanas), el número mínimo de lecturas en la agrupación es 2, R-va se establece de manera que los rangos de las varianzas en todas las agrupaciones se encuentran dentro del intervalo más bajo del 5% (en el cálculo de la varianza, se excluyen las secuencias de lectura ubicadas en dos extremos en el intervalo del 20% en longitud de la agrupación), R-li se establece de manera que los rangos de los coeficientes de correlación en todas las agrupaciones se encuentran dentro del intervalo más bajo del 5%, el conjunto de control incluye 90 muestras normales, y V-con es 5.

Ejemplo experimental I

Este ejemplo facilita un estudio sobre una familia con síndrome de Cri du Chat. En este ejemplo, los dos individuos objetivo pertenecen a la misma familia, en el que “FA” representa al padre, y “SON” representa al hijo.

1. Se realizó una secuenciación del genoma completo respectivamente en los dos individuos objetivo con un multiplicador bajo, en la que la profundidad de secuenciación de “FA” fue de 2,2, y de 3,1 para “SON”.

2. Después, los resultados de secuenciación de los dos individuos objetivo se alinearon respectivamente con la secuencia de referencia HG19 usando el software de alineación SOAP, para obtener dos archivos FA.sin y SON.sin.

3. Los dos archivos FA.sin y SON.sin se agruparon, filtraron y analizaron, para obtener las siguientes agrupaciones de resultados y parámetros de salida relevantes:

“FA”:

numeración de dos cromosomas en los que residen agrupaciones emparejadas de resultados: chr12, chr5

intervalos de posición de dos extremos de agrupaciones emparejadas de resultados: 14779615-14780233, 23314785-23314205

extensiones de dos extremos de agrupaciones emparejadas de resultados: 618, 580

número de pares de lecturas que soportan el par de agrupaciones de resultados: 5

compacidad (varianza) de los extremos izquierdo y derecho: 90,59, 87,01

si las agrupaciones están ubicadas en una región N: no

intervalo de puntos de ruptura: chr12: 14779968-14780233, chr5: 23314205-23314455

variación en RPK de regiones relevantes en los cromosomas: en la figura 6, el eje horizontal representa la posición (unidad: 10 Kpb) en el cromosoma, y el eje longitudinal representa la RPK. La curva se representa gráficamente basándose en los datos de FA.pair, y pa y pb representan las posiciones de puntos de ruptura. Puede observarse a partir de la figura 6 que la variación en RPK de “FA” es insignificante.

“SON”:

numeración de dos cromosomas en los que residen agrupaciones emparejadas de resultados: chr12, chr5

intervalos de posición de dos extremos de agrupaciones emparejadas de resultados: 14779618-14779968, 23314455-23314830

extensiones de dos extremos de agrupaciones emparejadas de resultados: 350, 375

número de pares de lecturas que soportan el par de agrupaciones de resultados: 6

compacidad (varianza) de los extremos izquierdo y derecho: 22,43, 18,44

si las agrupaciones están ubicadas en una región N: no

intervalo de puntos de ruptura: chr12: por encima de 14779968, chr5: por debajo de 23314455

variación en RPK de regiones relevantes en el cromosoma: en la figura 7, el eje horizontal representa la posición (unidad: 10 Kpb) en el cromosoma, y el eje longitudinal representa la RPK. La curva se representa gráficamente basándose en los datos de SON.pair, y pa y pb representan las posiciones de puntos de ruptura. Puede observarse a partir de la figura 7 que la variación en RPK de “SON” es evidente. A partir de la revisión de la RPK calculada puede saberse que la RPK en el brazo corto del cromosoma 5 de SON sólo es 0,5 veces el promedio, y la RPK en el brazo corto del cromosoma 12 es 0,5 veces mayor que el promedio.

A partir de los resultados de análisis puede determinarse indudablemente que se produce una translocación equilibrada en “FA” y se produce una translocación desequilibrada en “SON”. El intervalo de puntos de ruptura analizado a partir del resultado de “FA” se encuentra dentro de 300 pb. Para llevar a cabo adicionalmente el estudio sobre posiciones de puntos de ruptura, se elimina una secuencia correspondiente de la secuencia de referencia HG19, y se designa un cebador, para una secuenciación de Sanger y una validación mediante qPCR. Las posiciones exactas obtenidas finalmente de los puntos de ruptura son Chr12: 14780019, Chr5: 23314435.

