(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 656 023**

(51) Int. CI.:

C12Q 1/68 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **14.05.2012 PCT/CN2012/075478**

(87) Fecha y número de publicación internacional: **21.11.2013 WO13170429**

(96) Fecha de presentación y número de la solicitud europea: **14.05.2012 E 12876917 (1)**

(97) Fecha y número de publicación de la concesión europea: **13.12.2017 EP 2851431**

(54) Título: **Procedimiento, sistema y medio legible por ordenador para determinar la información de bases en una región predeterminada del genoma del feto**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
22.02.2018

(73) Titular/es:

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

(72) Inventor/es:

CHEN, SHENGPEI;
GE, HUIJUAN;
LI, XUCHAO;
YI, SHANG;
WANG, JIAN;
WANG, JUN;
YANG, HUANMING y
ZHANG, XIUQING

(74) Agente/Representante:

DURAN-CORRETJER, S.L.P

DESCRIPCIÓN

Procedimiento, sistema y medio legible por ordenador para determinar la información de bases en una región predeterminada del genoma del feto

SECTOR TÉCNICO

La presente invención se refiere, en general, a un procedimiento de determinación de información de bases de una región predeterminada en un genoma fetal, y a un sistema y a un medio legible por ordenador del mismo.

ANTECEDENTES

Las enfermedades genéticas son un tipo de enfermedades causadas por cambios de materiales genéticos, que tienen las características de ser congénitas, familiares, permanentes y hereditarias. Las enfermedades genéticas pueden clasificarse en 3 clases: enfermedad monogenética, trastorno poligenético y anomalía cromosómica, en la que la enfermedad monogenética es sobre todo debida a una anormalidad de la función genética causada por la herencia dominante o recesiva de un único gen causante de la enfermedad; mientras que el trastorno poligenético es un tipo de enfermedad causada por una pluralidad de cambios en los genes, que pueden estar influenciados por el entorno externo hasta cierto grado; y la anormalidad cromosómica incluye la anormalidad en el número y anormalidad en la estructura, siendo el ejemplo más común el Síndrome de Down que resulta de la trisomía 21, de la cual un paciente niño presenta rasgos congénitos, tales como mongolismo y forma del cuerpo anormal, etc. Dado que hasta ahora no hay tratamientos terapéuticos eficaces para enfermedades genéticas, se pueden realizar sólo pertinentemente tratamientos de apoyo o de remisión con fármacos con alto coste, lo que puede llevar cargas pesadas, tanto en la economía como en el ánimo de la sociedad y las familias. De este modo, es extremadamente necesario hacer un trabajo preventivo mediante la detección del estado patológico con un feto antes de nacer para lograr el propósito de una buena atención prenatal y postnatal. El feto en sí mismo no se utiliza en la presente invención.

Sin embargo, el procedimiento de detección relacionado aún debe mejorarse.

El documento WO 2007/092473 A2 da a conocer un cribado genético fetal no invasivo mediante el análisis digital que implica el análisis de la sangre materna para una enfermedad genética, en el que se analiza el ADN mezclado del feto y de la madre en la sangre materna para distinguir la mutación fetal o la anormalidad genética de la línea base del ADN materno.

El documento US 2012/0095697 A1 da a conocer un procedimiento para determinar el número de copias de una región genómica en una posición de detección de una secuencia diana en una muestra, en el que se corrige la preferencia de la cobertura de la secuencia y se puede normalizar frente a una muestra de línea base. Se realizan la segmentación, evaluación y obtención de resultados del modelo oculto de Markov (HMM) para estimar un valor total del número de copias y el valor del número de copias específico de una región para una pluralidad de regiones.

El documento WO 2012/021149 A1 se refiere al modelo oculto de Markov y a otros procedimientos de aprendizaje mediante máquinas que se utilizan juntos en la detección de nanoporos con la utilización de bloqueadores de canales que son transductores de eventos o informadores de eventos, permitiendo así la detección de la transducción de nanoporos y la biosensibilización de nanoporos.

W. Timp y otros en Biophysical Journal, volumen 102, No. 10, mayo 2012, L37-L39, dan a conocer la secuencia de ADN basada en nanoporos. Se demuestra un procedimiento para decodificar mediciones eléctricas de nanoporos con una resolución de 3 pb en una secuencia de ADN utilizando un modelo oculto de Markov.

Mingjie Lin y otros en la International Conference on Reconfigurable Computing and FPGAs, Reconfig '09, IEEE, 2009, págs. 95-100, dan a conocer un procedimiento de computación reconfigurable basado en el aprendizaje de redes bayesianas dinámicas propuesto para requerir bases en la pirosecuenciación a partir de los datos de expresión de genes en micromatrices.

Kuo-ching Liang y otros en IEEE/ACM TRANSACTIONS ON COMPUTACIONAL BIOLOGY AND BIOINFORMATICS, volumen 4, No. 3, julio-septiembre 2007, págs. 430-440 y Kuo-ching Ling y otros en la 40ª Annual Conference on Information Sciences and Systems, IEEE, 2006, págs. 1599-1604 dan a conocer más procedimientos matemáticos para el análisis de la secuencia de ADN utilizando el modelo oculto de Markov.

CARACTERÍSTICAS DE LA INVENCIÓN

La presente invención busca resolver, como mínimo, uno de los problemas existentes en la técnica relacionada, hasta, como mínimo, cierto grado.

Según un primer aspecto de la presente invención, se da a conocer un procedimiento de determinación de la

información de bases de una región predeterminada en un genoma fetal, según la reivindicación 1. Según la presente invención, el procedimiento comprende las etapas, tal como se definen en la reivindicación 1. La formación del genoma de la descendencia es igual a una recombinación aleatoria con el genoma de la generación parental (es decir, un intercambio de recombinación de haplotipos y una combinación aleatoria de gametos). Para el plasma de una mujer embarazada, si se presupone un haplotipo fetal (una recombinación de haplotipos parentales) como estados ocultos, los datos de secuenciación del plasma se pueden utilizar como observaciones (secuencia de observación), probabilidades de transición, probabilidades de símbolos de observación y se puede deducir la distribución del estado inicial gracias a los datos anteriores, a continuación, se puede determinar la recombinación más posible de haplotipos fetales utilizando un modelo oculto de Markov basado en el algoritmo de Viterbi a fin de obtener más información del feto antes del nacimiento. De este modo, según las realizaciones de la presente divulgación, gracias al modelo oculto de Markov, por ejemplo, utilizando el algoritmo de Viterbi, y con referencia a la información genética de un individuo relacionado, se puede determinar la secuencia de ácidos nucleicos de una región predeterminada en un genoma fetal, mediante lo cual se puede realizar de manera eficaz una detección genética prenatal con información genética del genoma fetal.

Según un segundo aspecto de la presente invención, se da a conocer un sistema para determinar la información de bases de una región predeterminada en un genoma fetal, según la reivindicación 9. Según la presente invención, el sistema comprende las etapas, tal como se definen en la reivindicación 9. Utilizando el sistema se puede implementar de manera efectiva el procedimiento anterior de determinación de la información de bases de una región predeterminada en un genoma fetal, el cual determina la secuencia de ácidos nucleicos de una región predeterminada en un genoma fetal gracias al modelo oculto de Markov, por ejemplo utilizando el algoritmo de Viterbi, y con referencia a la información genética de un individuo relacionado, mediante lo cual se puede realizar de manera eficaz una detección genética prenatal con información genética del genoma fetal.

Según un tercer aspecto de la presente invención, se da a conocer un medio legible por ordenador, según la reivindicación 14. Según la presente invención, el medio legible por ordenador, que incluye una pluralidad de instrucciones, está adaptado para determinar la información de bases de una región predeterminada basándose en el resultado de la secuenciación de un feto que se combina con la información genética de un individuo relacionado utilizando un modelo oculto de Markov que provoca que un aparato realice las etapas, tal como se define en las reivindicación 14. Utilizando el medio legible por ordenador de la presente divulgación se pueden ejecutar de manera eficaz la pluralidad de instrucciones mediante un procesador para determinar una secuencia de ácidos nucleicos de la región predeterminada en el genoma fetal gracias al modelo oculto de Markov, por ejemplo utilizando el algoritmo de Viterbi basado en los datos de secuenciación del feto que se combina con la información genética de un individuo relacionado, mediante lo cual se puede realizar de manera eficaz una detección genética prenatal con información genética del genoma fetal.

