(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 741 966**

(51) Int. CI.:

C12Q 1/6869 (2008.01)
C12Q 1/6809 (2008.01)
G16B 30/00 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **31.12.2011 PCT/CN2011/002244**

(87) Fecha y número de publicación internacional: **04.07.2013 WO13097062**

(96) Fecha de presentación y número de la solicitud europea: **31.12.2011 E 11878559 (1)**

(97) Fecha y número de publicación de la concesión europea: **31.07.2019 EP 2772549**

(54) Título: **Método para detectar una variación genética**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
12.02.2020

(73) Titular/es:

BGI GENOMICS CO., LTD. (100.0%)
21 Hongan 3rd Street, BGI Park Building 7 Floor 7-14, Yantian District
Shenzhen , CN

(72) Inventor/es:

CHEN, SHENGPEI;
ZHANG, CHUNLEI;
CHEN, FANG;
XIE, WEIWEI;
PAN, XIAOYU;
WANG, JIAN;
WANG, JUN;
YANG, HUANMING y
ZHANG, XIUQING

(74) Agente/Representante:

SÁEZ MAESO, Ana

DESCRIPCIÓN

Método para detectar una variación genética.

Campo Técnico

La presente invención se refiere al campo de la detección de variación genética, y en particular, a la detección de una variación del número de copias, por ejemplo, microdeleción/ microduplicación y aneuploidía.

Antecedentes de la Técnica

Una variación del número de copias (CNV) se refiere a una mutación submicroscópica de un fragmento de ADN en un intervalo de kb a Mb, que está marcado por el aumento o disminución del número de copias. La investigación sobre la relación entre la variación del número de copias y la enfermedad tiene una larga historia. Para algunas variaciones en el número de copias de la mutación de la línea germinal (es decir, las variaciones en el número de copias generadas debido a las variaciones de un feto en sí, que están ausentes en ambos padres), se cree que cuanto más grande es un fragmento, más fácil se produce una anomalía congénita, y para por ejemplo, las enfermedades de aneuploidía cromosómica (por ejemplo, T21, T18, etcétera) y los síndromes de microdeleción/microduplicación cromosómica son enfermedades reconocidas que se relacionan con la variación del número de copias por mutación de la línea germinal.

Los síndromes de microdeleción/microduplicación cromosómica humana son un tipo de enfermedad de fenotipos complejos y cambiantes causados por la ocurrencia de deleciones o duplicaciones de microfragmentos, es decir, variaciones en el número de copias en fragmentos de ADN, en cromosomas humanos con una incidencia relativamente alta en lactantes perinatales y lactantes neonatales, y puede conducir a enfermedades y anomalías graves, por ejemplo, cardiopatía congénita o malformación cardíaca, retraso grave del crecimiento, apariencia o deformidad de las extremidades, etcétera. Además, los síndromes de microdeleción también son una de las principales razones que causan retraso mental además deSíndrome de Down y síndrome de X frágil (Knight SJL (edición): Genetics of Mental Retardation. Monogr Hum Genet. Basel, Karger, 2010, vol 18, página 101 - 113 (DOI: 10.1159/000287600)). En los últimos años, la enfermedad cardíaca congénita que encabeza la lista en las estadísticas sobre la incidencia de defectos congénitos y el retraso mental, la parálisis cerebral y la sordera congénita, que se clasifican en las clínicas ambulatorias de asesoramiento y diagnóstico genético, están relacionadas con los síndromes de microdeleción. Los síndromes de microdeleción comunes incluyen el síndrome de microdeleción 22q11, el síndrome de cri du chat, el síndrome de Angelman, la deleción de AZF, etcétera.

Aunque la incidencia de cada síndrome de microdeleción es muy baja, en donde las incidencias del síndrome de microdeleción 22q11 relativamente común, el síndrome de cri du chat, el síndrome de Angelman, el síndrome de Miller-Dieker, etcétera. son 1: 4,000 (nacimientos vivos), 1: 50,000, 1: 10,000 y 1: 12,000 respectivamente. Debido a la limitación de las técnicas de detección clínica, un gran número de pacientes con síndromes de microdeleción no pueden detectarse en el pesquisaje y diagnóstico prenatal. E incluso cuando se busca una razón en retrospectiva después de la ocurrencia de caracterizaciones clínicas típicas, meses o incluso años después del nacimiento de un bebé, la causa de la enfermedad tampoco se puede diagnosticar debido a la limitación de las técnicas de detección (https://decipher.sanger.ac.uk/syndromes). Debido a que no puede lograrse una cura radical para algunos tipos de síndromes de microdeleción ocurriendo la muerte durante los meses o años posteriores al nacimiento, se genera una gran carga mental y económica para la sociedad y las familias. De acuerdo con estadísticas incompletas, los pacientes con "síndrome de la marioneta feliz" (síndrome de Angelman) en todo el mundo han alcanzado los 15 mil, y el número de pacientes con los otros tipos de síndromes de microdeleción cromosómica también está aumentando año tras año. Por lo tanto, la detección previa al embarazo de microdeleciones/microduplicaciones cromosómicas realizadas en pacientes clínicamente sospechosos y padres con antecedentes de embarazo anormal relacionado es propicio para proporcionar asesoramiento genético y proporcionar una base para la toma de decisiones clínicas. El diagnóstico prenatal temprano durante el embarazo puede prevenir eficazmente el nacimiento de una paciente infantil o proporcionar una base para proporcionar un enfoque de tratamiento para una paciente infantil después del nacimiento (Bretelle F, y otros Prenatal and postnatal diagnosis of 22q11.2 deletion syndrome. Eur J Med Genet. 2010 Nov-Dic;53(6):367-70).

Sin embargo, este tipo de enfermedades no se pueden detectar mediante métodos clínicos de rutina, como el método de cariotipo cromosómico y similares (con una resolución superior a 10 M) debido al nivel micro de variación en el cromosoma (Malcolm S. Microdeletion and microduplication syndromes. Prenat Diagn. 1996 Dic;16(13): 1213-9). Actualmente, para el diagnóstico prenatal de los síndromes de microdeleción/microduplicación, los métodos de líquido amniótico fetal invasivo u otros tejidos se adoptan principalmente para realizar el diagnóstico molecular. Actualmente, los métodos de diagnóstico molecular invasivo incluyen principalmente cariotipo cromosómico de alta resolución, FISH (hibridación fluorescente in situ), Matriz CGH (hibridación genómica comparativa), MLPA (técnica de amplificación de sonda dependiente de la ligadura multiplex), PCR y similares. Entre ellos, el examen FISH se utiliza como un estándar de oro para el diagnóstico genético que puede detectar eficazmente la mayoría de las deleciones de fragmentos cromosómicos. Sin embargo, debido a que el muestreo invasivo necesita una cierta cirugía o cultivo celular, desde el punto de vista de la eficiencia temporal y el consumo de recursos, el mismo es adecuado para actuar como un indicador de diagnóstico, pero no como un método para el pesquisaje clínico universal.

En términos de métodos de detección no invasivos para los síndromes de microdeleción/microduplicación, también hay algunos intentos. Por ejemplo, en un estudio sobre la detección no invasiva de los síndromes de microdeleción en un feto publicado en noviembre de 2011, los investigadores realizaron una secuenciación de alta profundidad en el plasma de la madre durante el embarazo, generaron aproximadamente 243 millones de lecturas cortas y detectaron una microdeleción de alrededor de 4 Mb de 12p 11.22 a 12p 12.1 en el feto (David Peters y otros Noninvasive Prenatal Diagnosis of a Fetal Microdeletion Syndrome .N Engl J Med 2011; 365:1847-1848). Sin embargo, la generación de una cantidad tan grande de datos no es adecuada para uso clínico en términos de consumo de recursos o eficiencia temporal.

De la combinación de los contenidos mencionados anteriormente se puede saber que actualmente, entre los métodos de examen prenatal para los síndromes de microdeleción/microduplicación cromosómica, todavía no existe un método de detección universal factible. En este campo, se necesita un nuevo método confiable de detección para la variación del número de copias en un feto, con el fin de identificar sitios conocidos y explorar sitios desconocidos de manera novedosa.

Resumen de la invención

Con el desarrollo continuo de la técnica de secuenciación de alto rendimiento y la reducción continua en el costo de secuenciación, los estudios sobre técnicas de secuenciación en el pesquisaje prenatal hacen un análisis del pesquisaje de la variación del número de copias cromosómicas y la aneuploidía y otras variaciones genéticas, y en particular, la variación por aneuploidía fetal cromosómica, por secuenciación de alto rendimiento, que se aplica cada vez más. Para detectar la variación genética, la presente invención diseña un método de detección de la variación genética basado en la técnica de secuenciación de alto rendimiento que puede usar la detección de la variación del número de copias y la aneuploidía y otras variaciones genéticas y tiene las características de alto rendimiento, alta especificidad y localización precisa.

En consecuencia, la presente invención proporciona un método implementado por ordenador para detectar la variación genética como se establece en cualquiera de las reivindicaciones 1 a 15.

En la presente descripción se describe la adquisición de una muestra de prueba y la extracción de ADN, la secuenciación de alto rendimiento y el análisis de los datos obtenidos para obtener un resultado de detección.

En la presente descripción también se describe un método para detectar la variación genética, que comprende las siguientes etapas:

1) adquirir lecturas de una muestra de prueba, en donde el fragmento de las lecturas, por ejemplo, puede tener una longitud de 25-100 nt, y el número de fragmentos de las lecturas puede ser de al menos 1 millón;
2) alinear las lecturas con una secuencia del genoma de referencia;
3) dividir la secuencia del genoma de referencia en ventanas, calcular el número de lecturas que alinean con cada ventana, y adquirir la estadística para cada ventana en función del número de las lecturas;
4) y para un fragmento de la secuencia del genoma de referencia, sobre la base del cambio en las estadísticas de todas las ventanas al respecto en el fragmento de la secuencia del genoma de referencia, adquirir posiciones donde se produce un cambio significativo en las estadísticas de las ventanas en ambos lados, estas posiciones son posiciones donde los sitios de variación genética de la muestra de prueba están en la secuencia del genoma de referencia.

Dicho sitio de variación genética de acuerdo con el método de la invención es el punto medio entre un punto de inflexión donde dicha estadística cambia de ascendente a descendente y el siguiente mismo punto de inflexión, y hay al menos 50, al menos 70, al menos 100, preferentemente 100 longitudes de ventana entre dos sitios de variación genética; y el sitio mencionado anteriormente, el punto de inflexión y el punto medio se refieren a una posición cromosómica correspondiente a una ventana correspondiente a la estadística, y pueden representarse por el punto inicial, el punto medio, el punto final y cualquier otra posición de la ventana.