Ejemplo experimental II

Este ejemplo facilita un estudio sobre cardiopatía congénita. En este ejemplo, el individuo objetivo es un paciente con cardiopatía congénita, y designado como “XX”.

1. Se realizó una secuenciación del genoma completo en el individuo objetivo con un multiplicador bajo, en la que la profundidad de secuenciación fue de 2,7.

2. Después, el resultado de secuenciación se alineó con la secuencia de referencia HG19 usando el software de alineación SOAP, para obtener XX.sin.

3. XX.sin se agrupó, filtró y analizó, para obtener las siguientes agrupaciones de resultados y parámetros de salida relevantes:

“XX”:

numeración de dos cromosomas en los que residen agrupaciones emparejadas de resultados: chr14, chr14

intervalos de posición de dos extremos de agrupaciones emparejadas de resultados: 73557040-73557288, 73670432-73670682

longitud estimada de fragmento de repetición: 113392

extensiones de dos extremos de agrupaciones emparejadas de resultados: 248, 250

número de pares de lecturas que soportan el par de agrupaciones de resultados: 4

compacidad (varianza) de los extremos izquierdo y derecho: 100,63, 100,59

si las agrupaciones están ubicadas en una región N: no

intervalo de puntos de ruptura: chr14: 73556540-73557040, chrl4: 73670682-73671182 (donde el tamaño del intervalo se estima basándose en 1 vez de L-lib, es decir 500 pb)

A partir del resultado de análisis puede determinarse indudablemente que se produce una repetición de aproximadamente 113 Kpb en longitud en el cromosoma 14 de “XX”, y la repetición se produce en tándem. Para llevar a cabo adicionalmente el estudio sobre posiciones de puntos de ruptura, se elimina una secuencia correspondiente de la secuencia de referencia HG19, y se designa un cebador, para una secuenciación de Sanger y una validación mediante qPCR. La razón de multiplicación de qPCR es mayor de 1, lo que sugiere una repetición. Las posiciones exactas de los puntos de ruptura obtenidas finalmente a partir de la secuenciación de Sanger son Chr14: 73557008, Chr14: 73670820, lo que confirma que sí se produce una repetición de 113812 pb en el cromosoma 14 de “XX”, y el fragmento de repetición se inserta en el extremo del fragmento en tándem.

REIVINDICACIONES

1. Método implementado por ordenador para detectar anomalías estructurales cromosómicas, que comprende:

adquirir un resultado de secuenciación del genoma completo de un individuo objetivo o individuos objetivo, en el que el resultado de secuenciación incluye múltiples pares de lecturas, consistiendo cada par de lecturas en dos secuencias de lectura ubicadas respectivamente en dos extremos de un fragmento de cromosoma determinado, y cada par de lecturas se deriva por separado de las cadenas positiva y negativa del fragmento de cromosoma correspondiente, o tanto de la cadena positiva como de la negativa del fragmento de cromosoma correspondiente;

alinear el resultado de secuenciación con una secuencia de referencia, para obtener un conjunto de apareamiento anómalo, en el que el conjunto de apareamiento anómalo incluye un primer tipo de pares de lecturas en el que las dos secuencias de lectura en el primer tipo de par de lecturas se aparean respectivamente con cromosomas diferentes de la secuencia de referencia;

agrupar las secuencias de lectura en el conjunto de apareamiento anómalo basándose en las posiciones apareadas con las mismas, en el que cada agrupación contiene secuencias de lectura de un sólo extremo procedentes de un grupo de pares de lecturas, y las secuencias de lectura del otro extremo correspondientes residen en otra agrupación;

filtrar las agrupaciones resultado del agrupamiento, incluyendo calcular la compacidad de cada agrupación y eliminar por filtración las agrupaciones que tienen una compacidad que no cumple un requisito de compacidad preestablecido (R-va) y las agrupaciones emparejadas con ellas, en el que la compacidad se indica por la varianza, en el que se calcula la varianza de la posición de cada lectura en una agrupación con respecto al centro o al centro de gravedad de la agrupación, y en el que cuanto menor es la varianza, mayor es la compacidad; y

obtener las agrupaciones de resultados filtradas que contienen el primer tipo de pares de lecturas, y basándose en dichas agrupaciones de resultados filtradas determinar la aparición de anomalía estructural cromosómica de tipo translocación.