En las siguientes descripciones se proporcionarán, en parte, aspectos y ventajas adicionales de realizaciones de la presente divulgación, serán evidentes, en parte, a partir de las siguientes descripciones, o se aprenderán a partir de la práctica de las realizaciones de la presente divulgación.

DESCRIPCIÓN BREVE DE LOS DIBUJOS

Estos y otros aspectos y ventajas de las realizaciones de la presente divulgación serán evidentes y se entenderán más fácilmente a partir de las siguientes descripciones realizadas con referencia a los dibujos que se acompañan, en los que:
la figura 1 es un diagrama de flujo que muestra un procedimiento de análisis que utiliza modelo oculto de Markov según una realización de la presente divulgación; y
la figura 2 es un diagrama esquemático que muestra un sistema para determinar la información de bases de una región predeterminada en un genoma fetal, según una realización de la presente divulgación.

DESCRIPCIÓN DETALLADA

Se hará referencia en detalle a las realizaciones de la presente divulgación. Los elementos iguales o similares y los elementos que tienen funciones iguales o similares se designan con números de referencia similares a lo largo de las descripciones. Las realizaciones descritas en el presente documento con referencia a los dibujos son explicativas, ilustrativas, y se utilizan para entender, en general, la presente divulgación. Las realizaciones no se interpretarán para limitar la presente divulgación.

Cabe indicar que términos, tales como "primero" y "segundo", se utilizan en el presente documento para fines de descripción y no pretenden indicar o implicar una importancia o significancia relativa. De este modo, las características definidas con "primero", "segundo" pueden incluir explícita o implícitamente una o más de las características. Además, en la descripción de la presente divulgación, a menos que se indique lo contrario, "una/la pluralidad de" significa dos o más.

Procedimiento de determinación de la información de bases de una zona predeterminada en un genoma fetal

En un primer aspecto de la presente divulgación, se da a conocer un procedimiento de determinación de información de bases de una región predeterminada en un genoma fetal. Según las realizaciones de la presente divulgación, el procedimiento puede comprender:
en primer lugar, extraer una muestra de ADN genómico de un feto de sangre periférica de una mujer embarazada, construir una biblioteca de secuenciación basada en una muestra de ADN genómico de un feto. Según la presente divulgación, la fuente de la muestra de ADN genómico del feto es sangre periférica de una mujer embarazada. Utilizando la sangre periférica de una mujer embarazada como fuente de la muestra de ADN genómico del feto se puede realizar de manera eficaz la obtención de la muestra de ADN genómica del feto mediante un muestreo no invasivo, mediante lo cual el genoma fetal se puede monitorizar de manera eficaz con la suposición de que no tiene influencia en el desarrollo normal del crecimiento fetal. En cuanto a los procedimientos y procesos de construcción de una biblioteca de secuenciación para la muestra de ácidos nucleicos, un experto en la materia puede seleccionarlos de manera apropiada dependiendo de la diferente tecnología de secuenciación. El proceso detallado se puede referir al procedimiento proporcionado por el fabricante del secuenciador, tal como Illumina Company, por ejemplo, se hace referencia a Multiplexing Sample Preparation Guide (parte #1005361; febrero de 2010) o Paired-End SamprePrep Guide (parte #1005063; febrero de 2010) de Illumina Company, que se incorporan en el presente documento como referencia. Según las realizaciones de la presente divulgación, los procedimientos y dispositivos para la extracción de un ácido nucleico de una muestra biológica no están sometidos a ningún tipo de restricción especial, que se pueden llevar a cabo utilizando un kit de extracción de ácidos nucleicos comercial.

Después de construirse, se aplica la biblioteca de secuenciación obtenida a un secuenciador para obtener un resultado de la secuenciación correspondiente que consiste en una pluralidad de datos de secuenciación. Según las realizaciones de la presente divulgación, los procedimientos y dispositivos para la secuenciación no están sometidos a ningún tipo de restricción especial, incluyendo, pero sin limitarse a los mismos, el procedimiento de terminación de cadena (Sanger); es preferente un procedimiento de secuenciación de alto rendimiento. De este modo, utilizando características de secuenciación de alto rendimiento y profundas de estos aparatos, puede mejorarse adicionalmente la eficacia, mediante lo cual se puede mejorar adicionalmente la precisión y exactitud del análisis posterior con los datos de secuenciación, tal como una prueba estadística. El procedimiento de secuenciación de alto rendimiento incluye, pero sin limitarse a las mismas, una tecnología de secuenciación Next-Generation o una tecnología de secuenciación única. La plataforma de secuenciación Next-Generation (Metzker ML. Sequencing technologies-the next generation. *Nat Rev Genet*. Enero 2010; 11 (1): 31-46) incluye, pero sin limitarse a las mismas, las plataformas de secuenciación Illumina-Solexa (GATM, HiSeq2000TM) y así sucesivamente), ABI-Solid y Roche-454 (pirosecuenciación); la plataforma (tecnología) de secuenciación única incluye, pero sin limitarse a las mismas, secuenciación True Single Molecule DNA de Helicos Company, secuenciación de una única molécula en tiempo real (SMRT®) de Pacific Biosciences Company y la tecnología de secuenciación por nonaporos de Oxford Nanopore Technologies (Rusk, Nicole (01-04-2009). Cheap Third-Generation Sequencing. *Nature Methods* 6 (4): 244-245), etc. Con el desarrollo gradual de la tecnología de secuenciación, un experto en la materia puede comprender otros procedimientos de secuenciación y también se pueden utilizar aparatos para la secuenciación del genoma completo. Según los ejemplos específicos de la presente divulgación, la biblioteca de secuenciación del genoma completo se puede someter a secuenciación, como mínimo, mediante un procedimiento seleccionado de Illumina-Solexa, ABI-SOLID, Roche-454 y un aparato de secuenciación de una única molécula.

Opcionalmente, después de obtenerse, el resultado de la secuenciación puede alinearse con una secuencia de referencia para determinar los datos de secuenciación que corresponden a la región predeterminada. La expresión "región predeterminada" utilizada en el presente documento debe entenderse en sentido amplio, refiriéndose a cualquier región de una molécula de ácido nucleico que contiene un posible evento predeterminado. Para el análisis de SNP, puede ser una región que contiene un sitio de SNP. Para el análisis de aneuploidía cromosómica, la región predeterminada se refiere a todo el cromosoma a analizar o una parte del mismo, es decir, la selección de los datos de secuenciación que derivan del cromosoma. Los procedimientos de selección de los datos de secuenciación que derivan de una región correspondiente en el resultado de la secuenciación no están sometidos a ninguna restricción especial. Según las realizaciones de la presente divulgación, todos los datos de secuenciación obtenidos pueden alinearse con una secuencia de referencia con un ácido nucleico conocido para obtener los datos de secuenciación que derivan de la región predeterminada. Además, según las realizaciones de la presente divulgación, la región predeterminada también puede ser una pluralidad de puntos de dispersión que no son discontinuos en un genoma. Según las realizaciones de la presente divulgación, un tipo de la secuencia de referencia utilizada puede no estar sometido a ningún tipo de restricción especial, la cual puede ser cualquiera de las secuencias conocidas contenidas en una región diana. Según las realizaciones de la presente divulgación, la secuencia de referencia puede utilizar un genoma de referencia humano conocido. Por ejemplo, según las realizaciones de la presente divulgación, el genoma de referencia humano es NCBI 36.3, HG18. Además, según las realizaciones de la presente divulgación, los procedimientos de alineación no están sometidos a ninguna restricción especial. Según ejemplos específicos, se puede utilizar SOAP para la alineación.