Como se describe en la presente descripción, existe la etapa 5) de pesquisaje de los sitios de variación genética para obtener sitios de variación genética posteriores al pesquisaje.

Por ejemplo, la etapa 5) mencionada anteriormente es para cada sitio de variación genética y comprende realizar estadísticas de la diferencia entre dos grupos numéricos que consisten en estadísticas de ventanas contenidas en el fragmento entre el sitio de variación genética y su sitio de variación genética precedente y en el fragmento entre el sitio de variación genética y su sitio de variación posterior, y eliminar el sitio de variación genética cuyo valor de diferencia de significación es máximo y mayor que un umbral preestablecido; y repetir el proceso mencionado anteriormente, hasta que los valores de significación de la diferencia de los sitios de variación genética sean todos más pequeños que el umbral preestablecido, en donde dicha significación de la diferencia, por ejemplo, puede realizarse mediante la prueba de ejecución, eliminando el sitio de variación genética cuyo valor de significación en la prueba de ejecución es máximo y mayor que el umbral preestablecido; y repetir el proceso mencionado anteriormente, hasta que los valores de significación de los sitios de variación genética en la prueba de ejecución sean todos más pequeños que el umbral preestablecido.

También como se describe, el umbral preestablecido utilizado en la etapa 5) mencionado anteriormente se obtiene mediante las siguientes etapas:

a) adquirir los sitios de variación genética de acuerdo con el método de la presente invención sustituyendo la muestra de prueba con una muestra de control;
b) para cada sitio de variación genética, realizar estadísticas sobre la diferencia entre dos grupos numéricos que consisten en estadísticas de ventanas contenidas en el fragmento entre el sitio de variación genética y su sitio de variación genética anterior y en el fragmento entre el sitio de variación genética y su sitio de variación posterior, y eliminar el sitio de variación genética que es el menos significativo; y
c) repetir la etapa b) mencionada anteriormente, hasta que el número de puntos de inflexión candidatos restantes sea igual al valor esperado N_c en donde $N_c = L_c/T$, L_c es la longitud de la secuencia del genoma, la máxima precisión teórica T es el tamaño del fragmento que puede detectarse teóricamente, la precisión teórica final $T = W + S*N$ cuando el promedio de los tamaños de ventana es W, la longitud deslizante de las ventanas es S y el número de cada grupo de ventanas en la prueba de ejecución es N, y entre los valores de significación de todos los puntos de ruptura candidatos restantes, el mínimo es el umbral de significación.

Como se describe en la presente descripción, también hay un método para detectar la variación genética, que comprende las siguientes etapas:

1) adquirir los sitios de variación genética en un fragmento de la secuencia del genoma de referencia de acuerdo con el método de la presente invención;
2) y una etapa para realizar una selección basada en la confianza estadística en fragmentos entre dichos sitios de variación genética.

La etapa de la selección basada en la confianza en la etapa 2) mencionada anteriormente puede comprender:

i) calcular la probabilidad de distribución de las estadísticas a través del patrón de distribución de las estadísticas para las ventanas, y establecer un umbral;
ii) y comparar el promedio de las estadísticas de ventanas en el fragmento entre los sitios de variación genética posteriores al pesquisaje con dicho umbral, y determinar si el fragmento entre los sitios genéticos es anómalo sobre la base del resultado de la comparación.

De otra manera, la etapa de selección basada en la confianza en la etapa 2) mencionada anteriormente comprende:

i) calcular la probabilidad de distribución de las estadísticas a través del patrón de distribución de las estadísticas para las ventanas, y establecer un primer umbral y un segundo umbral;
ii) y comparar el promedio de las estadísticas de ventanas en el fragmento entre los sitios de variación genética posteriores al pesquisaje con dichos primer umbral y segundo umbral,
en donde, si las estadísticas para ventanas en el fragmento son más pequeñas que el primer umbral, el fragmento es una deleción de fragmentos, y si las mismas son mayores que el segundo umbral, el fragmento es una duplicación de fragmentos,
en donde dicho primer umbral es un valor del estadístico donde la probabilidad acumulativa de la ocurrencia del estadístico es menor o igual a 0.1, preferentemente menor o igual a 0.01, con la máxima preferencia 0.05, y/o dicho segundo umbral puede ser un valor del estadístico donde la probabilidad acumulativa de la ocurrencia del estadístico es mayor o igual a 0.9, preferentemente mayor o igual a 0.99, con la máxima preferencia 0.95.

En la presente descripción se describe más a fondo un medio legible por ordenador, que lleva una serie de códigos ejecutables, que pueden ejecutar el método de detección genética de la invención y como se describe en la presente.

También se describe un método para detectar la variación genética fetal, que comprende las siguientes etapas:

adquirir una muestra materna que contiene ácido nucleico fetal;
secuenciar dicha muestra materna;
y una etapa de detección de la variación genética usando el método como se describe en la presente descripción. Dicha muestra materna puede ser sangre periférica materna.

La superioridad de la presente invención, en comparación con los métodos actuales para detectar la variación genética, incluye principalmente los siguientes puntos:

(1) Es clínicamente factible: se usa solo alrededor de 5 M de datos de secuenciación, y pueden detectarse alrededor de 5 Mb de fragmentos de CNV, mientras que un método reportado usó casi 243 M, por lo tanto, nuestro método reduce el costo y el consumo de tiempo de generación de datos en gran medida.
(2) Es extensible: se puede aumentar la precisión al expandir el número de grupos de control además de aumentar la cantidad de secuenciación, para reducir la presión sobre la cantidad inicial de ADN.

(3) Es más estable y más completo: no hay detalles de la operación en sí mismos que se señalen claramente en los artículos informados, mientras que la presente invención diseña la corrección de los grupos de datos, la preferencia por las condiciones de fragmentación y varios otros aspectos.

Descripción de las Figuras

La Figura 1 es un breve diagrama de flujo sobre el análisis de variación genética de cromosomas en un ejemplo de la presente invención.
La Figura 2A es un cariograma cromosómico digital de S67.
La Figura 2B es un cariograma cromosómico digital de S10.
La Figura 2C es un cariograma cromosómico digital de S14.
La Figura 2D es un cariograma cromosómico digital de S18.
La Figura 2E es un cariograma cromosómico digital de S49.
La Figura 2F es un cariograma cromosómico digital de S55.
La Figura 2G es un cariograma cromosómico digital de S82.
La Figura 2H es un cariograma cromosómico digital de S103.

Breve descripción de las tablas

La Tabla 1 es una lista de resultados de la CNV de todas las muestras en el caso de implementación.
La Tabla 2 muestra los resultados de detección de aCGH y cariotipificación de todas las muestras en el caso de implementación.
La Tabla 3. muestra los resultados de la prueba en el presente caso de implementación y los resultados de la detección de cariotipo estándar.

Modalidades

El tipo de ácido nucleico no está particularmente limitado, el cual puede ser ácido desoxirribonucleico (ADN), y también puede ser ácido ribonucleico (ARN), preferentemente ADN. Los expertos en la técnica entenderán que, para el ARN, el mismo puede convertirse por medios convencionales en ADN con una secuencia correspondiente para realizar la detección y el análisis posteriores. Además, la propiedad de la muestra de prueba tampoco se limita particularmente. Puede adoptarse una muestra de ADN genómico, y una parte del ADN genómico también puede utilizarse como muestra de prueba. La fuente de la muestra de prueba no se limita particularmente. Una muestra de una mujer embarazada puede adoptarse como muestra de prueba, de la que se puede extraer una muestra de ácido nucleico que contiene información genética fetal, y luego puede detectarse y analizarse la información genética fetal y el estado fisiológico. Los ejemplos de una muestra de una mujer embarazada que pueden usarse incluyen, pero no se limitan a, sangre periférica de la mujer embarazada, orina de la mujer embarazada, células trofoblásticas exfoliadas del cuello uterino de la mujer embarazada, moco cervical de la mujer embarazada, y glóbulos rojos nucleados fetales. Los inventores descubrieron que a través de la extracción de la muestra de ácido nucleico de la muestra de una mujer embarazada mencionada anteriormente, la variación genética en el genoma de un feto puede analizarse de manera efectiva para realizar el diagnóstico prenatal no invasivo o la detección en el feto. Ventajosamente, se puede realizar una detección no invasiva de la variación genética fetal, por ejemplo, dicha muestra es sangre periférica de una mujer embarazada, además, es posible la detección invasiva, por ejemplo, dicha muestra puede ser de sangre del cordón umbilical; dicho tejido puede ser tejido placentario o tejido coriónico; y dichas células pueden ser células de líquido amniótico no cultivadas o cultivadas y células progenitoras de vellosidades. Un sujeto a ser ensayado y un sujeto normal son de la misma especie. Además, la detección de variación de la presente invención no se usa necesariamente para el diagnóstico de enfermedades o propósitos relacionados, porque con la presencia de polimorfismo, la presencia de algunas variaciones en relación con un genoma de referencia no representa el riesgo de padecer una enfermedad o estado de salud. Por lo tanto, la detección de variación de la presente invención puede ser simplemente para uso en investigación científica sobre polimorfismo genético.

En la presente invención, una muestra de control se opone a la muestra de prueba. Por ejemplo, en un método relacionado con la detección de una enfermedad, la muestra de control se refiere a una muestra normal. Por ejemplo, en una modalidad de la presente invención, la muestra de prueba es sangre periférica materna, y la muestra de control correspondiente es sangre periférica de una madre normal que concibe un feto normal.

Como se describe en la presente descripción, el método y el aparato para extraer la muestra de ácido nucleico de la muestra de prueba tampoco se limita particularmente, y pueden usarse estuches de extracción de ácido nucleico comercializados para realizarlo.

En el método de la presente invención, dichas ventanas tienen el mismo número de lecturas únicas de referencia. Las lecturas únicas de referencia se refieren a un fragmento cromosómico con una secuencia única, este fragmento puede ubicarse definitivamente en una sola posición cromosómica, y las lecturas únicas de referencia cromosómica pueden construirse en base a una secuencia descrita del genoma cromosómico de referencia, por ejemplo, hg18 o hg19. Un proceso para adquirir las lecturas únicas de referencia generalmente incluye las etapas de cortar el genoma de referencia en lecturas de cualquier longitud fija, alinear estas lecturas nuevamente con el genoma de referencia y seleccionar las lecturas que están alineadas únicamente con el genoma de referencia como lecturas únicas de referencia. Dicha longitud

fija depende de la longitud de las secuencias en el resultado de secuenciación por un secuenciador, refiriéndose a la longitud promedio para detalles. Las longitudes en los resultados de secuenciación obtenidos por diferentes secuenciadores son diferentes, y específicamente para cada ejecución de secuenciación, las longitudes en los resultados de secuenciación también pueden ser diferentes, y existen ciertos factores subjetivos y de experiencia existentes en la selección de la longitud.