2. Método según la reivindicación 1, en el que

la filtración de las agrupaciones resultado del agrupamiento comprende además:

calcular la correlación lineal de dos agrupaciones emparejadas, y filtrar las agrupaciones emparejadas que tienen una correlación lineal que no cumple un requisito preestablecido R-li; y/o

alinear agrupaciones emparejadas con un conjunto de control preestablecido que comprende una pluralidad de muestras normales, y filtrar las agrupaciones emparejadas que tienen un recuento de aciertos de muestras normales que alcanza un umbral preestablecido V-con.

3. Método según la reivindicación 1, que comprende además:

buscar las agrupaciones de resultados que contienen el primer tipo de pares de lecturas, si dos secuencias de lectura adyacentes tienen posiciones opuestas en pares de lecturas respectivos, tomar el intervalo entre posiciones en que las dos secuencias de lectura se aparean como un intervalo de puntos de ruptura; y si tales secuencias de lectura no existen, adquirir la posición de la secuencia de lectura más interior, y tomar el intervalo obtenido extendiéndose hacia el exterior desde la posición en una longitud preestablecida como un intervalo de puntos de ruptura.

4. Método según la reivindicación 1, en el que

el conjunto de apareamiento anómalo comprende además un segundo tipo de pares de lecturas según la descripción en que dos secuencias de lectura en el segundo tipo de par de lecturas se aparean con el mismo cromosoma de la secuencia de referencia, pero la longitud L-pr del fragmento de cromosoma calculada según las posiciones apareadas con las mismas es negativa; y

las agrupaciones de resultados filtradas que contienen el segundo tipo de pares de lecturas se obtienen adicionalmente para determinar la aparición de anomalía estructural cromosómica de tipo de repetición en tándem.

5. Método según la reivindicación 4, que comprende además:

buscar las agrupaciones de resultados que contienen el segundo tipo de pares de lecturas, tomar el

intervalo entre dos posiciones apareadas con las mismas que son las más alejadas en distancia en las agrupaciones emparejadas como un intervalo de aparición de repetición, y tomar el intervalo obtenido extendiéndose hacia el exterior desde las dos posiciones respectivamente en una longitud preestablecida como un intervalo de puntos de ruptura.

6. Método según la reivindicación 1, en el que

el conjunto de apareamiento anómalo comprende además un tercer tipo de pares de lecturas según la descripción en que dos secuencias de lectura en el tercer tipo de par de lecturas se aparean con el mismo cromosoma de la secuencia de referencia, pero la longitud L-pr del fragmento de cromosoma calculada según las posiciones apareadas con las mismas es mayor que un tamaño de biblioteca L-lib, y la desviación está por encima de un umbral preestablecido V-lib, en el que V-lib es preferiblemente del 5% x L-lib al 15% x L-lib, y más preferiblemente del 10% x L-lib; y

las agrupaciones de resultados filtradas que contienen el tercer tipo de pares de lecturas se obtienen adicionalmente para determinar la aparición de anomalía estructural cromosómica de tipo deleción.

7. Método según la reivindicación 6, que comprende además:

buscar las agrupaciones de resultados que contienen el tercer tipo de pares de lecturas, tomar el intervalo entre dos posiciones apareadas con las mismas que son las más próximas en distancia en las agrupaciones emparejadas como un intervalo de aparición de deleción, y tomar el intervalo obtenido extendiéndose hacia el interior desde las dos posiciones respectivamente en una longitud preestablecida como un intervalo de puntos de ruptura.

8. Método según una cualquiera de las reivindicaciones 1 a 7, en el que

la alineación del resultado de secuenciación con la secuencia de referencia comprende además:

adquirir un conjunto de apareamiento normal, en el que el conjunto de apareamiento normal incluye pares de lecturas según la descripción en que dos secuencias de lectura en el par de lecturas se aparean con el mismo cromosoma de la secuencia de referencia, y la relación de las cadenas positiva y negativa de las posiciones apareadas con las mismas es conforme con la de en el par de lecturas, y la longitud L-pr del fragmento de cromosoma calculada según las posiciones apareadas con las mismas se desvía del tamaño de biblioteca L-lib usado en la secuenciación en un valor menor que el umbral preestablecido V-lib, en el que V-lib es preferiblemente desde el 5% x L-lib hasta el 15% x L-lib, y más preferiblemente del 10% x L-lib; y

calcular estadísticamente el número RPU de lecturas en el conjunto de apareamiento normal por longitud unitaria, y adquirir la variación de RPU en comparación con el valor promedio, para ayudar en la determinación de la aparición de anomalías estructurales, en el que la variación de RPU en comparación con el valor promedio se indica preferiblemente si la variación de RPU está por encima de un umbral preestablecido V-rm, y V-rm es preferiblemente del 10-30% y más preferiblemente del 20%.