A continuación, se determina una parte de una secuencia de ácidos nucleicos de la región predeterminada basándose en los datos de secuenciación correspondientes a la región predeterminada; y se determinan otras partes de la secuencia de ácidos nucleicos basándose en la parte determinada de la secuencia de ácidos nucleicos

de la región predeterminada utilizando el algoritmo de Viterbi para obtener la secuencia de ácidos nucleicos de la región predeterminada. Según la presente invención, la información de bases de la región predeterminada se determina basándose en el resultado de la secuenciación del feto combinándose con la información genética de un individuo relacionado utilizando un modelo oculto de Markov. Según las realizaciones de la presente divulgación, la información de bases de la región predeterminada se determina utilizando el modelo oculto de Markov llevado a cabo basándose en el algoritmo de Viterbi. De este modo, se puede realizar de manera eficaz una detección genética prenatal con la información genética del genoma fetal.

Haciendo referencia a la figura 1, se describe a continuación en detalle una guía para el análisis utilizando el algoritmo de Viterbi en base a un modelo oculto de Markov.

En el sentido genético, la expresión "un individuo relacionado" se refiere a individuos que tienen una relación genética con un feto. Por ejemplo, según las realizaciones de la presente divulgación, "un individuo relacionado" puede ser una generación de alquiler de un feto, tal como los padres. De este modo, la formación del genoma de la descendencia es igual a una recombinación aleatoria con el genoma de la generación parental (es decir, un intercambio de recombinación de haplotipos y una combinación aleatoria de gametos). Para el plasma de una mujer embarazada, si se presupone un haplotipo fetal (una recombinación de haplotipos parentales) como estados ocultos, los datos de secuenciación del plasma se pueden utilizar como observaciones (secuencia de observación), probabilidades de transición, probabilidades de símbolos de observación y se puede deducir la distribución del estado inicial gracias a los datos anteriores, a continuación, se puede determinar la recombinación más posible de haplotipos fetales utilizando un modelo oculto de Markov basado en el algoritmo de Viterbi a fin de obtener más información del feto antes del nacimiento.

Las etapas de análisis se muestran a continuación en detalle:

Marcador:

I. El número de sitios a detectar es N.

II. Los haplotipos de los padres se registran respectivamente como $FH = \{fh_0, fh_1\}$ y $MH = \{mh_0, mh_1\}$
en los que
$mh_k = \{m_{1,k}, \ldots, m_{i,k}, \ldots, m_{N,k}\}$, $fh_k = \{f_{1,k}, \ldots, f_{i,k}, \ldots, f_{N,k}\}$,
$\forall fh_{i,k}, mh_{i,k} \in \{A, C, G, T\}$,
$k \in \{0, 1\}$, $i = 1, 2, 3, \ldots, N$.

III. El haplotipo fetal desconocido se registra como $H = \{h_0, h_1\}$, en particular, h_0 y h_1 representan, respectivamente, la herencia de la madre y del padre.

$$h_0 = \{m_{1,x_1}, \ldots, m_{i,x_i}, \ldots, m_{N,x_N}\} \quad , \quad h_1 = \{f_{1,y_1}, \ldots, f_{i,y_i}, \ldots, f_{N,y_N}\}$$

en las que $x_i \in \{0, 1\}$, $y_i \in \{0, 1\}$,

Los subíndices x_i e y_i presentan, respectivamente, pares de secuencias, y $q_i = \{x_i, y_i\}$ representa los estados ocultos que deben decodificarse.

Mientras todos los estados ocultos posibles que se presentan constituyen un conjunto Q.

IV. Los datos de secuenciación se registran como $S = \{s_1, \ldots, s_i, \ldots, s_N\}$
en los que $s_i = \{n_{i,A}, n_{i,C}, n_{i,G}, n_{i,G}\}$ representa la información de secuenciación de un sitio que contiene el número de cuatro bases A, C, T y G.

V. La concentración fetal promedio y la tasa de error de secuenciación promedio se registran, respectivamente, como ε y e.

Etapa 1, construir un vector de distribución de probabilidades de un estado inicial y una matriz de transición de la recombinación de haplotipos:

I. La distribución de probabilidades de los estados iniciales se registra como $\pi = \{\pi_j\}$ ($j \in Q$).

Según las realizaciones de la presente divulgación, en las circunstancias de no tener datos de referencia, se puede suponer que $\pi_j = \Pr(q_1 = j) \triangleq \frac{1}{4}$, es decir, las posibilidades de que cada estado oculto esté presente en el primer sitio son iguales.

II. Según las realizaciones de la presente divulgación, la probabilidad de recombinación de haplotipos se registra como $p_r = re/N$, en la que *re* representa los tiempos promedio de recombinaciones de gametos humanos, variando los datos anteriores de 25 a 30.

III. Según las realizaciones de la presente divulgación, una matriz de transición de la recombinación de haplotipos se registra como $A = \{a_{jk}\}$ ($j, k \in Q$), en la que a_{jk} representa la probabilidad de transición de estados ocultos, es decir,

$$a_{jk} = \Pr\left(q_i = k \middle| q_{i-1} = j\right) = \begin{cases} \left(1-p_r\right)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ \left(1-p_r\right)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad o \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases},$$

Los subíndices x_i e y_i de haplotipos fetales $h_0 = \{m_{1,x_1}, ..., m_{i,x_i}, ..., m_{N,x_N}\}$ y $h_1 = \{f_{1,y_1}, ..., f_{i,y_i}, ..., f_{N,y_N}\}$ constituyen un par de secuencias, $q_i = \{x_i, y_i\}$ constituyen los estados ocultos que deben decodificarse. Por ejemplo, $x_i = 0$ representa "en un cromosoma materno, un alelo en el locus correspondiente es $m_{i,0}$".

Etapa 2, la construcción de una matriz de probabilidades de las observaciones.

Según las realizaciones de la presente divulgación, la matriz de probabilidades de las observaciones se registra como $B = \{b_{i,j}(s_i)\}$ (i = 1, 2, 3,..., N, $j \in Q$), en la que $b_{i,j}(s_i)$ representa "una posibilidad observada de esta información de secuenciación en un sitio i, considerando un haplotipo materno y un haplotipo fetal (estado j, $j = \{x_i, y_i\}$)", es decir,

$$b_{i,j}(s_i) = \Pr\left(s_i \middle| q_i = j, \{m_0, m_1\}\right)$$
$$= \frac{(n_{i,A}+n_{i,C}+n_{i,G}+n_{i,T})!}{n_{i,A}!n_{i,C}!n_{i,G}!n_{i,T}!}\cdot\left(P_{i,A}\right)^{n_{i,A}}\cdot\left(P_{i,C}\right)^{n_{i,C}}\cdot\left(P_{i,G}\right)^{n_{i,G}}\cdot\left(P_{i,T}\right)^{n_{i,T}},$$

en la que $P_{i,base}$ representa "una posibilidad de una base en un sitio i, considerando un haplotipo materno y un haplotipo fetal (estado j, $j = \{x_i, y_i\}$)", es decir,

$$P_{i,base} = \Pr\left(base \middle| q_i = j, \{m_0, m_1\}\right)$$
$$= \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta\left(base, m_k\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, m_{x_i}\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, f_{y_i}\right),$$

en la que, una función indicadora es

$$\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}.$$

Dicha etapa es aplicar el parámetro HMM, calculando una distribución de probabilidades de observaciones en cada sitio $b_{i,j}(s_i)$, es decir, calculando una posibilidad que presentan los datos de secuenciación actuales (observaciones) en el plasma de la mujer embarazada, suponiendo diferentes haplotipos fetales en cada sitio.

Etapa 3, la construcción de una matriz de probabilidades parciales y un cursor de inversión (tomando un ejemplo de construcción de una matriz de probabilidades unidimensional):

Definición: probabilidad parcial $\delta_i(q_i) = \left(\max_{q_{i-1}\in Q}\delta_i(q_i)\cdot a_{q_{i-1}q_i}\right)\cdot b_{i,q_i}(s_i)$,

Definición: cursor de inversión $\Psi_i(q_i) = \arg\max_{q_{i-1}\in Q}\delta_i(q_i)\cdot a_{q_{i-1}q_i}$.