En un ejemplo de la presente invención, la longitud de las lecturas únicas de referencia se selecciona de acuerdo con las longitudes reales de las secuencias en el resultado de la secuencia, por ejemplo, 25-100 pb, y para el sistema illumina/Solexa, por ejemplo, opcionalmente 50 pb, y luego el número de lecturas únicas de referencia contenidas en cada ventana se controla en 800,000-900,000. En el método de la presente invención, dichas ventanas pueden tener una superposición o no tener superposición entre ellas. En un ejemplo de la presente invención, la distancia entre ventanas adyacentes es 1-100 kb, preferentemente 5-20 kb, con mayor preferencia 10 kb. Esta distancia puede ajustarse de acuerdo con la abundancia de ADN en la muestra fetal. El principio del ajuste es que cada ventana corresponde a una estadística y una posición cromosómica, lo que también significa que la distancia entre ventanas determina la precisión de la detección. Cuanto mayor sea la precisión, mayor será el fondo derivado de la madre y más difícil será la discriminación de las fuentes de variaciones genéticas.

En el método de la presente invención, dicho estadístico puede ser el número de lecturas en sí mismo, pero preferentemente es un estadístico después de la corrección de errores (por ejemplo, corrección de GC) y/o estandarización de datos, cuyo propósito es que el estadístico cumpla con una distribución común en las estadísticas, por ejemplo, la distribución normal o estándar normal. El análisis estadístico posterior de las estadísticas puede, por lo tanto, facilitarse. En un ejemplo de la presente invención, se realiza el proceso de estandarización frente al número promedio de lecturas de todas las ventanas. En un ejemplo de la presente invención, la estandarización incluye un proceso para evaluar el valor Z de aquí en adelante. En una modalidad, dicha estadística se ajusta aproximadamente a la distribución normal obtenida por el proceso de estandarización en el número de lecturas que están alineadas a una ventana. En una modalidad, dicha estandarización se basa en el número promedio de lecturas que están alineadas con todas las ventanas. En una modalidad, dicha estadística es una estadística de distribución normal estándar aproximada.

En la presente invención, las lecturas se refieren a fragmentos de secuencia emitidos por un secuenciador, preferentemente alrededor de 25-100 nt.

Las moléculas de ADN se pueden adquirir mediante el uso del método de cristalización por precipitación, el método de cromatografía en columna, el método de esferas magnéticas, el método SDS y otros métodos de extracción de ADN de rutina, preferentemente usando el método de esferas magnéticas. El llamado método de esferas magnéticas se refiere a las moléculas de ADN aisladas obtenidas a partir de que la sangre, los tejidos o las células se someten a la acción de una solución de lisis celular y proteinasa K, utilizando esferas magnéticas específicas para realizar una adsorción de afinidad reversible en las moléculas de ADN y después de las proteínas, los lípidos y otras impurezas se eliminan lavando con un líquido de enjuague, eluyendo las moléculas de ADN de las esferas magnéticas con un líquido de purificación. Las perlas magnéticas se conocen bien en la técnica y están disponibles comercialmente, por ejemplo, de Tiangen.

Generalmente, se lleva a cabo la secuenciación directa de las moléculas de ADN obtenidas de las muestras y las etapas posteriores, y el ADN extraído puede usarse para las etapas posteriores sin ser procesado. En algunos ejemplos preferidos, solo pueden estudiarse fragmentos con bandas principales electroforéticas concentradas en el tamaño de 50-700 pb, preferentemente 100-500 pb, más preferentemente 150-300 pb, particularmente aproximadamente 200 pb. En algunos ejemplos más preferidos, las moléculas de ADN pueden dividirse en fragmentos con bandas principales electroforéticas concentradas en un cierto tamaño, por ejemplo, 50-700 pb, preferentemente 100-500 pb, con mayor preferencia 150-300 pb, particularmente cerca de 200 pb, y entonces se realizan las siguientes etapas. El tratamiento de romper al azar dichas moléculas de ADN puede usar digestión enzimática, atomización, ultrasonido o el método HydroShear. Preferentemente, se usa el método de ultrasonido, por ejemplo, la serie S de Covaris Corporation (basada en la técnica AFA, en la que cuando la energía del sonido/energía mecánica liberada por un sensor pasa a través de una muestra de ADN, el gas se disuelve para formar burbujas; después de eliminar la energía, las burbujas explotan y se genera la capacidad de fracturar las moléculas de ADN; mediante el establecimiento de una determinada intensidad de energía e intervalo de tiempo y otras condiciones, las moléculas de ADN pueden fraccionarse en tamaños dentro de un cierto rango; por ejemplo, para un principio y un método específicos, consulte las instrucciones para la serie S de Covaris Corporation).

Como se describe en la presente invención, dicho punto de ruptura o punto de ruptura candidato es un sitio de variación genética potencial o existente, y por convención, el sitio se expresa como la posición en el genoma de referencia. Los dos conceptos, sitio de variación genética y punto de ruptura, son intercambiables en un caso particular, y simplemente diferentes en expresión, y ambos pueden usarse para representar la coordenada de posición de una variación genética potencial o definitivamente existente en el genoma de referencia en varias etapas.

El método de secuenciación puede adoptarse para adquirir las lecturas de la muestra de prueba, y dicha secuenciación puede realizarse a través de cualquier método de secuenciación, que incluye, entre otros, el método de terminación de cadena didesoxi; preferentemente un método de secuenciación de alto rendimiento, que incluye, pero no se limita a,

técnicas de secuenciación de segunda generación o técnicas de secuenciación de moléculas individuales (Rusk, Nicole (2009-04-01). Secuenciación económica de tercera generación. Nature Methods 6 (4): 2446 (4).

Plataformas para dicha secuenciación de segunda generación (Metzker ML. Sequencing technologies-the next generation. Nat Rev Genet. 2010 enero;11(1):31-46) incluyen, pero no se limitan a, la plataforma de secuenciación Illumina-Solexa (GATM, HiSeq2000TM, etc.), ABI-Solid y Roche-454 (pirosecuenciación); y las plataformas (técnicas) para la secuenciación de moléculas individuales incluyen, pero no se limitan a, secuenciación de ADN de molécula única verdadera de Helicos Corporation, secuenciación de molécula única en tiempo real (SMRTTM) de Pacific Biosciences Corporation y técnica de secuenciación de nanoporos de Oxford Nanopore Technologies Corporation, etcétera.

El tipo de secuenciación puede ser secuenciación de un solo extremo y secuencia de pares, y la longitud de la secuencia puede ser de 50 pb, 90 pb o 100 pb. En una modalidad de la presente invención, dicha plataforma de secuenciación es Illumina/Solexa, el tipo de secuenciación es secuenciación de extremos pareados, y se obtiene una secuencia de molécula de ADN de tamaño de 100 pb con la relación posicional de extremo par.

Como se describe, la profundidad de secuenciación de la secuenciación puede determinarse de acuerdo con el tamaño del fragmento de variación cromosómica fetal que se detectará, y cuanto mayor sea la profundidad de secuenciación, mayor será la sensibilidad de detección, es decir, menor será el fragmento de deleción y duplicación detectable. La profundidad de secuenciación puede ser 1 - 30 ×, es decir, la cantidad total de datos es 1-30 veces la longitud del genoma humano, por ejemplo, en una modalidad de la presente invención, la profundidad de secuenciación es 0.1 ×, es decir, 2 veces (2.5 × 108 pb).

Cuando las moléculas de ADN a analizar provienen de una pluralidad de muestras de prueba, se puede agregar una secuencia etiqueta diferente a cada muestra, para discriminar las muestras en el proceso de secuenciación (Micah Hamady, Jeffrey J Walker, J Kirk Harris y otros Error-correcting barcoded primers forpyrosequencing hundreds of samples in multiplex. Nature Methods, 2008, marzo, Vol.5 No.3), de este modo se realiza la secuenciación simultánea de la pluralidad de muestras. Las secuencias etiquetas son para discriminar diferentes secuencias, pero no afectarán otras funciones de las moléculas de ADN a las que se agregaron las secuencias etiquetas. La longitud de una secuencia etiqueta puede ser de 4-12 pb.

Dicha secuencia de referencia genómica humana es preferentemente una secuencia de referencia genómica humana en la base de datos NCBI, por ejemplo, la secuencia de referencia genómica humana compilación 36 en la base de datos NCBI (hg18; NCBI versión 36).

Como se describe en la presente descripción, dicha alineación puede ser una alineación sin desajuste permitido, y también puede ser una alineación con 1 desajuste de base permitido. La alineación de secuencia puede realizarse a través de cualquier programa de alineación de secuencia, por ejemplo, el Paquete de análisis de oligonucleótidos cortos (SOAP) y la alineación BWA (Burrows-Wheeler Aligner) que están disponibles para los expertos en la técnica, y las lecturas están alineadas con la secuencia del genoma de referencia para obtener las posiciones de las lecturas en el genoma de referencia. La alineación de la secuencia puede realizase usando los parámetros predeterminados proporcionados por el programa, o los expertos en la técnica seleccionan los parámetros de acuerdo con la necesidad. En una modalidad de la presente invención, el software de alineación usado es SOAPaligner/soap2.

El algoritmo de dicho software es una serie de programas para la detección de la variación del número de copias en un feto desarrollado por el instituto BGI en Shenzhen, que se denominan colectivamente FCAPS. Puede realizar corrección de datos, estandarización y fragmentación en una muestra de prueba y un conjunto de control a través de los datos generados por la técnica de secuenciación de nueva generación, y estimar el alcance y el tamaño de las variaciones en el número de copias en un feto.

En algunos ejemplos particulares, para la etapa 1), la adquisición de lecturas de una muestra de prueba es después de la extracción de ADN de plasma de la muestra de prueba y la muestra de control de acuerdo con el manual de operación para el Estuche Tiangen DP327-02, se construye una biblioteca de acuerdo con el procedimiento modificado de construcción de la biblioteca estándar de Illumina/Solexa. Para obtener detalles sobre la construcción de una biblioteca de secuenciación del genoma completo, consulte las reglas de la directiva provistas por el fabricante del secuenciador, por ejemplo, Illumina Corporation, por ejemplo, consulte la Guía de preparación de muestras de multiplexación (Parte#1005361; febrero 2010) o la guía Paired-End SamplePrep (Parte#1005063; febrero 2010) por Illumina Corporation. En este proceso, los adaptadores usados para la secuenciación se agregan a ambos extremos de las moléculas de ADN que se concentran a 200 pb, se agrega una secuencia etiqueta diferente a cada muestra, por lo tanto, los datos para una pluralidad de muestras pueden discriminarse en los datos obtenidos por un secuencia única, y con el uso del método de secuenciación de segunda generación, secuenciación Illumina/Solexa (pueden usarse otros métodos de secuenciación como ABI/SOLiD para lograr el mismo efecto o similar), se obtienen lecturas con cierto tamaño de fragmento para cada muestra.