9. Método según una cualquiera de las reivindicaciones 1 a 7,

en el que la alineación del resultado de secuenciación con la secuencia de referencia comprende además

adquirir un conjunto sin apareamiento, que incluye secuencias de lectura que no pueden aparearse con la secuencia de referencia, e incluye secuencias de lectura sin apareamiento emparejadas o secuencias de lectura sin apareamiento de un sólo extremo; y

tras obtener las agrupaciones de resultados, el método comprende además

adquirir secuencias de lectura de un sólo extremo en un intervalo establecido periféricamente con respecto al intervalo de puntos de ruptura determinado, recuperar secuencias de lectura emparejadas con las mismas del conjunto sin apareamiento como secuencias de corrección, truncar todas las secuencias de corrección para dar N secciones, en el que N es preferiblemente 2, alinear de nuevo las subsecuencias obtenidas tras truncar las secuencias de corrección con la secuencia de referencia, y ensamblar la región de puntos de ruptura según el resultado de apareamiento normal.

10. Método según una cualquiera de las reivindicaciones 1 a 7, en el que

durante el cálculo de la compacidad de cada agrupación, del 5 al 25% de las secuencias de lectura ubicadas en dos extremos de la agrupación se excluyen del cálculo; y/o

R-va se establece de manera que los rangos de las varianzas en todas las agrupaciones se encuentran dentro del intervalo más bajo del 2%-10% y preferiblemente del 5%.

11. Método según la reivindicación 2, en el que

durante el cálculo de la correlación lineal de dos agrupaciones emparejadas, la correlación lineal se indica mediante el coeficiente de correlación, y R-li se establece de manera que los rangos de los coeficientes de correlación en todas las agrupaciones se encuentran dentro del intervalo más alto del 2%-10% y preferiblemente del 5%; y/o

la proporción de V-con con respecto al número de muestras normales en el conjunto de control es del 3-10%, y preferiblemente del 5-6%.

12. Método según la reivindicación 1, en el que

el tamaño de biblioteca L-lib usado en la secuenciación es mayor que o igual a 300 pb y preferiblemente de 500 pb o 5 Kpb, y/o

la longitud de las secuencias de lectura es mayor que o igual a 25 pb y preferiblemente dentro de ± el 10% de 50 pb.

13. Dispositivo para detectar anomalías estructurales cromosómicas, que comprende:

una unidad de entrada de datos, configurada para introducir datos;

una unidad de salida de datos, configurada para emitir datos;

una unidad de almacenamiento, configurada para almacenar datos, y que contiene un programa ejecutable en la misma; y

un procesador, en conexión de datos con la unidad de entrada de datos, la unidad de salida de datos y la unidad de almacenamiento, y configurado para ejecutar el programa ejecutable, en el que la ejecución del programa incluye realizar el método según una cualquiera de las reivindicaciones 1 a 12.

14. Medio de almacenamiento legible por ordenador, configurado para almacenar un programa ejecutable por un ordenador, en el que la ejecución del programa comprende realizar el método según una cualquiera de las reivindicaciones 1 a 12.