Las expresiones de "probabilidad parcial $\delta_i(q_i)$" y "cursor de inversión $\Psi_i(q_i)$" utilizadas en el presente documento siguen las definiciones clásicas del algoritmo de Viterbi. Para las descripciones detalladas de la definición del parámetro se puede hacer referencia a Lawrence R. Rabiner, PROCEEDINGS OF THE IEEE, volumen 77, No. 2, febrero de 1989, que se incorpora en el presente documento como referencia.

Etapa 4, determinar un estado final y retroceder por una ruta opcional

Determinación del estado final, $q_N^* = \arg\max_{q_N\in Q}\delta_N(q_N)$.

El haplotipo fetal más posible $q_i^* = \Psi_i(q_i)$ (i = 1, 2, 3,..., N-1) se obtiene mediante retroceso por la ruta opcional basándose en el cursor de inversión.

Etapa 5, salida de resultados

De este modo, se puede analizar de manera eficaz la secuencia del genoma fetal. En comparación con otro procedimiento existente de detección prenatal, el procedimiento de la presente divulgación puede tener las siguientes ventajas técnicas, que incorpora principalmente la precisión y cantidad de información genética que puede obtenerse:

1) Según las realizaciones de la presente divulgación, un sitio a detectar no se limita a un sitio parental, para un sitio materno, es decir, un sitio heterocigoto materno, si un feto hereda un sitio de patopoiesia materna también puede detectarse de forma excelente, con una precisión de hasta el 95% o más; y se puede detectar una pluralidad de tipos de anormalidad que agranda el intervalo de detección de la enfermedad.

2) Según las realizaciones de la presente divulgación, la información de una pluralidad de sitios y enfermedades se puede obtener mediante una sola secuenciación; mientras que las secuencias de genes, que tienen una baja cobertura en el plasma de la mujer embarazada que no se puede determinar con precisión solamente mejorando la profundidad de la secuenciación, se pueden obtener mediante el procedimiento de la presente divulgación con un resultado preciso y fiable.

3. Según las realizaciones de la presente divulgación, se puede realizar una representación gráfica con una enfermedad genética, algunas enfermedades relacionadas se pueden deducir directamente con la información de otros sitios, con una gran cantidad de información obtenida en una vez, lo cual tiene un significado más instructivo para la detección clínica.

Además, según las realizaciones de la presente divulgación, el procedimiento de determinación de la información de bases de una región predeterminada en un genoma fetal, no limitado a ciertos sitios polimórficos genéticos, tales como SNP o STR, está adaptado para todos los sitios polimórficos genéticos, el cual se puede utilizar en paralelo para una pluralidad de sitios para verificarse entre sí. Además de aplicar la información genómica de detección no invasiva prenatal de un feto, consiguiendo el propósito de la detección de la enfermedad, el procedimiento de la presente divulgación también se puede utilizar en la identificación de paternidad prenatal no invasiva, es decir, la determinación de la identidad del padre del feto antes del nacimiento, la disposición de asistencia para disputas que implican responsabilidades y obligaciones de crianza, bienes y casos de agresión sexual, etc.

Sistema para determinar la información de bases de una región predeterminada en un genoma fetal

En otro aspecto de la presente divulgación, se da a conocer un sistema para determinar la información de bases de una región predeterminada en un genoma fetal. Según las realizaciones de la presente divulgación, en referencia a la figura 2, el sistema -1000- puede comprender: un aparato de construcción de una biblioteca -100-, un aparato de secuenciación -200- y un aparato de análisis -400-.

Según las realizaciones de la presente divulgación, aparato de construcción de una biblioteca -100- está adaptado para construir una biblioteca de secuenciación basándose en una muestra de ADN genómico de un feto. Según las realizaciones de la presente divulgación, el aparato de secuenciación -200- está conectado al aparato de construcción de una biblioteca -100- y está adaptado para someter la biblioteca de secuenciación a secuenciación para obtener un resultado de la secuenciación del feto que consiste en una pluralidad de datos de secuenciación. Según las realizaciones de la presente divulgación, el sistema -1000- puede comprender también un aparato de extracción de muestras de ADN adaptado para extraer la muestra de ADN genómico del feto a partir de sangre periférica de mujer embarazada. De este modo, el sistema puede estar adaptado para la detección prenatal no invasiva.

Según las realizaciones de la presente divulgación, opcionalmente, el sistema puede comprender también un aparato de alineación -300-. Según las realizaciones de la presente divulgación, el aparato de alineación -300- está conectado al aparato de secuenciación -200- y está adaptado para alinear el resultado de la secuenciación del feto con una secuencia de referencia, para determinar el resultado de la secuenciación que deriva de la región predeterminada. Según las realizaciones de la presente divulgación, los procedimientos y dispositivos para la secuenciación no están sometidos a ningún tipo de restricción especial, incluyendo, pero sin limitarse a los mismos, procedimiento de terminación de cadena (Sanger); es preferente un procedimiento de secuenciación de alto rendimiento. De este modo, utilizando características de secuenciación de alto rendimiento y profundas de estos aparatos, puede mejorarse adicionalmente la eficacia, mediante lo cual se puede mejorar adicionalmente la precisión y exactitud del análisis posterior con los datos de secuenciación, tal como una prueba estadística. El procedimiento de secuenciación de alto rendimiento incluye, pero sin limitarse a las mismas, una tecnología de secuenciación Next-Generation o una tecnología de secuenciación única. La plataforma de secuenciación Next-Generation (Metzker ML. Sequencing technologies-the next generation. *Nat Rev Genet*. Enero 2010; 11 (1): 31-46) incluye, pero sin limitarse a las mismas, las plataformas de secuenciación Illumina-Solexa (GATM, HiSeq2000TM), etc), ABI-Solid y Roche-454 (pirosecuenciación); la plataforma (tecnología) de secuenciación única incluye, pero sin limitarse a las mismas, secuenciación True Single Molecule DNA de Helicos Company, secuenciación de una única molécula en tiempo real (SMRT®) de Pacific Biosciences Company y la tecnología de secuenciación por nonaporos de Oxford Nanopore

Technologies (Rusk, Nicole (01-04-2009). Cheap Third-Generation Sequencing. *Nature Methods* 6 (4): 244-245), etc. Con el desarrollo gradual de la tecnología de secuenciación, un experto en la materia puede comprender otros procedimientos de secuenciación y también se pueden utilizar aparatos para la secuenciación del genoma completo. Según los ejemplos específicos de la presente divulgación, la biblioteca de secuenciación del genoma completo se puede someter a secuenciación, como mínimo, mediante un procedimiento seleccionado de Illumina-Solexa, ABI-SOLID, Roche-454 y un aparato de secuenciación de una única molécula. Según las realizaciones de la presente divulgación, un tipo de secuencia de referencia utilizada puede no estar sometido a ningún tipo de restricción especial, la cual puede ser cualquiera de las secuencias conocidas contenidas en una región diana. Según las realizaciones de la presente divulgación, la secuencia de referencia puede utilizar un genoma de referencia humano conocido. Por ejemplo, según las realizaciones de la presente divulgación, el genoma de referencia humano es NCBI 36.3, HG18. Además, según las realizaciones de la presente divulgación, los procedimientos de alineación no están sometidos a ninguna restricción especial. Según ejemplos específicos, se puede utilizar SOAP para la alineación.

Según las realizaciones de la presente divulgación, el aparato de análisis -400- está conectado al aparato de secuenciación y está adaptado para determinar la información de bases de la región predeterminada basándose en el resultado de la secuenciación del feto en combinación con información genética de un individuo relacionado utilizando un modelo oculto de Markov.