En algunos ejemplos particulares, para la etapa 2) alineación: las lecturas en la etapa 1) en el método de la presente invención están alineadas con SOAP2 con la secuencia genómica estándar de referencia humana en la base de datos NCBI para obtener la información posicional en el genoma de la secuencia de ADN secuenciada. Para evitar la

perturbación del análisis de CNV mediante secuencias repetidas, solo las lecturas que están alineadas con la secuencia de referencia genómica humana de forma exclusiva se seleccionan para el análisis posterior.

En algunos ejemplos particulares, la etapa 3), dividir en ventanas y adquirir estadísticas para las ventanas, comprende las siguientes etapas:

a) para la muestra de prueba y la muestra de control, proporcionar ventanas con la longitud de w en la secuencia genómica de referencia, calculando el contenido de GC en cada ventana y calculando el número relativo de fragmentos de lecturas que caen en cada ventana; y b) corregir y estandarizar el número relativo de lecturas de fragmentos de la muestra de prueba mencionada anteriormente frente al número relativo de lecturas de fragmentos de la muestra control.

En algunos ejemplos particulares, la corrección de GC basada en el conjunto de muestra de control se realiza en la muestra de prueba: porque existe un cierto sesgo de GC entre/dentro de los lotes de secuenciación, lo que hace que se produzca una desviación del número de copias en la región de alta GC o en la región de baja GC en el genoma, el número relativo corregido de lecturas en cada ventana obtenida por la corrección GC de los datos de secuenciación basados en el conjunto de muestras de control puede eliminar este sesgo y mejorar la precisión de detección de la variación del número de copias. El número relativo corregido de lecturas en cada ventana está estandarizado: la variación del número de copias en un feto se detecta usando plasma de la madre embarazada, y con el efecto del fondo de ADN de la madre, la variación en el feto es relativamente difícil de destacar, por lo que se exige reducir el ruido del fondo de ADN de la madre y amplificar la señal de la variación del número de copias en el feto a través de la estandarización. En un ejemplo, dicha corrección de GC comprende las siguientes etapas: a) adquirir lecturas que están alineadas a cada ventana de acuerdo con el método de la presente invención sustituyendo la muestra de prueba con una muestra control y calculando el número relativo de las lecturas para cada ventana; b) adquirir la relación funcional entre el contenido de GC de las lecturas que están alineadas con cada ventana y el número relativo de las lecturas para dicha ventana; y c) para cada ventana, usando el contenido de GC de las lecturas de la muestra de prueba alineadas con la ventana y la relación funcional mencionada anteriormente, y corrigiendo el número relativo de lecturas de la muestra de prueba para la ventana para obtener el número relativo corregido de lecturas para la ventana.

En algunos ejemplos particulares, la etapa 3), dividir en ventanas y adquirir estadísticas para las ventanas, comprende las siguientes etapas:

a) calcular el número relativo de lecturas de la muestra de prueba y la de la muestra control: para la muestra de prueba y la muestra control, proporcionar ventanas con la longitud de w en la secuencia de referencia genómica humana, calcular el número de lecturas, $r_{i,j}$, caer en cada ventana en la etapa 2) en el método de la presente invención, donde los subíndices i y j representan el número de serie de la ventana y el número de serie de la muestra respectivamente, y calcular el contenido de GC, $GC_{i,j}$, de cada ventana y calcular el número relativo de lecturas, $R_{i,j} = \log_2\left(\frac{r_{i,j}}{\overline{r}_j}\right)$, donde el número promedio de lecturas es $\overline{r}_j = \frac{1}{n}\sum_{i=1}^{n} r_{i,j}$,

b) corrección y estandarización de datos:
①en un sistema de coordenadas con el contenido de GC como abscisa y el número relativo de lecturas R como ordenada, realizar un ajuste lineal en $R_{i,j}$ y $GC_{i,j}$ de la muestra control para obtener la pendiente a_i y la intercepción b_i,
②para cada ventana de la muestra de prueba, calcular el número relativo corregido de lecturas

$$\hat{R}^{\%}_{i,j} = a_i \times GC_{i,j} + b_i$$

③Para cada ventana de la muestra de prueba, calcular la estadística $Z_{i,j}$:

$$Z_{i,j} = \left(R_{i,j} - \hat{R}^{\%}_{i,j} - media_{\ j}\right) / SD_j, \quad donde \quad media_{\ j} = \frac{1}{n}\sum_{i=1}^{n}\left(R_{i,j} - \hat{R}^{\%}_{i,j}\right),$$

$$SD_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(R_{i,j} - \hat{R}^{\%}_{i,j} - media_{\ j}\right)^2}$$

En algunos ejemplos particulares, la adquisición de posiciones donde los sitios de variación genética de la muestra de prueba están en la secuencia del genoma de referencia en la etapa 4) se realiza a través de las siguientes etapas:

①inicio: para un punto final de cada ventana, si la tendencia de cambio de la estadística Z se cambia entre la ventana anterior y la siguiente ventana del punto, y la distancia entre el punto y un punto anterior donde la tendencia de cambio de la estadística Z se cambia entre la ventana anterior y la siguiente ventana es al menos n ventanas (n es un número entero de 10-500, preferentemente 50-300, por ejemplo 100), entonces el punto es un punto de inflexión candidato, por ejemplo, el punto medio entre ese punto de inflexión donde la estadística Z pasa de ascendente a descendente entre la

ventana anterior y la siguiente ventana y el siguiente mismo punto de inflexión es un punto de inflexión candidato, o el punto medio entre ese punto de inflexión donde la estadística Z pasa de descendente a ascendente entre la ventana anterior y la siguiente ventana y el siguiente punto de inflexión es un punto de inflexión candidato bk (k = 1, 2, ..., s; s es un número entero de > 0);

②iteración óptima: para estudiar la variación del número de copias o la aneuploidía en un fragmento de la secuencia del genoma, todos los puntos de corte candidatos clasificados del fragmento de la secuencia del genoma se registran como $B_c = \{b_1, b_2, ..., b_s\}$, en el que existen dos fragmentos, uno a la izquierda y otro a la derecha, para cada punto de inflexión candidato b_k y dichos fragmentos son una región desde el punto de inflexión anterior al punto de inflexión y una región desde el punto de inflexión hasta el siguiente punto de inflexión; el valor p (p_k) obtenida sometiendo $Z_{i,j}$ para todas las ventanas en estos dos fragmentos a una prueba (por ejemplo, se considera la prueba de ejecución que es una prueba no paramétrica y que usa el estado uniforme de distribución después de mezclar elementos de dos grupos para evaluar la importancia de la diferencia entre estos dos grupos) como "el significado de b_k como punto de inflexión"; un punto de inflexión candidato con el máximo p_k excluido; y esta etapa se repite hasta que todos los valores p sean más pequeños que el valor p de terminación (p_{final}) para la secuencia del genoma;

③adquisición del valor de terminación p: en el proceso de prueba, las etapas a) a c) mencionadas anteriormente ① se llevan a cabo usando otra muestra control como muestra de prueba, para un fragmento de una secuencia del genoma, todos los puntos de inflexión candidatos clasificados del fragmento de la secuencia del genoma se registran como $B_c = \{b_1, b_2, ..., b_s\}$, en el que existen dos ventanas, una a la izquierda y otra a la derecha, para cada punto de inflexión candidato b_k; el valor p (p_k) obtenido al someter a todos los $Z_{i,j}$ en estas dos ventanas a la prueba de ejecución se considera "la importancia de b_k como punto de inflexión"; un punto de inflexión candidato con el máximo p_k excluido; las dos ventanas a la izquierda y a la derecha se fusionan hasta que el número de puntos de inflexión candidatos sea igual al valor esperado N_c ($N_c = L_c/T$, donde L_c es la longitud de la secuencia del genoma c, T (precisión final teórica) es el tamaño del fragmento que se puede detectar teóricamente; y cuando el tamaño de la ventana es W, la longitud de deslizamiento de las ventanas es S, y el número de cada grupo en la prueba de ejecución es N, la precisión teórica máxima T = W+S*N); y en el conjunto de los puntos de inflexión candidatos, el valor p mínimo es el valor p de terminación (p_{final}) de la secuencia genómica.

En algunos ejemplos particulares, la etapa de realizar una selección basada en la confianza en fragmentos entre dichos sitios de variación genética es: para un fragmento entre sitios de variación genética en la secuencia del genoma de referencia, el promedio de $Z_{i,j}$ en el fragmento se calcula y registra como $\overline{Z}$ en donde si $\overline{Z}$ del fragmento es menor que -1.28, entonces el fragmento es una eliminación de fragmento, y si el mismo es mayor que 1.28, entonces el fragmento es una duplicación de fragmento.

En la presente invención, la prueba de ejecución es una prueba no paramétrica en la que, de acuerdo con el estado uniforme de distribución de elementos en dos grupos después de mezclar los dos grupos, adquiere el valor de significancia P para evaluar estos dos grupos. Ver http://support.sas.com/kb/33/092.html.

En la presente invención, durante la prueba con la muestra control como muestra de prueba, la secuenciación o los experimentos en la práctica conducirán a la existencia de diferencias en el número de fragmentos para secuenciar con respecto a los diferentes fragmentos que se alinearon en todo el genoma, por lo que en el proceso del ensayo, estas diferencias serán discriminadas, y solo fragmentos en ambos extremos de un punto de inflexión aún no han alcanzado el nivel de variación. Al comienzo del ensayo, estas diferencias no pueden ser discriminadas de manera relativamente significativa por los puntos de inflexión candidatos, por lo que se necesita definir un valor de N, lo que garantiza que cuando el número de puntos de inflexión sea el valor de N, los experimentos pueden discriminar estas diferencias relativamente bien, y luego puede ser más preciso usar el umbral obtenido aquí al analizar la muestra de prueba.

Para la determinación del umbral del valor $\overline{Z}$: las estadísticas se realizan en la muestra control de acuerdo con las etapas a) y b), y luego el valor Z en cada ventana cumple con la distribución normal, y -1.28 y 1.28 son cuantiles donde la probabilidad acumulativa en la distribución normal es 0.05 y 0.95, respectivamente. De acuerdo con la necesidad, los expertos en la técnica también pueden seleccionar el valor $\overline{Z}$ como un valor con un absoluto mayor o un absoluto menor, que corresponde a una mayor probabilidad acumulativa y menor en la distribución normal, respectivamente; sin embargo, -1.28 y 1.28 son los umbrales más preferidos establecidos para la presente invención por los inventores a través de una gran cantidad de experimentos, y un umbral con un absoluto absoluto mayor que los dos valores aumentará la tasa de falsos negativos/falsos positivos en una detección resultado.