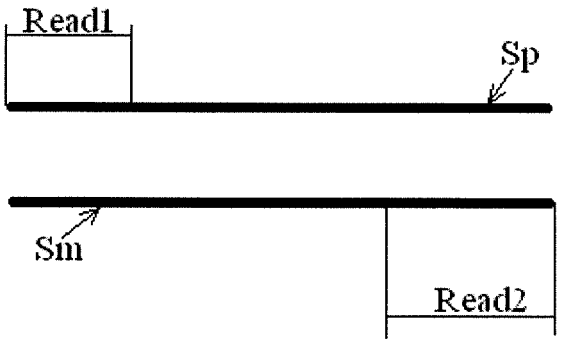

FIG. 1

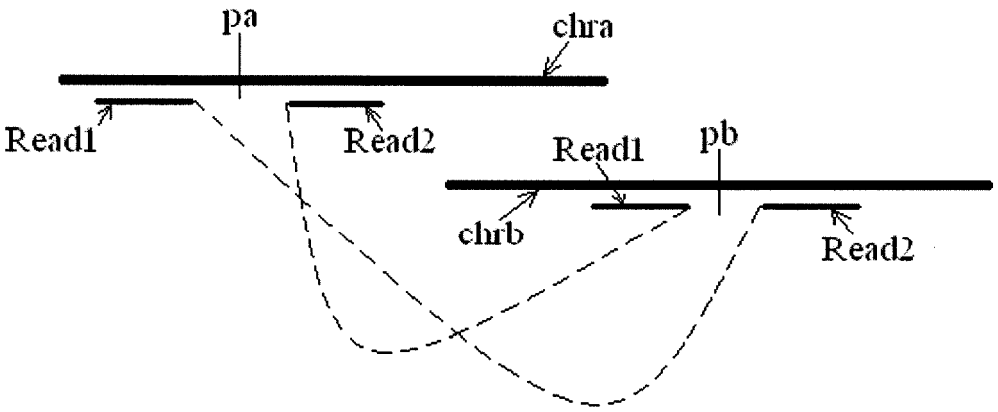

FIG. 2

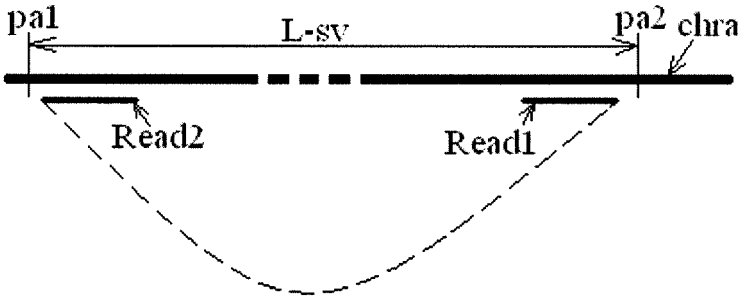

FIG. 3

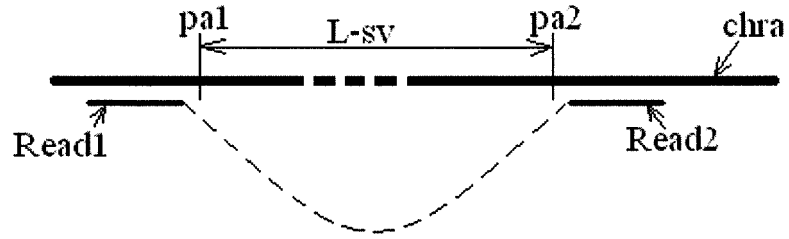

FIG. 4

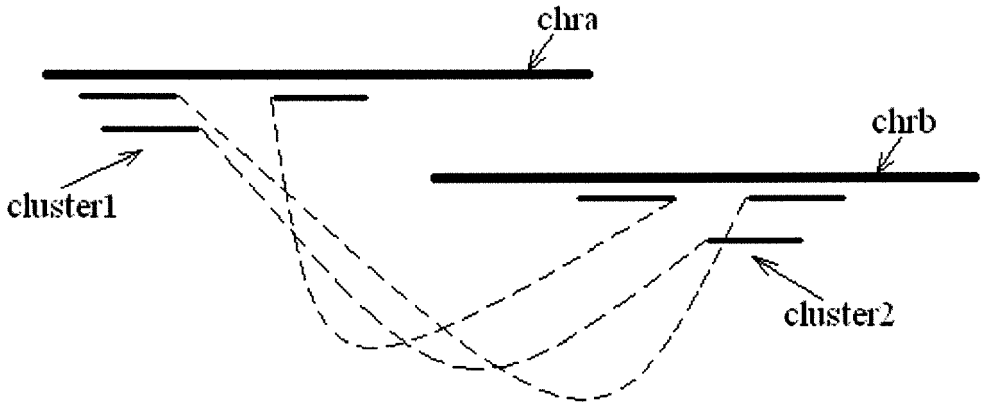

FIG. 5