Según las realizaciones de la presente divulgación, en el algoritmo de Viterbi, se utiliza 0,25 como una distribución de probabilidades de un estado inicial, se utiliza re/N como una probabilidad de recombinación, siendo re 25~30, de manera preferente, siendo re 25 y siendo N la longitud de la región predeterminada,

$$a_{jk} = \Pr\left(q_i = k \middle| q_{i-1} = j\right) = \begin{cases} \left(1-p_r\right)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ \left(1-p_r\right)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad o \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases},$$

se utiliza como una matriz de transición de la recombinación, siendo p_r re/N.

Según las realizaciones de la presente divulgación, el aparato de alineación está adaptado para determinar una base que tiene la probabilidad más alta basándose en una fórmula de

$$P_{i,base} = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta\left(base, m_k\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, m_{x_i}\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, f_{y_i}\right)$$

en la que

$$\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}.$$

El análisis con datos de secuenciación, que se describe en detalle anteriormente, está también adaptado al sistema para determinar la información de bases de una región predeterminada en un genoma fetal, el cual se omite por razones de brevedad.

De este modo, utilizando el sistema se puede implementar de manera eficaz el procedimiento anterior de determinación de la información de bases de una región predeterminada en un genoma fetal, el cual puede determinar la secuencia de ácidos nucleicos de una región predeterminada en un genoma fetal gracias al modelo oculto de Markov, por ejemplo utilizando el algoritmo de Viterbi, y con referencia a la información genética de un individuo relacionado, mediante lo cual se puede realizar de manera eficaz una detección genética prenatal con la información genética del genoma fetal.

Además, según las realizaciones de la presente divulgación, la región predeterminada es un sitio en el que se ha determinado previamente que tiene un polimorfismo genético y el polimorfismo genético es, como mínimo, un polimorfismo seleccionado de polimorfismo de nucleótido único y STR.

El término "conectado" debe entenderse ampliamente, el cual puede referirse a una conexión directa o una conexión indirecta, siempre que se consiga la conexión funcional anterior.

Cabe señalar que un experto en la materia puede entender que las características y ventajas del procedimiento de determinación de la información de bases de una región predeterminada en un genoma fetal descrito anteriormente también pueden adaptarse al sistema para determinar la información de bases de una región predeterminada en un genoma fetal, lo cual se omite por razones de brevedad.

Medio legible por ordenador

En un aspecto adicional de la presente divulgación, se da a conocer un medio legible por ordenador. Según las realizaciones de la presente divulgación, el medio legible por ordenador incluye una pluralidad de instrucciones adaptadas para determinar la información de bases de una región predeterminada basándose en un resultado de la secuenciación de un feto en combinación con la información genética de un individuo relacionado utilizando un modelo oculto de Markov. De este modo, utilizando el medio legible por ordenador se puede implementar de manera eficaz el procedimiento anterior de determinación de la información de bases de una región predeterminada en un genoma fetal, el cual puede determinar la secuencia de ácidos nucleicos de una región predeterminada en un genoma fetal gracias al modelo oculto de Markov, por ejemplo utilizando el algoritmo de Viterbi, y con referencia a la información genética de un individuo relacionado, mediante lo cual se puede realizar de manera eficaz una detección genética prenatal con la información genética del genoma fetal.

Según las realizaciones de la presente divulgación, la pluralidad de instrucciones está adaptada para determinar la información de bases de la región predeterminada utilizando el modelo oculto de Markov basándose en el algoritmo de Viterbi. Según las realizaciones de la presente divulgación, en el algoritmo de Viterbi, se utiliza 0,25 como una distribución de probabilidades de un estado inicial, se utiliza re/N como una probabilidad de recombinación, siendo re 25~30, de manera preferente, siendo re 25 y siendo N la longitud de la región predeterminada,

$$a_{jk} = \Pr\left(q_i = k \middle| q_{i-1} = j\right) = \begin{cases} \left(1-p_r\right)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ \left(1-p_r\right)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad o \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases},$$

se utiliza como una matriz de transición de la recombinación, siendo p_r re/N.

Según las realizaciones de la presente divulgación, la pluralidad de instrucciones están además adaptadas para determinar una base que tiene la probabilidad más alta basándose en una fórmula de

$$P_{i,base} = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta\left(base, m_k\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, m_{x_i}\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, f_{y_i}\right)$$

en la que

$$\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}.$$

El análisis con datos de secuenciación, que se describe en detalle anteriormente, está también adaptado al medio legible por ordenador, el cual se omite por razones de brevedad.

Además, según las realizaciones de la presente divulgación, la región predeterminada es un sitio previamente determinado que tiene un polimorfismo genético y el polimorfismo genético es, como mínimo, uno seleccionado de polimorfismo de nucleótido único y STR.

En cuanto a la memoria descriptiva, "medio legible por ordenador" puede ser cualquier dispositivo adaptativo para incluir, almacenar, comunicar, propagar o transferir programas para utilizar con el sistema, dispositivo o equipo de ejecución de instrucciones o en combinación con los mismos. Los ejemplos más específicos del medio legible por ordenador comprenden, pero no sin limitarse a los mismos: una conexión electrónica (un dispositivo electrónico) con uno o más cables, una carcasa para ordenador portátil (un dispositivo magnético), una memoria de acceso aleatorio (RAM), una memoria de sólo lectura (ROM), una memoria de sólo lectura programable y borrable (EPROM o una memoria flash), un dispositivo de fibra óptica y disco compacto portátil de memoria de sólo lectura (CD-ROM). Además, el medio legible por ordenador puede ser incluso un papel u otro medio apropiado capaz de imprimir programas sobre el mismo, esto es porque, por ejemplo, el papel u otro medio apropiado se pueden escanear ópticamente y, a continuación, editarse, descifrarse o procesarse con otros procedimientos apropiados cuando sea necesario para obtener los programas de una manera eléctrica y, a continuación, los programas se pueden almacenar en las memorias de los ordenadores.

Se debe entender que cada parte de la presente divulgación se puede realizar mediante el hardware, software, firmware o su combinación. En las realizaciones anteriores, se puede realizar una pluralidad de etapas o procedimientos mediante el software o firmware almacenado en la memoria y ejecutarse mediante el sistema de ejecución de instrucciones apropiado. Por ejemplo, si se realiza mediante el hardware, del mismo modo que en otra

realización, las etapas o procedimientos se pueden realizar mediante una de las siguientes técnicas conocidas en el sector o una combinación de las mismas: un circuito lógico discreto que tiene un circuito de puertas lógicas para realizar una función lógica de una señal de datos, un circuito integrado específico de la aplicación que tiene un circuito de puertas lógicas en una combinación apropiada, una matriz de puertas programables (PGA), una matriz de puertas programables por campo (FPGA), etc.

Los expertos en la materia entenderán que todas las etapas o partes de las mismas en el procedimiento de ejemplo anterior de la presente divulgación se pueden conseguir mediante el control del hardware relacionado con los programas. Los programas se pueden almacenar en un medio de almacenamiento legible por ordenador y los programas comprenden una etapa o una combinación de etapas en las realizaciones del procedimiento de la presente divulgación cuando se ejecutan en un ordenador.

Además, cada célula de función de las realizaciones de la presente divulgación puede estar integrada en un módulo de procesamiento, o estas células pueden tener una existencia física separada, o dos o más células están integradas en un módulo de procesamiento. El módulo integrado se puede realizar en forma de módulos de función de hardware o de software. Cuando el módulo integrado se realiza en forma de un módulo de función de software y se comercializa o se utiliza como un producto independiente, el módulo integrado se puede almacenar en un medio de almacenamiento legible por ordenador.

Se hará referencia en detalle a los ejemplos de la presente divulgación. Los expertos en la materia entenderán que los siguientes ejemplos son explicativos. Si la tecnología o las condiciones específicas no están especificadas en los ejemplos, se realizará una etapa según las técnicas o condiciones descritas en la bibliografía en la técnica (por ejemplo, en referencia a J. Sambrook y otros (traducido por Huang PT), Molecular Cloning: a Laboratory Manual, 3ª edición, Science Press) o según las instrucciones del producto. Si no se especifican los fabricantes de los reactivos o instrumentos, los reactivos o instrumentos pueden estar disponibles comercialmente, por ejemplo, de la empresa Illumina.