El pesquisaje no invasivo de la CNV fetal en una población adecuada es propicio para proporcionar asesoramiento genético y proporcionar una base para la toma de decisiones clínicas; y el diagnóstico prenatal puede prevenir efectivamente el nacimiento de un paciente infantil. La población adecuada puede ser todas las mujeres embarazadas sanas.

A continuación se detallarán las modalidades de la presente invención junto con ejemplos. Aquellos sin condiciones específicas indicadas en los ejemplos se realizan de acuerdo con las condiciones de rutina o las condiciones propuestas por los fabricantes. Los reactivos o instrumentos usados sin los fabricantes indicados son todos productos de rutina disponibles en el mercado. El número de artículo del fabricante de cada reactivo o estuche se encuentra entre los siguientes corchetes. Los adaptadores y las secuencias etiqueta usadas para la secuenciación se derivan del estuche de oligonutida de preparación de muestras multiplexado de Illumina Corporation.

Ejemplo I. Detección de la variación del número de copias de fragmentos grandes fetales en 1 caso de plasma materno, y detección de la variación de aneuploidía fetal en 9 casos de plasma materno

1. Extracción de ADN:

De acuerdo con el flujo operativo para el estuche TiangenDP327-02, se extrajo el ADN de los 8 casos de muestras de plasma mencionados anteriormente (ver la Tabla 1 para los números de muestra), se construyó una biblioteca para el ADN extraído de acuerdo con el flujo modificado de construcción de biblioteca estándar Illumina/Solexa, se agregaron adaptadores usados para la secuenciación a ambos extremos de las moléculas de ADN con bandas principales concentradas a 200 pb, se agregó una secuencia etiqueta diferente a cada muestra, y luego se realizó la hibridación con adaptadores complementarios en la superficie de la celda de flujo. Una capa de cebadores de cadena sencilla se unió a través de la superficie de la celda de flujo, y después de convertirse en cadenas simples, los fragmentos de ADN se "fijaron" en un extremo del chip mediante la complementación con bases de cebador en la superficie del chip; y el otro extremo (5' o 3') fue aleatoriamente complementario a otro cebador cercano y también fue "fijado" para formar un "puente", la amplificación se repitió durante 30 ciclos, y cada molécula se amplificó aproximadamente 1,000 veces para formar un conglomerado de ADN monoclonal. Luego, a través de la secuenciación de extremos pareados en IlluminaHiseq2000, se obtuvieron secuencias de fragmentos de ADN de aproximadamente 50 pb de longitud.

Específicamente, aproximadamente 10 ng de ADN obtenido de las muestras de plasma mencionadas anteriormente se sometieron al flujo de construcción modificado de biblioteca estándar Illumina/Solexa, y se observaron las instrucciones del producto para el flujo específico (las instrucciones de construcción de la biblioteca estándar Illumina/Solexa proporcionadas por http://www.illumina.com/). Se determinó el tamaño de la biblioteca de ADN y se determinó que los fragmentos insertados eran de aproximadamente 200 pb a través de 2100Bioanalyzer (Agilent), y la secuenciación en el ordenador pudo realizarse después de la cuantificación precisa por QPCR.

2. Secuenciación: en este ejemplo, las muestras de ADN obtenidas de los 10 casos de plasma mencionados anteriormente se manipularon de acuerdo con las instrucciones para ClusterStation y Hiseq2000 (secuenciación PE) publicadas oficialmente por Illumina/Solexa para obtener la cantidad de datos de aproximadamente 0.36 G de cada muestra para realizar la secuenciación en el ordenador, y cada muestra se discriminó de acuerdo con dichas secuencias etiqueta. Las secuencias de ADN obtenidas por secuenciación se alinearon con la secuencia de referencia genómica humana versión 36 en la base de datos NCBI (hg18; NCBIBuild 36) de la manera en que no se permitía la falta de coincidencia utilizando el software de alineación SOAP2 (obtenido de soap.genomics.org.cn) para obtener ubicaciones de las secuencias de ADN para secuenciar en dicho genoma.

3. Análisis de Datos

a) Cálculo del número relativo de lecturas para la muestra de prueba: la longitud de las lecturas únicas de referencia se seleccionó como 50 pb, se contó el número de lecturas únicas de referencia, la secuencia de referencia genómica humana se dividió en ventanas con el mismo número de lecturas únicas de referencia (840,000), el promedio de todos los tamaños de ventana fue de 1 Mb, y la distancia entre ventanas adyacentes fue de S = 10 kb. El número real de lecturas, $r_{i,j}$, se contó la caída en cada ventana en la etapa 2 mencionado anteriormente, donde los subíndices *i* y *j* representan el número de serie de la ventana y el número de serie de la muestra respectivamente, y el contenido de GC de cada ventana, $GC_{i,j}$, se calculó y el número relativo de lecturas, $R_{i,j} = \log_2\left(\frac{r_{i,j}}{\overline{r}_j}\right)$, se calculó, donde el número promedio de lecturas es $\overline{r}_j = \frac{1}{n}\sum_{i=1}^{n} r_{i,j}$;

b) Corrección y estandarización de datos:

① en un sistema de coordenadas con el contenido de GC como abscisa y el número relativo de lecturas R como ordenada, realizar un ajuste lineal en $R_{i,j}$ y $GC_{i,j}$ de la muestra control para obtener la pendiente a_i y la intercepción b_i,

② para cada ventana de la muestra de prueba, calcular el número relativo corregido de lecturas

$$R^{0}_{i,j} = a_i \times GC_{i,j} + b_i,$$

③ para cada ventana de la muestra de prueba, calcular el número relativo estandarizado de lecturas $Z_{i,j}$:

$$Z_{i,j} = \left(R_{i,j} - \overline{R}_{i,j} - media_{\ j}\right) / SD_j \text{, donde } media_{\ j} = \frac{1}{n}\sum_{i=1}^{n}\left(R_{i,j} - \overline{R}_{i,j}\right),$$

$$SD_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(R_{i,j} - \overline{R}_{i,j} - media_{\ j}\right)^2}.$$

c) Fusión de ventanas
① inicio: la posición del punto de partida de cada ventana en la secuencia del genoma de referencia se registró como la posición del estadístico Z. Luego, en correspondencia con la posición cromosómica en el genoma de referencia, el valor de Z tuvo una tendencia de cambio. Se encontró la posición correspondiente al punto de inflexión del valor Z (es decir, un punto crítico donde el valor Z se convierte de una tendencia creciente en una tendencia decreciente o se convierte de una tendencia decreciente en una tendencia creciente). Para cualquier cromosoma, las posiciones con la distancia entre ellas de al menos 100 ventanas se seleccionaron secuencialmente comenzando desde el punto de partida de la primera ventana, y estas posiciones se registraron como puntos de inflexión candidatos b_k (k = 1, 2, ..., s; s es un entero > 0);
② iteración óptima: para estudiar el análisis de variación del número de copias o la aneuploidía en cualquier cromosoma en el genoma (en este ejemplo, solo se estudiaron los cromosomas humanos 1-22), todos los puntos de inflexión candidatos clasificados de cada cromosoma se registraron como $B_c = \{b_1, b_2, ..., b_s\}$, en donde existen dos fragmentos, uno a la izquierda y otro a la derecha, para cada punto de inflexión candidato b_k y dichos fragmentos eran una región desde el punto de inflexión anterior al punto de inflexión y una región desde el punto de inflexión hasta el siguiente punto de inflexión; el valor p (p_k) obtenido al someter a todos los $Z_{i,j}$ en estos dos fragmentos de la prueba de ejecución se consideró "la importancia de b_k como punto de inflexión "; un punto de inflexión candidato con el máximo p_k fue excluido; y esta etapa se repitió hasta que todos los valores p fueron menores que el valor de terminación p (p_{final}) para el cromosoma;
③ adquisición del valor p de terminación: en el proceso de prueba, las etapas a) a c) mencionados anteriormente ① se realizaron usando la muestra control como muestra de prueba, para el cromosoma c, todos los puntos de inflexión candidatos clasificados del cromosoma cth se registraron como $B_c = \{b_1, b_2, ..., b_s\}$, en el que existen dos ventanas, una a la izquierda y otra a la derecha, para cada punto de inflexión candidato b_k; el valor p (p_k) obtenido al someter a todos los $Z_{i,j}$ en estas dos ventanas a la prueba de ejecución se consideró como "la importancia de b_k como punto de inflexión "; se excluyó un punto de inflexión candidato de menor importancia, hasta que el número de puntos de inflexión candidatos fuera igual al valor esperado N_c ($N_c = L_c/T$, donde L_c es la longitud del cromosoma y la máxima precisión teórica T = 2 Mb); y en el conjunto de los puntos de inflexión candidatos, el valor mínimo p fue el valor de terminación p (p_{final}) del cromosoma, ver la siguiente tabla;

Valores relevantes usados en el ejemplo

Cromosoma	Longitud del cromosoma (pb)	N_c	p_{final}
1	247,249,719	123	4.45E-131
2	242,951,149	121	2.30E-169
3	199,501,827	99	4.98E-149
4	191,273,063	95	5.11E-172
5	180,857,866	90	4.47E-141
6	170,899,992	85	1.99E-127
7	158,821,424	79	2.70E-145
8	146,274,826	73	2.31E-131
9	140,273,252	70	2.99E-121
10	135,374,737	67	1.22E-206
11	134,452,384	67	2.99E-121
12	132,349,534	66	2.60E-127
13	114,142,980	57	2.10E-178
14	106,368,585	53	2.51E-67
15	100,338,915	50	2.77E-128

16	88,827,254	44	2.70E-84
17	78,774,742	39	1.27E-89
18	76,117,153	38	1.07E-143
19	63,811,651	31	1.71E-120
20	62,435,964	31	1.56E-95
21	46,944,323	23	3.96E-89
22	49,691,432	24	2.38E-111

d) filtración de fragmentos después de la fusión de ventanas: para filtrar aún más los fragmentos obtenidos después de la fusión de ventanas, el promedio de $Z_{i,j}$ en el fragmento fue calculado y recodificado como $\overline{Z}$, y si $\overline{Z}$ del fragmento era menor que -1,28 o mayor que 1,28, entonces el fragmento era una variación del número de copias. Ver la Tabla 1 para los resultados.

4) Visualización de resultados, ver Figura 2.