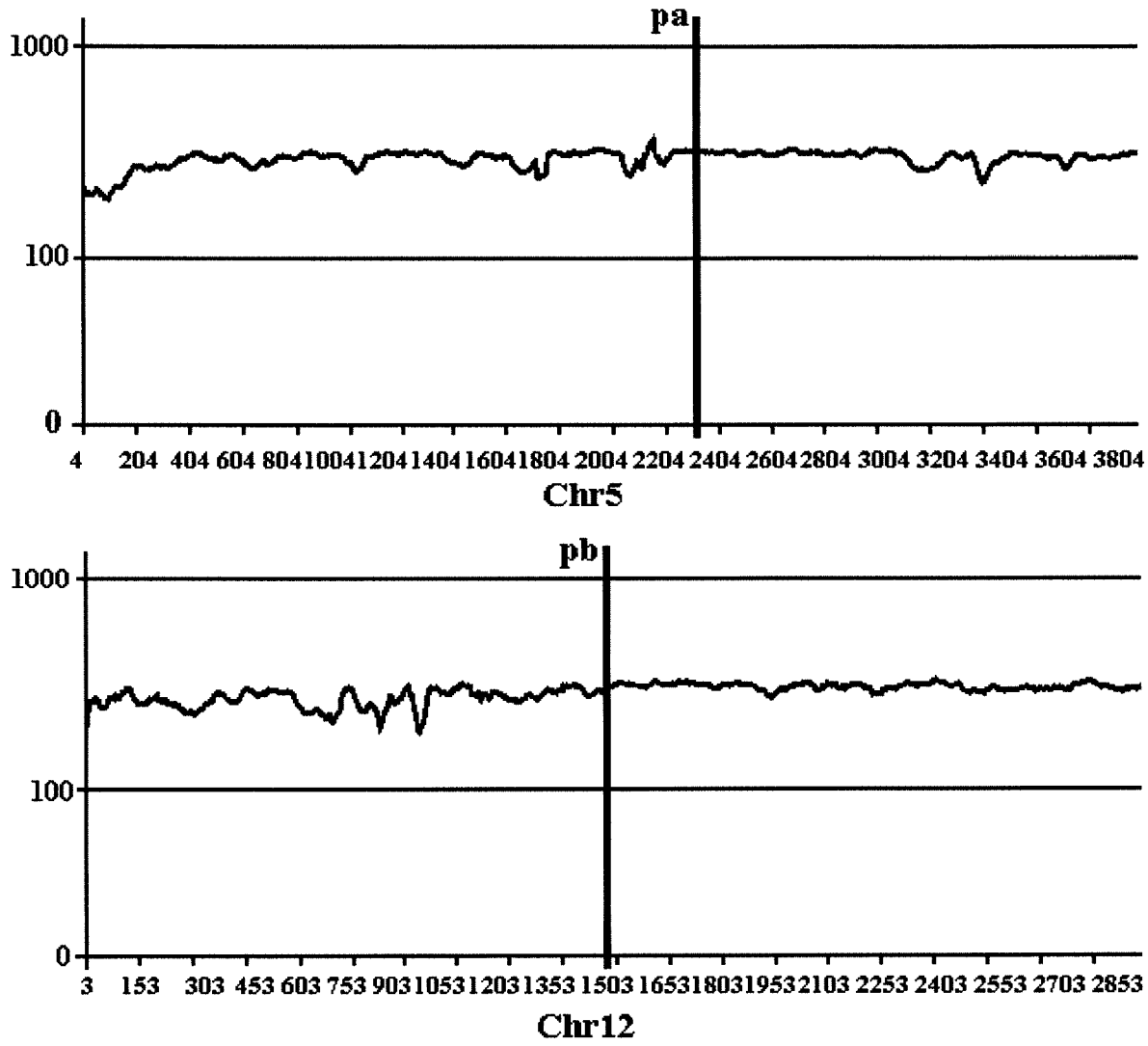
pa
1000
100
0
4 204 404 604 804 1004 1204 1404 1604 1804 2004 2204 2404 2604 2804 3004 3204 3404 3604 3804
Chr5
pb
1000
100
0
3 153 303 453 603 753 903 1053 1203 1353 1503 1653 1803 1953 2103 2253 2403 2553 2703 2853
Chr12