Procedimiento general

El procedimiento, según las realizaciones de la presente divulgación, comprende principalmente las siguientes etapas:
1) tomar muestras de forma no invasiva de una muestra de una mujer embarazada que contiene materiales genéticos fetales, extraer el ADN genómico de las mismas;
2) extraer y purificar la muestra de ADN genómico de los miembros de la familia del feto, tales como los padres o abuelos del mismo;
3) construir una biblioteca de secuenciación con cada material genético según un requisito para una plataforma de secuenciación diferente;
4) filtrar los datos de secuenciación obtenidos con criterios de filtración basados en el valor de la calidad, la contaminación del adaptador y así sucesivamente;
5) ensamblar las secuencias de alta calidad obtenidas, según se requiera, alinear un resultado ensamblado con una secuencia de referencia del genoma humano para obtener secuencias asignadas de manera exclusiva para el análisis utilizando el modelo.

Modelo de análisis

Marcador:

I. El número de sitios a detectar es N.

II. Los haplotipos de los padres se registran respectivamente como $FH = \{fh_0, fh_1\}$ y $MH = \{mh_0, mh_1\}$ en las que
$mh_k = \{m_{1,k}, \ldots, m_{i,k}, \ldots, m_{N,k}\}$, $fh_k = \{f_{1,k}, \ldots, f_{i,k}, \ldots, f_{N,k}\}$,
$\forall fh_{i,k}, mh_{i,k} \in \{A, C, G, T\}$,
$k \in \{0, 1\}$, $i = 1, 2, 3, \ldots, N$.

III. El haplotipo fetal desconocido se registra como $H = \{h_0, h_1\}$, en particular, h_0 y h_1 representan, respectivamente, la herencia de la madre y del padre.

$$h_0 = \{m_{1,x_1}, \ldots, m_{i,x_i}, \ldots, m_{N,x_N}\} \quad , \quad h_1 = \{f_{1,y_1}, \ldots, f_{i,y_i}, \ldots, f_{N,y_N}\}$$

en las que $x_i \in \{0, 1\}$, $y_i \in \{0, 1\}$,

Los subíndices x_i e y_i presentan, respectivamente, pares de secuencias, y $q_i = \{x_i, y_i\}$ representa los estados ocultos que deben decodificarse.

Si bien, todos los estados ocultos posibles que se presentan constituyen un conjunto Q.

IV. Los datos de secuenciación se registran como $S = \{s_1, \ldots, s_i, \ldots, s_N\}$
en los que $s_i = \{n_{i,A}, n_{i,C}, n_{i,G}, n_{i,G}\}$ representa la información de secuenciación de un sitio que contiene el número de cuatro bases A, C, T y G.

V. La concentración fetal promedio y la tasa de error de secuenciación promedio se registran, respectivamente, como ε y e.

Etapa 1, construir un vector de distribución de probabilidades de un estado inicial y una matriz de transición de la recombinación de haplotipos:

I. La distribución de probabilidades de los estados iniciales se registra como $\pi = \{\pi_j\}$ (j ∈ Q).

Según las realizaciones de la presente divulgación, en las circunstancias de no tener datos de referencia, se puede suponer que $\pi_j = \Pr(q_1 = j) \triangleq \frac{1}{4}$, es decir, las posibilidades de que cada estado oculto esté presente en el primer sitio son iguales.

II. Según las realizaciones de la presente divulgación, la probabilidad de recombinación de haplotipos se registra como $p_r = re/N$, en la que *re* representa los tiempos promedio de recombinaciones de gametos humanos, variando los datos anteriores de 25 a 30.

III. Según las realizaciones de la presente divulgación, una matriz de transición de la recombinación de haplotipos se registra como $A = \{a_{jk}\}$ ($j, k \in$ Q), en la que a_{jk} representa la probabilidad de transición de estados ocultos, es decir,

$$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad o \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases},$$

Los subíndices x_i e y_i de haplotipos fetales $h_0 = \{m_{1,x_1}, \ldots, m_{i,x_i}, \ldots, m_{N,x_N}\}$ y $h_1 = \{f_{1,y_1}, \ldots, f_{i,y_i}, \ldots, f_{N,y_N}\}$ constituyen un par de secuencias, $q_i = \{x_i, y_i\}$ constituyen los estados ocultos que deben decodificarse. Por ejemplo, $x_i = 0$ representa "en un cromosoma materno, un alelo en el locus correspondiente es $m_{i,0}$".

Etapa 2, la construcción de una matriz de probabilidades de las observaciones:

Según las realizaciones de la presente divulgación, la matriz de probabilidades de las observaciones se registra como $B = \{b_{i,j}(s_i)\}$ (i = 1, 2, 3,..., N, $j \in$ Q), en la que $b_{i,j}(s_i)$ representa "una posibilidad observado de esta información de secuenciación en un sitio i, considerando un haplotipo materno y un haplotipo fetal (estado j, $j = \{x_i, y_i\}$)", es decir,

$$b_{i,j}(s_i) = \Pr(s_i | q_i = j, \{m_0, m_1\}) = \frac{(n_{i,A}+n_{i,C}+n_{i,G}+n_{i,T})!}{n_{i,A}!n_{i,C}!n_{i,G}!n_{i,T}!} \cdot (P_{i,A})^{n_{i,A}} \cdot (P_{i,C})^{n_{i,C}} \cdot (P_{i,G})^{n_{i,G}} \cdot (P_{i,T})^{n_{i,T}},$$

en la que $P_{i,base}$ representa "una posibilidad de una base en un sitio i, considerando un haplotipo materno y un haplotipo fetal (estado j, $j = \{x_i, y_i\}$)", es decir,

$$P_{i,base} = \Pr(base | q_i = j, \{m_0, m_1\}) = \sum_{k \in \{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta(base, m_k) + \frac{1}{2}\varepsilon \cdot \Delta(base, m_{x_i}) + \frac{1}{2}\varepsilon \cdot \Delta(base, f_{y_i}),$$

en la que, una función indicadora es $\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}$.

Etapa 3, la construcción de una matriz de probabilidades parciales y un cursor de inversión (tomando un ejemplo de construcción de una matriz de probabilidades unidimensional):

Definición: probabilidad parcial $\delta_i(q_i) = \left(\max_{q_{i-1} \in Q} \delta_i(q_i) \cdot a_{q_{i-1}q_i}\right) \cdot b_{i,q_i}(s_i)$,

Definición: cursor de inversión $\Psi_i(q_i) = \arg\max_{q_{i-1} \in Q} \delta_i(q_i) \cdot a_{q_{i-1}q_i}$.

Etapa 4, determinar un estado final y retroceder por una ruta opcional

Determinación del estado final, $q_N^* = \arg\max_{q_N \in Q} \delta_N(q_N)$.

El haplotipo fetal más posible $q_i^* = \Psi_i(q_i)$ (i = 1, 2, 3,..., N-1) se obtiene mediante el retroceso por la ruta opcional basándose en el cursor de inversión.

Etapa 5, salida de resultados

EJEMPLO 1

Obtención y tratamiento de las muestras

(1) La muestra recogida incluía: sangre periférica extraída de un padre y una madre embarazada dentro de una familia, y la sangre del cordón umbilical fetal después del nacimiento, todos ellas se recogieron en un tubo que contenía EDTA para anticoagulación; se recogió saliva de los cuatro abuelos utilizando un kit OG-250 de recogida de saliva/purificación de ADN Oragene® DNA.

(2) El ADN de saliva extraído de los cuatro abuelos se sometió al genotipado utilizando chips de genes HD Human610-Quad BeadChip Infinium®.