Tabla 1. Una lista de resultados de la CNV de todas las muestras en el caso de implementación

No.	Cromosoma	Punto de inicio de la CNV	Punto final de la CNV	Tamaño de la CNV	Análisis del resultado	Regiones y bandas involucradas
S67	4	181,243,323	191,250,465	10.1M	Deleción	4q34.3→q35.2
	7	34,983	17,074,358	17M	Duplicación	7p21.1→p22.3
S10	18	1	76,117,153	76.1M	Duplicación	18p11.32→q23
S14	21	1	46,944,323	46.9M	Duplicación	21p13→q22.3
S18	18	1	76,117,153	76.1M	Duplicación	18p11.32→q23
S49	13	1	114,142,980	114.1M	Duplicación	13p13→q34
S55	21	1	46,944,323	46.9M	Duplicación	21p13→q22.3
S82	21	1	46,944,323	46.9M	Duplicación	21p13→q22.3
S103	13	1	114,142,980	114.1M	Duplicación	13p13→q34

Los resultados del análisis de CNV en la presente invención se compararon con los resultados mediante el chip CGH a continuación, y los resultados de comparación se muestran en la siguiente Tabla 2. Para los resultados con el chip CGH, se usó el estuche de micromatrices CGH del genoma humano (Agilent Technologies Inc.).

El proveedor obtuvo el mismo de acuerdo con el protocolo, y las etapas se describen brevemente de la siguiente a continuación:

El ADN de una persona sana con el mismo sexo que la muestra a analizar o el ADN mixto de personas sanas masculinas y femeninas se usó como ADN de referencia, el ADN de referencia y el ADN a analizar se etiquetó con las fluoresceínas, Cy3 y Cy5, respectivamente, y luego se hibridó con sondas, y si la relación de intensidad de fluorescencia del ADN a analizar con respecto al ADN de referencia fuera 1, entonces podría entenderse que las cantidades de ADN a analizar y las del ADN de referencia son iguales, y si la proporción no fue igual a 1, entonces se indicó que hay deleciones o amplificaciones en el ADN a ensayar. Las resoluciones de varios tipos de Matrices CGH dependen del intervalo y la longitud de las sondas en la micromatriz. Flujo: líquido de cultivo celular restante después de que se recogió el examen cromosómico de la banda G, y se extrajeron los ADN genómicos de la muestra a analizar y la de la muestra control. Después de la purificación, la muestra a analizar y la muestra de referencia se marcaron con fluorescencia de manera diferente, y luego las muestras se mezclaron con ADN de Cot-1 bloqueando la hibridación no específica, se denaturalizó, se prealineó y se hibridó con la micromatriz, y finalmente, el ADN que no se unió a la micromatriz de manera objetivo específica se eluyó, y después del escaneo y análisis por software se obtuvo la relación de intensidad de fluorescencia de las dos señales en el objetivo en cada micromatriz, lo cual reflejó el cambio en el número de copias de una secuencia correspondiente o gen entre el ADN genómico de la muestra a analizar y el ADN genómico de la muestra de referencia.

Tabla 2. Comparación de los resultados de detección del ejemplo de la presente invención y los resultados por el chip CGH

No.	Cromosoma	Resultados por chip CGH	Resultados de detección del método de la presente invención	Análisis de resultados
S67	4	Deleción: 181,170,528-190,958,960	Deleción: 181,243,323-191,250,465	Consistente
	7	Duplicación: 204,709-16,283,844	Duplicación: 34,983-17,074,358	Consistente

Los resultados del análisis de CNV en la presente invención se compararon con los resultados del cariotipo estándar como se muestra a continuación, y los resultados de comparación se muestran en la siguiente Tabla 3. Las etapas del cariotipo estándar fueron las siguientes:

(1) El líquido amniótico obtenido por centesis se centrifugó durante 5 minutos (a una velocidad de rotación de 800-1,000 revoluciones/minuto), y luego se inoculó en una campana de inoculación. El sobrenadante se pipeteó y se retuvo para otros exámenes, 0.5 ml de líquido amniótico y células de líquido amniótico precipitado permanecieron en el tubo de centrífuga, y las células exfoliadas fetales precipitadas y las células amnióticas se pipetearon uniformemente en una suspensión celular, y se inocularon en tres matraces de cultivo que contenían una solución de cultivo.
(2) Los matraces de cultivo se colocaron en una incubadora de dióxido de carbono.
(3) 5-7 días después de la inoculación, las células viables en el líquido amniótico se adhirieron al fondo de los matraces y comenzaron a crecer, y el estado de crecimiento de las células se pudo observar con un microscopio invertido. Si se hubiera producido adherencia, la solución de cultivo podría cambiarse, se agregaron 3-5 ml de la solución de cultivo reciente, y la solución se cambió una vez cada 2-3 días después. Las células adherentes incluían células epitelioides, células similares a fibroblastos y células de líquido amniótico que eran un tipo de células con la morfología que caía entre las células epitelioides y los fibroblastos. Los tres tipos de células mencionados anteriormente formaron clones, y si el estado de crecimiento era bueno, 11-14 días después de la inoculación, podría haber más de diez clones de escamas grandes en el fondo de los matraces, a ojo descubierto también se podía ver escamas de los clones en el fondo de los matraces, y los núcleos celulares eran grandes y redondos. En este momento, se puede preparar la preparación disgregada, o la denominada cosecha. Un día antes de la cosecha, se debe cambiar la misma con la solución de cultivo fresco para aumentar la división del núcleo.
(4) Cosecha: la cosecha se realizó en promedio 14-20 días después del cultivo. Se añadió colchicina de 0.04 nanogramos/mililitro a los matraces de cultivo, las células se detuvieron en la metafase y se cultivaron durante 5-15 horas, y se pudo ver bajo un microscopio invertido que había muchos núcleos celulares en la fase mitótica, y las celdas eran redondas, grandes y brillantes como un copo de perlas brillantes, y estaban interconectadas. La cantidad de colchicina añadida podría ser diferente en varios laboratorios.
(5) Tripsinización: la solución de cultivo en los matraces de cultivo se vertió en tubos de centrífuga, se colocaron 0.5 ml de solución de digestión con tripsina con EDTA al 0.02 % o 0.5 ml de pronasa al 0.15 % en el fondo de los matraces de cultivo, los clones celulares en el fondo de los matraces se pipetearon suavemente con una pipeta de vidrio curvada larga, se vio bajo un microscopio invertido que las células clonadas flotaban, las mismas se pipetearon en los tubos de centrífuga, y luego, las células que aún no habían flotado se lavaron con 0.5-1 ml de la solución de Hank, se continuó pipeteando con la pipeta larga y se vertieron en los tubos de centrífuga después de hacer que se separaran por completo. La centrifugación se realizó durante 5 minutos a una velocidad de 800-1.000 revoluciones/minuto, se retiró el sobrenadante y las células se reservaron para su uso.
(6) Tratamiento hipotónico: Se añadieron 4 ml de solución de KCl 0.075 M a 37 °C suavemente a los tubos de centrífuga y células mencionados anteriormente, se agitó suavemente el fondo de los tubos con un dedo o las células precipitadas se dispersaron con una pipeta puntiaguda, se colocaron en un baño de agua a 37 °C durante 16 minutos (el tiempo del tratamiento hipotónico se pudo determinar de acuerdo con sus propias experiencias en varios laboratorios), y se centrifugó durante 5 minutos, se retiró el sobrenadante, la solución de fijación fresca (metanol:ácido acético glacial = 3:1) se dejó caer suavemente a lo largo de las paredes de los tubos, se agitó suavemente el fondo de los tubos con un dedo, se separaron las células uniformemente, se fijaron durante 15 minutos y se centrifugaron, se cambió la solución de fijación y después de un segundo la fijación del tiempo se realizó durante 30 minutos, durante la misma noche.
(7) Soplado de la muestra: después de la centrifugación, se retiró el sobrenadante y se retuvieron 0.5 ml y se preparó en una suspensión celular, o se retiró completamente el sobrenadante y se añadieron 0.5 ml de solución de fijación recién preparada, y después se pipeteó cuidadosamente con un tubo de vidrio largo y delgado, se pipeteó una gota, se dejó caer sobre un portaobjetos de vidrio sacado del agua helada y se dispersó soplando suavemente, y después de colocar el portaobjetos de vidrio en el aire y secarlo, se observó el estado de dispersión de los cromosomas bajo un microscopio, y luego se continuó soplando. El portaobjetos de vidrio seco podría teñirse directamente con Giemsa.
(8) Bandeo: si la morfología cromosómica era buena, se podría realizar el bandeo Giemsa, denominado bandeo G. El portaobjetos de vidrio se horneó primero a 65 °C durante 1 hora, o se horneó a 37 °C durante 24 horas, el portaobjetos de vidrio se colocó en solución de tripsina al 0.25 % durante 20-25 segundos a temperatura ambiente, se sometió a solución salina fisiológica dos veces, se colocó en solución de Giemsa al 2% durante 5-10 minutos, se sacó, se lavó con

agua corriente y se secó al aire, y los cromosomas se pudieron observar bajo un microscopio para realizar la cariotipificación.

Tabla 3. Comparación de los resultados de detección en el presente caso de implementación y los resultados de detección de cariotipo estándar

No.	Cariotipificación estándar	Resultados de detección de la presente invención	Análisis de resultados
S10	T18	T18	Consistente
S14	T21	T21	Consistente
S18	T18	T18	Consistente
S49	T13	T13	Consistente
S55	T21	T21	Consistente
S82	T21	T21	Consistente
S103	T13	T13	Consistente

Aunque las modalidades particulares de la presente invención se han detallado, los expertos en la técnica entenderán que, de acuerdo con todas las enseñanzas que se divulgaron, esos detalles pueden estar sujetos a diversas modificaciones y sustituciones.