FIG. 6

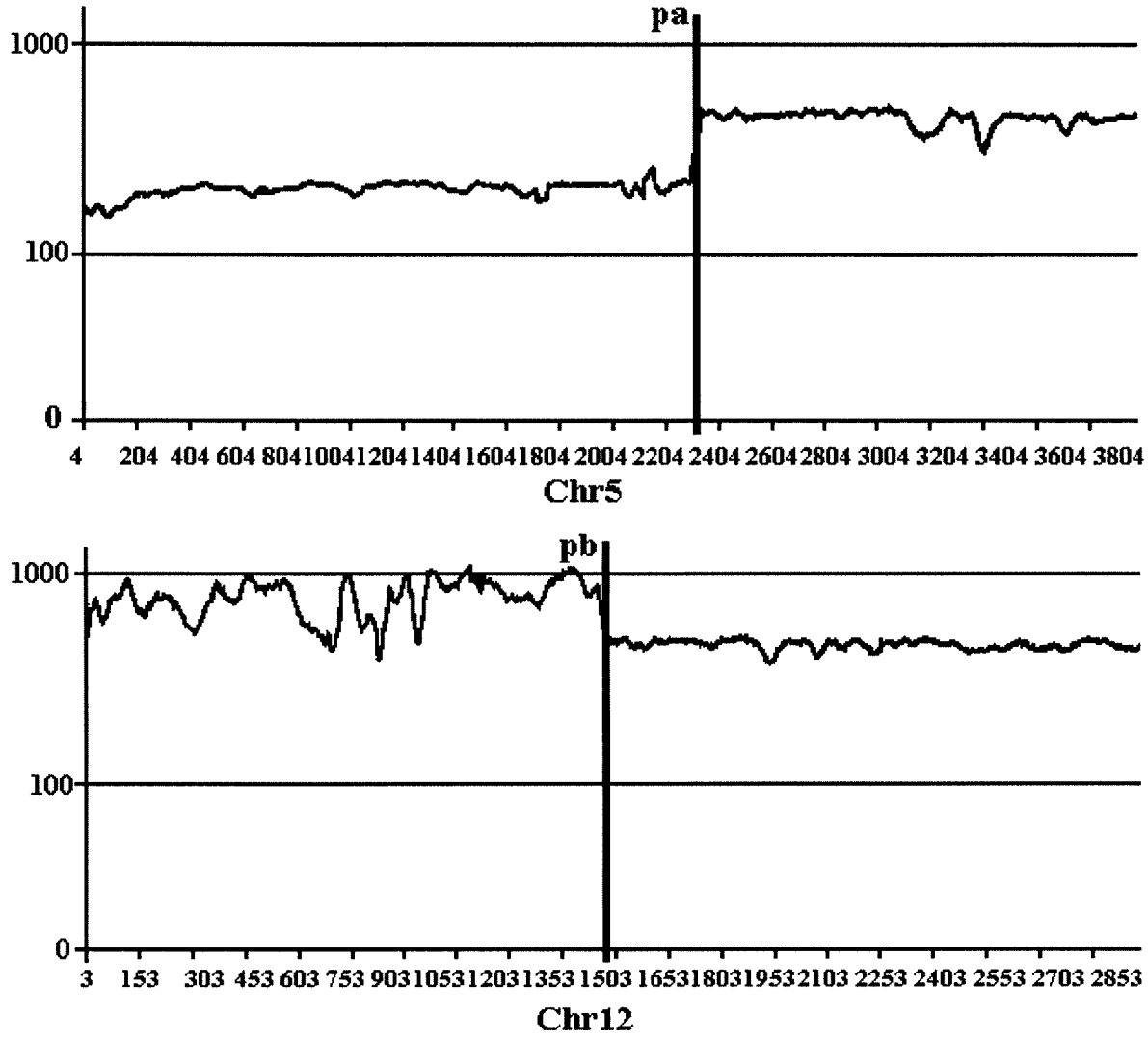
pa
1000
100
0
4 204 404 604 804 1004 1204 1404 1604 1804 2004 2204 2404 2604 2804 3004 3204 3404 3604 3804
Chr5
pb
1000
100
0
3 153 303 453 603 753 903 1053 1203 1353 1503 1653 1803 1953 2103 2253 2403 2553 2703 2853
Chr12

FIG. 7