(3) La sangre periférica recogida de la madre embarazada se centrifugó a 1.600 g a 4°C durante 10 minutos para separar las células y el plasma de la sangre. A continuación, el plasma obtenido se centrifugó a 16.000 g a 4°C durante 10 minutos para eliminar adicionalmente los leucocitos residuales para obtener un plasma final de la madre embarazada. A continuación, se extrajo el ADN genómico del plasma final de la madre embarazada utilizando el kit TIANamp Micro DNA (Tiangen) para obtener una mezcla de ADN genómico de la madre y el feto de la misma. A continuación, se extrajo el ADN genómico materno de los leucocitos residuales extraídos. El ADN de plasma obtenido se sometió a la construcción de bibliotecas basándose en el requisito para el secuenciador HiSeg2000® de Illumina®. Las bibliotecas construidas se sometieron a una prueba de distribución utilizando el Bioanalizador 2100 de Agilent® para cumplir el requisito para los intervalos de los fragmentos. A continuación, dos bibliotecas se sometieron a cuantificación utilizando el procedimiento de Q-PCR. Las bibliotecas calificadas se sometieron a secuenciación utilizando el secuenciador HiSeg2000® de Illumina®, con un ciclo de secuenciación de índice PE101 (es decir, secuenciación con un índice de 101 pb en el extremo del par), en el que la configuración y las operaciones de los parámetros se basaron en las especificaciones de Illumina® (obtenidas en http://www.lumina.com/support/documentation.ilmn).

(4) Se extrajeron sangre periférica de los padres, leucocitos extraídos de sangre periférica materna y sangre de cordón umbilical fetal con sus respectivos ADN genómico utilizando el kit TIANamp Micro DNA (TIANGEN).

A excepción de la muestra de ADN de plasma, todas las muestras de ADN obtenidas necesitaron fragmentarse utilizando Covaris® para tener una longitud de 500 pb. Los fragmentos de ADN obtenidos y la muestra de ADN de plasma se sometieron a la construcción de bibliotecas basándose en el requisito para secuenciador HiSeg2000® de Illumina®, con un procedimiento detallado:

Sistema de reacción de reparación de los extremos:

10x tampón de T4 polinucleótido quinasa	10 µl
dNTP (10 mM)	4 µl
T4 ADN polimerasa	5 µl
fragmentos Klenow	1 µl
T4 polinucleótido quinasa	5 µl
fragmentos de ADN	30 µl
ddH_2O	hasta 100 µl

Después de reaccionar a 20°C durante 30 minutos, se utilizó el kit de purificación de PCR (QIAGEN) en el reciclaje de productos reparados en los extremos. A continuación, los productos reparados en los extremos reciclados se disolvieron finalmente en 34 µl de tampón EB.

Sistema de reacción para la adición de base A en el extremo:

10x tampón Klenow	5 µl
dATP (1 mM)	10 µl
Klenow (3'-5' exo)	3 µl
ADN	32 µl

Después de incubar a 37°C durante 30 minutos, los productos obtenidos se purificaron mediante el kit de purificación de PCR MinElute® (QIAGEN) y se disolvieron en 12 µl de tampón EB para obtener muestras de ADN al que se ha añadido base A en el extremo.

Sistema de reacción adaptador ligante:

2x tampón ligante de ADN rápido	25 µl
mezcla de oligos adaptadores PEI (20 µM)	10 µl
T4 ADN ligasa	5 µl
Muestra de ADN al que se ha añadido base A en el extremo	10 µl

Después de reaccionar a 20°C durante 15 minutos, se utilizó el kit de purificación de PCR (QIAGEN) en el reciclaje de productos ligados. Los productos ligados se disolvieron finalmente en 32 µl de tampón EB.

Sistema de reacción de PCR:

Producto ligado	10 µl
Mezcla de ADN polimerasas de Phusion	25 µl
Cebador de PCR (10 pmol/µl)	1 µl
Índice N (10 pmol/µl)	1 µl
Agua UltraPure®	13 µl

El procedimiento de reacción se muestra a continuación:

98°C	30 s	
98°C	10 s	10 ciclos
65ºC	30 s	
72ºC	30 s	
72°C	5 minutos	
4ºC	Mantenimiento	

Se utilizó el kit de purificación de PCR (QIAGEN) en el reciclaje de productos de PCR, los cuales finalmente se disolvieron en 50 µl de tampón EB.

Las bibliotecas construidas se sometieron a una prueba de distribución utilizando el Bioanalizador 2100 de Agilent® para cumplir el requisito para los intervalos de fragmentos. A continuación, se sometieron dos bibliotecas a cuantificación utilizando el procedimiento Q-PCR. Las bibliotecas calificadas se sometieron a secuenciación utilizando el secuenciador HiSeg2000® de Illumina®, con un ciclo de secuenciación de índice PE101 (es decir, secuenciación con un índice de 101 pb en el extremo del par), en el que la configuración y las operaciones de los parámetros se basan en las especificaciones de Illumina® (obtenidas en http://www.lumina.com/support/documentation.ilmn)

(5) Genotipado de secuenciación de genomas parentales y maternos
a. Los datos de secuenciación se alinearon a un genoma de referencia humano (Hg19, NCBI 36.3) utilizando SOAP2.
b. Los datos obtenidos se sometieron a la construcción de la secuencia consenso (CNS) utilizando SOAPsnp (se utilizaron miles de datos de planificación para los datos de pedigrí Southern Han (CHS)
c. Se extrajeron los genotipos de un sitio marcador.

(6) Determinación de los haplotipos de los padres
a. Construcción de una matriz de genotipos por grupos que contiene los genotipos de los ancestros y de los padres, es decir, la extracción de genotipos en el sitio del marcador de los padres, los ancestros y pedigrí Southern Han.
b. Deducción de los haplotipos de los padres utilizando BEAGLE.

(7) Determinación del haplotipo fetal
a. Alinear los datos de secuenciación de plasma a un genoma de referencia humano ((Hg19, NCBI 36.3) utilizando SOAP2;
b. Construir un vector de probabilidad de estados iniciales y una matriz de transición de la recombinación de haplotipos,
construir el vector de probabilidad de estados iniciales: tomar un modelo de datos que no son de referencia, es decir, las probabilidades de cada estado inicial eran iguales, siendo 0,25.
construir la matriz de transición de la recombinación de haplotipos: de forma conservadora, *re* = 25 (otras eran iguales que las descritas en el "procedimiento general");

c. Calcular la información de secuenciación de cada sitio y la construcción de una matriz de probabilidades de las observaciones (otras eran iguales que las descritas en el "procedimiento general");
d. La construcción de una matriz de probabilidades parciales y un cursor de inversión (otras eran iguales que las descritas en el "procedimiento general");
e. Determinar un estado final y retroceder por una ruta opcional; y
f. Obtener resultados

Según los resultados del genotipado, la exactitud del mismo se muestra a continuación:

			madre						total		
			homocigosis			heterocigosis					
			número de sitio	número exacto	exactitud	número de sitio	número exacto	exactitud	número de sitio	número exacto	exactitud
autosoma	padre	homocigosis	199.552	199.552	100,00%	66.238	63.968	96,57%	265.790	263.520	99,15%
		heterocigosis	65.409	64.735	98,97%	41.849	39.944	95,45%	107.258	104.679	97,60%
			264.961	264.287	99,75%	108.087	103.912	96,14%	373.048	368.199	98,70%
cromosoma X			4.881	4.881	100,00%	1.718	1.478	86,03%	6.599	6.359	96,36%

Aplicabilidad industrial

El procedimiento de determinación de la información de bases de una región predeterminada en un genoma fetal, el sistema para determinar la información de bases de una región predeterminada en un genoma fetal y un medio legible por ordenador, según la presente invención, se pueden aplicar de manera eficaz en el análisis de la secuencia de ácidos nucleicos de la región predeterminada en el genoma fetal.