REIVINDICACIONES

1. Un método implementado por ordenador para detectar la variación genética, que comprende las siguientes etapas:

1) adquirir lecturas de una muestra de prueba;
2) alinear las lecturas con una secuencia del genoma de referencia;
3) dividir la secuencia del genoma de referencia en ventanas, calcular el número de lecturas que alinean con cada ventana, y adquirir la estadística para cada ventana en función del número de las lecturas;
4) para un fragmento de la secuencia del genoma de referencia, sobre la base del cambio en las estadísticas de todas las ventanas contenidas en el fragmento de la secuencia del genoma de referencia, adquirir posiciones donde se produce un cambio significativo en las estadísticas de las ventanas en ambos lados, estas posiciones son posiciones donde los sitios de variación genética de la muestra de prueba están en la secuencia del genoma de referencia; y
5) examinar los sitios de variación genética para obtener sitios de variación genética posteriores al pesquisaje, en donde para cada sitio de variación genética, se realizan pruebas estadísticas sobre la diferencia entre dos grupos numéricos que consisten ene estadísticas de ventanas contenidas en el fragmento entre el sitio de variación genética y su sitio de variación genética precedente y en el fragmento entre el sitio de variación genética y su sitio de variación genética posterior, y eliminar el sitio de variación genética cuyo valor de diferencia significativo es máximo y mayor que un umbral preestablecido; y repetir el proceso mencionado anteriormente, hasta que los valores de significancia de la diferencia de los sitios de variación genética sean todos más pequeños que el umbral preestablecido; y
en donde el umbral preestablecido se adquiere mediante las siguientes etapas:

a) adquirir los sitios de variación genética de acuerdo con las etapas 1) a 4) sustituyendo la muestra de prueba con una muestra control,
b) para cada sitio de variación genética, realizar estadísticas sobre la diferencia entre dos grupos numéricos que consisten en estadísticas de ventanas contenidas en el fragmento entre el sitio de variación genética y su sitio de variación genética anterior y en el fragmento entre el sitio de variación genética y su sitio de variación posterior, y eliminar el sitio de variación genética que es el menos significativo; y
c) repetir la etapa b) mencionada anteriormente, hasta que un número de puntos de inflexión candidatos restantes sea igual al valor esperado N_c, en donde $N_c = L_c/T$, L_c es la longitud de la secuencia del genoma, la máxima precisión teórica T es el tamaño del fragmento que puede detectarse teóricamente, la precisión teórica final $T = W + S*N$ cuando el promedio de los tamaños de ventana es W, la longitud deslizante de las ventanas es S y el número de cada grupo de ventanas en la prueba de ejecución es N, y entre los valores de significación de todos los puntos de inflexión candidatos restantes, el mínimo es el umbral de significación .

2. Un método implementado por ordenador para detectar la variación genética como se reivindicó en la reivindicación 1, en donde la longitud de las lecturas es 25-100 nt.

3. Un método implementado por ordenador para detectar la variación genética como se reivindicó en la reivindicación 1, en donde la longitud de las lecturas es 35-100 nt.

4. Un método implementado por ordenador para detectar la variación genética como se reivindicó en la reivindicación 1, en donde el número de las lecturas es de al menos 1 millón.

5. Un método implementado por ordenador para detectar la variación genética como se reivindicó en la reivindicación 1, en donde las ventanas tienen el mismo número de lecturas únicas de referencia.

6. Un método implementado por ordenador para detectar la variación genética como se reivindicó en la reivindicación 1, en donde las ventanas tienen una superposición o no tienen superposición entre ellas.

7. Un método implementado por ordenador para detectar la variación como se reivindicó en la reivindicación 1, en donde la estadística se ajusta aproximadamente a la distribución normal obtenida por un proceso de estandarización en el número de lecturas que se alinean a una ventana.

8. Un método implementado por ordenador para detectar la variación genética como se reivindicó en la reivindicación 7, en donde la estandarización se basa en el número promedio de lecturas que están alineadas con todas las ventanas.

9. Un método implementado por ordenador para detectar la variación genética como se reivindicó en la reivindicación 1, en donde el sitio de variación genética es el punto medio entre un punto de inflexión donde la estadística cambia de ascendente a descendente y el siguiente mismo punto de inflexión, y hay al menos 50, al menos 70, al menos 100, preferentemente 100 longitudes de ventana entre dos sitios de variación genética.

10. Un método implementado por ordenador para detectar la variación genética como se reivindicó en la reivindicación 1, en donde en la etapa 5) el valor de diferencia de significación se realiza mediante la prueba de ejecución, eliminando el sitio de variación genética cuyo valor de importancia en la prueba de ejecución es máximo y mayor que el umbral preestablecido; y repitiendo el proceso mencionado anteriormente, hasta que los valores de significación de los sitios de variación genética en la prueba de ejecución sean todos más pequeños que el umbral preestablecido.

11. Un método implementado por ordenador para detectar la variación genética como se reivindicó en cualquier reivindicación precedente, que comprende las siguientes etapas:

1) adquirir los sitios de variación genética en un fragmento de la secuencia del genoma de referencia de acuerdo con el método de cualquier reivindicación anterior; y
2) realizar una selección basada en la confianza en fragmentos entre los sitios de variación genética.

12. Un método implementado por ordenador para detectar la variación genética como se reivindicó en la reivindicación 11, en donde la etapa 2) es:

i) calcular la probabilidad de distribución de las estadísticas a través del patrón de distribución de las estadísticas para las ventanas, y establecer un umbral; y
ii) comparar el promedio de las estadísticas de ventanas en el fragmento entre los sitios de variación genética posteriores al pesquisaje con el umbral, y determinar si el fragmento entre los sitios genéticos es anómalo sobre la base del resultado de la comparación.

13. Un método implementado por ordenador para detectar la variación genética como se reivindicó en la reivindicación 11, en donde la etapa 2) es:

i) calcular la probabilidad de distribución de las estadísticas a través del patrón de distribución de las estadísticas para las ventanas, y establecer un primer umbral y un segundo umbral; y
ii) comparar el promedio de las estadísticas de ventanas en el fragmento entre los sitios de variación genética posteriores al pesquisaje con el primer umbral y segundo umbral,
en donde, si las estadísticas para ventanas en el fragmento son más pequeñas que el primer umbral, el fragmento es una deleción de fragmentos, y si las mismas son mayores que el segundo umbral, el fragmento es una duplicación de fragmentos.

14. Un método implementado por ordenador para detectar la variación genética como se reivindicó en la reivindicación 13, en donde el primer umbral es un valor del estadístico donde la probabilidad acumulativa es 0.05, y/o el segundo umbral es un valor del estadístico donde la probabilidad acumulativa es 0.95.

15. Un método implementado por ordenador para detectar la variación genética como se reivindicó en cualquier reivindicación anterior, en donde la muestra de prueba es una muestra materna que contiene ácido nucleico fetal; preferentemente en donde la muestra materna es sangre periférica materna.

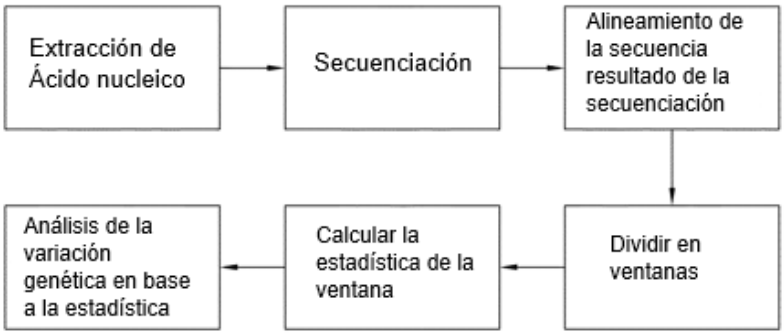

Figura 1

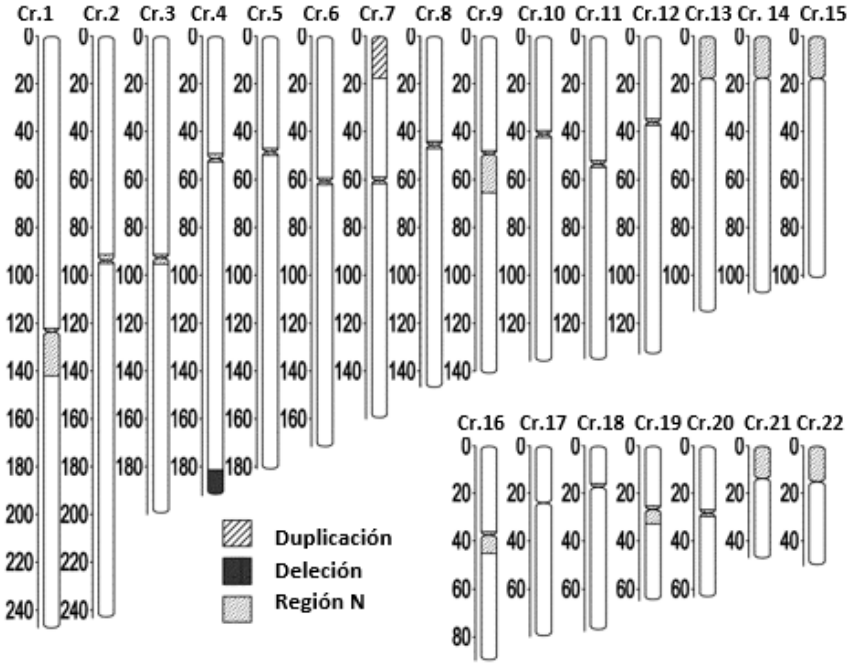

Figura 2A

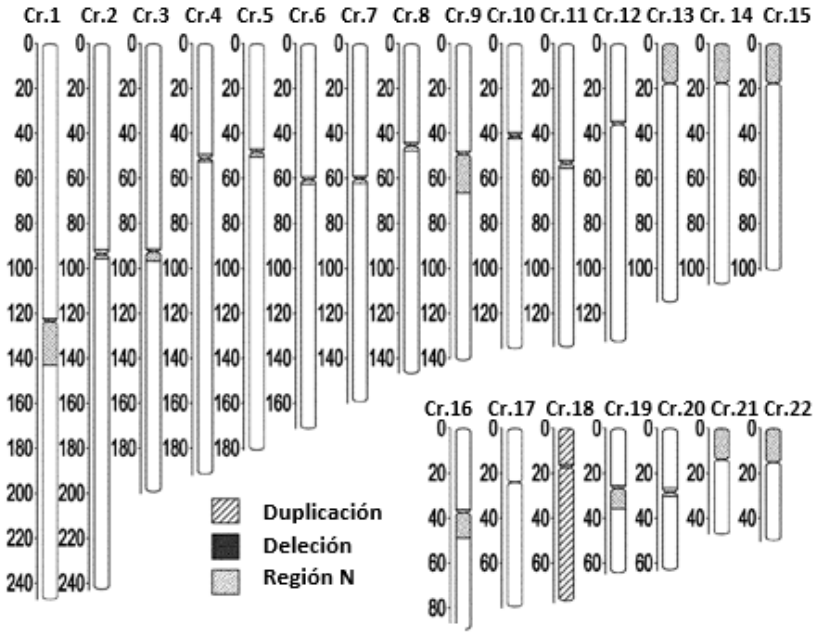

Figura 2B

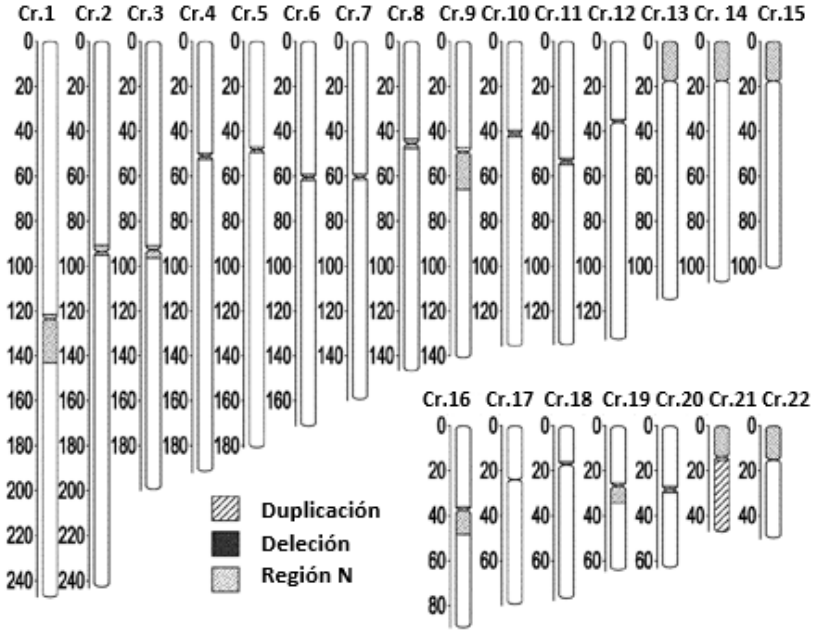

Figura 2C

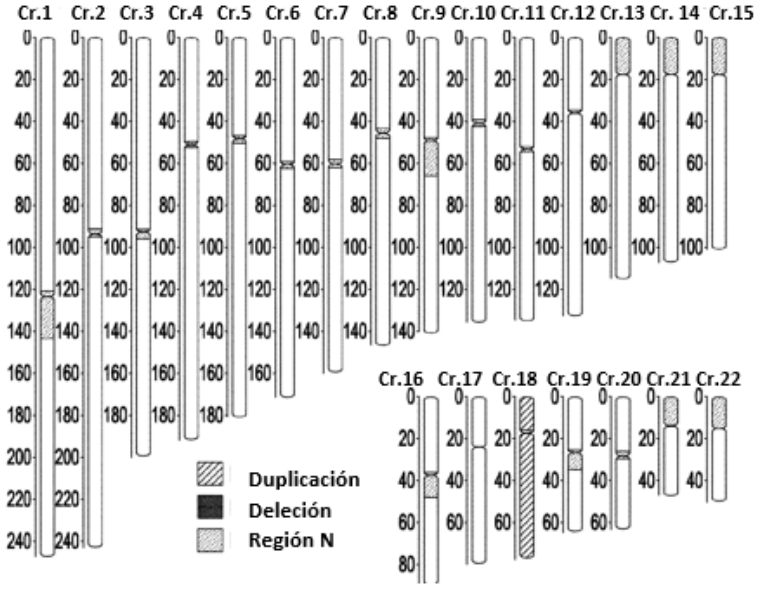

Figura 2D

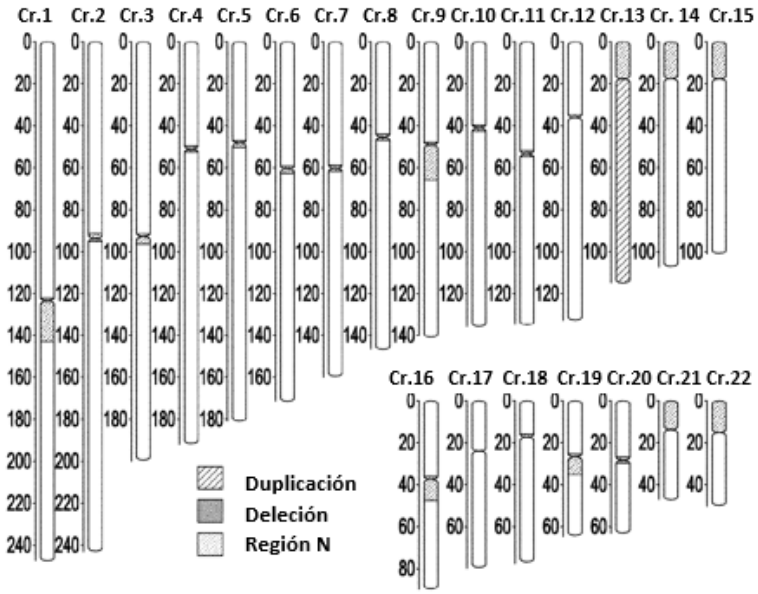

Figura 2E

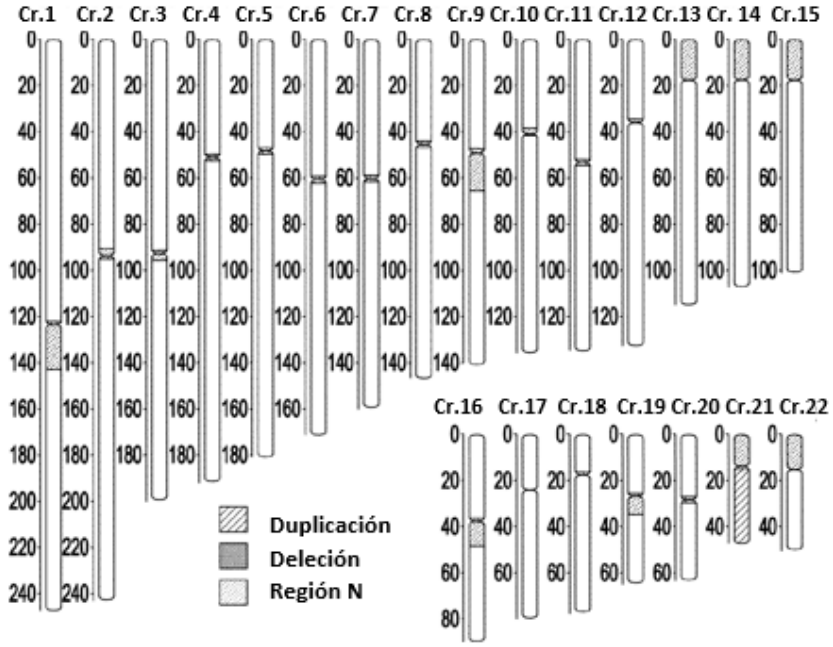

Figura 2F

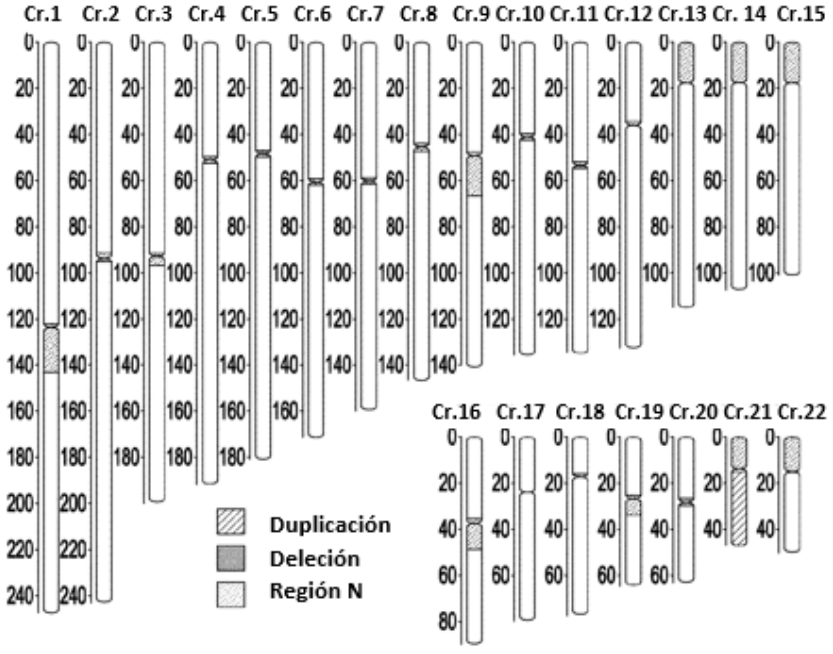

Figura 2G

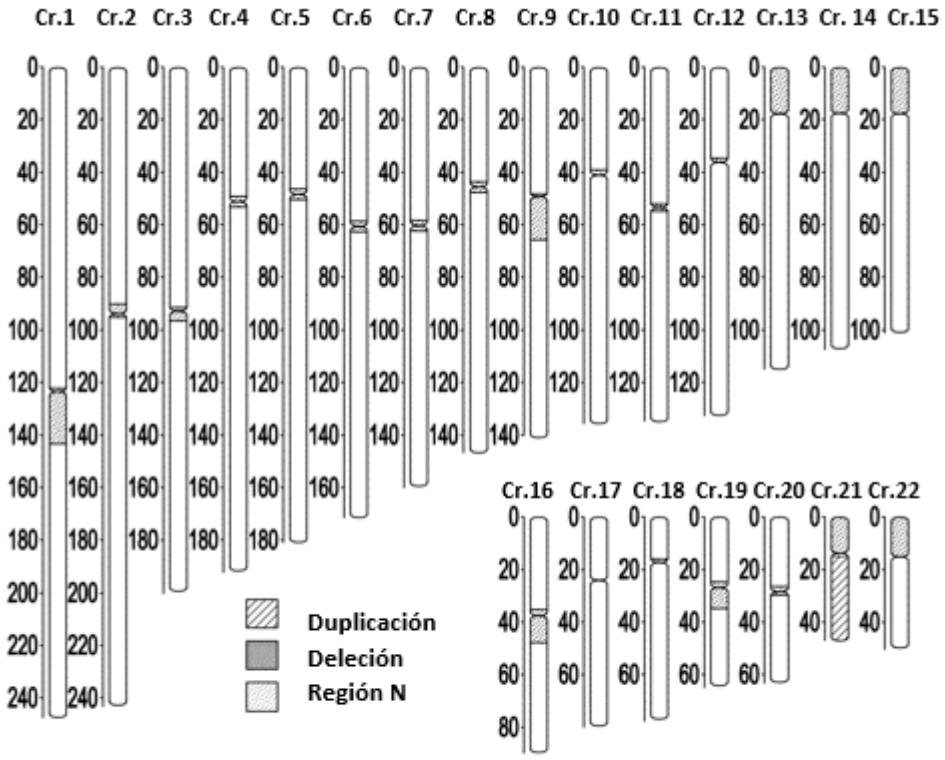
Cr.1 Cr.2 Cr.3 Cr.4 Cr.5 Cr.6 Cr.7 Cr.8 Cr.9 Cr.10 Cr.11 Cr.12 Cr.13 Cr. 14 Cr.15
Cr.16 Cr.17 Cr.18 Cr.19 Cr.20 Cr.21 Cr.22
Duplicación
Deleción
Región N

Figura 2H