La referencia a lo largo de la presente memoria descriptiva a "una realización", "algunas realizaciones", "otro ejemplo", "un ejemplo", "un ejemplo específico" o "algunos ejemplos", significa que un rasgo, estructura, material o característica particular descritos en relación con la realización o ejemplo se incluyen, como mínimo, en una realización o ejemplo de la presente divulgación. De este modo, las apariciones de frases, tales como "en algunas realizaciones", "en una realización", "en otro ejemplo", "en un ejemplo", "en un ejemplo específico" o "en algunos ejemplos" en varios lugares a lo largo de la presente memoria descriptiva no necesariamente se refieren a la misma realización o ejemplo de la presente divulgación. Además, los rasgos, estructuras, materiales o características particulares se pueden combinar en cualquier manera adecuada en una o más realizaciones o ejemplos.

REIVINDICACIONES

1. Procedimiento para determinar la información de bases de una región predeterminada en un genoma fetal, que comprende las siguientes etapas:
extraer una muestra de ADN genómico de un feto de sangre periférica de una mujer embarazada,
construir una biblioteca de secuenciación basándose en una muestra de ADN genómico del feto;
someter la biblioteca de secuenciación a secuenciación para obtener un resultado de la secuenciación del feto que consiste en una pluralidad de datos de secuenciación; y
determinar la información de bases de la región predeterminada basándose en el resultado de la secuenciación del feto en combinación con la información genética de un individuo relacionado utilizando un modelo oculto de Markov,
en el que la información de bases de la región predeterminada comprende la recombinación más posible de haplotipos fetales,
y la información de bases de la región predeterminada se obtiene mediante:
la construcción de un vector de distribución de probabilidades de un estado inicial y una matriz de transición de la recombinación de haplotipos;
la construcción de una matriz de probabilidades de las observaciones;
la construcción de una matriz de probabilidades parciales y un cursor de inversión;
la determinación de un estado final y el retroceso por una ruta opcional; y
la obtención de un resultado para analizar la secuencia de ADN del genoma fetal.

2. Procedimiento, según la reivindicación 1, en el que la biblioteca de secuenciación se somete a secuenciación, como mínimo, mediante un procedimiento seleccionado de Illumina-Solexa, ABI-Solid, Roche-454 y un aparato de secuenciación de una única molécula.

3. Procedimiento, según la reivindicación 1, que comprende, además, una etapa de alineación del resultado de la secuenciación del feto con una secuencia de referencia para determinar el resultado de la secuenciación que deriva de la región predeterminada,
en el que la etapa de alineación del resultado de la secuenciación del genoma fetal con la secuencia de referencia para determinar el resultado de la secuenciación que deriva de la región predeterminada comprende, además:
determinar una base que tiene la probabilidad más elevada basándose en la fórmula:

$$P_{i,base} = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta(base,m_k) + \frac{1}{2}\varepsilon\cdot\Delta(base,m_{x_i}) + \frac{1}{2}\varepsilon\cdot\Delta(base,f_{y_i})$$

en la que

$$\Delta(x,y) = \begin{cases} 1-e & x=y \\ e/3 & x\neq y \end{cases}$$

4. Procedimiento, según la reivindicación 3, en el que la secuencia de referencia es un genoma de referencia humano.

5. Procedimiento, según la reivindicación 1, en el que el individuo relacionado es un padre del feto.

6. Procedimiento, según la reivindicación 1, en el que la etapa de determinación de la información de bases de la región predeterminada utilizando el modelo oculto de Markov se realiza basándose en el algoritmo de Viterbi,
y en el algoritmo de Viterbi, se utiliza 0,25 como una distribución de probabilidades de un estado inicial, se utiliza re/N como las probabilidades de recombinación, siendo re 25~30 y siendo N la longitud de la región predeterminada,

$$a_{jk} = \Pr\left(q_i = k \middle| q_{i-1} = j\right) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad o \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases},$$

se utiliza como una matriz de transición de la recombinación, siendo p_r re/N.

7. Procedimiento, según la reivindicación 6, en el que re es 25.

8. Procedimiento, según la reivindicación 1, en el que la región predeterminada es un sitio en el que previamente se ha determinado que tiene un polimorfismo genético,
en el que el polimorfismo genético es, como mínimo, un polimorfismo seleccionado de polimorfismo de nucleótido único y STR.

9. Sistema para determinar la información de bases de una región predeterminada en un genoma fetal, que comprende:
un aparato de extracción de una muestra de ADN adaptado para extraer una muestra de ADN genómico de un feto de sangre periférica de mujer embarazada
un aparato de construcción de bibliotecas adaptado para construir una biblioteca de secuenciación basándose en una muestra de ADN genómico del feto;
un aparato de secuenciación conectado con el aparato de construcción de bibliotecas y adaptado para someter la biblioteca de secuenciación a secuenciación para obtener un resultado de la secuenciación del feto que consiste en una pluralidad de datos de secuenciación; y
un aparato de análisis conectado con el aparato de secuenciación y adaptado para determinar la información de bases de la región predeterminada basándose en el resultado de la secuenciación del feto en combinación con la información genética de un individuo relacionado utilizando un modelo oculto de Markov,
en el que la información de bases de la región predeterminada comprende la recombinación más posible de haplotipos fetales,
y la información de bases de la región predeterminada se obtiene mediante:
la construcción de un vector de distribución de probabilidades de un estado inicial y de una matriz de transición de la recombinación de haplotipos;
la construcción de una matriz de probabilidades de las observaciones;
la construcción de una matriz de probabilidades parciales y un cursor de inversión;
la determinación de un estado final y el retroceso por una ruta opcional; y
la obtención de un resultado para analizar la secuencia de ADN del genoma fetal.

10. Sistema, según la reivindicación 9, en el que el aparato de secuenciación es, como mínimo, un aparato seleccionado de Illumina-Solexa, ABI-Solid, Roche-454 y un aparato de secuenciación de una única molécula.

11. Sistema, según la reivindicación 9, que comprende, además, un aparato de alineación conectado al aparato de secuenciación y adaptado para alinear el resultado de la secuenciación del feto con una secuencia de referencia para determinar el resultado de la secuenciación que deriva de la región predeterminada,
en el que el aparato de alineación está adaptado para determinar una base que tiene la mayor probabilidad basándose en la fórmula

$$P_{i,base} = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta(base,m_k) + \frac{1}{2}\varepsilon\cdot\Delta(base,m_{x_i}) + \frac{1}{2}\varepsilon\cdot\Delta(base,f_{y_i})$$

en la que

$$\Delta(x,y) = \begin{cases} 1-e & x=y \\ e/3 & x\neq y \end{cases}.$$

12. Sistema, según la reivindicación 9, en el que el aparato de análisis está adaptado para determinar la información de bases de la región predeterminada utilizando un modelo oculto de Markov basándose en el algoritmo de Viterbi,
en el que en el algoritmo de Viterbi, se utiliza 0,25 como una distribución de probabilidades de un estado inicial, se utiliza re/N como las probabilidades de recombinación, siendo re 25~30 y siendo N la longitud de la región predeterminada,

$$a_{jk} = \Pr\left(q_i = k \middle| q_{i-1} = j\right) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad o \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases},$$

se utiliza como una matriz de transición de la recombinación, siendo p_r re/N.

13. Sistema, según la reivindicación 12, en el que re es 25.

14. Medio legible por ordenador, que comprende una pluralidad de instrucciones adaptado para determinar la información de bases de una región predeterminada basándose en un resultado de la secuenciación de un feto en combinación con la información genética de un individuo relacionado utilizando un modelo oculto de Markov que provoca que el aparato realice las etapas de:
extraer una muestra de ADN genómico de un feto de sangre periférica de una mujer embarazada,
construir una biblioteca de secuenciación basándose en una muestra de ADN genómico del feto;

someter la biblioteca de secuenciación a secuenciación para obtener un resultado de la secuenciación del feto que consiste en una pluralidad de datos de secuenciación; y
determinar la información de bases de la región predeterminada basándose en el resultado de la secuenciación del feto en combinación con la información genética de un individuo relacionado utilizando un modelo oculto de Markov, en el que la información de bases de la región predeterminada comprende la recombinación más posible de haplotipos fetales,
y la información de bases de la región predeterminada se obtiene mediante:
la construcción de un vector de distribución de probabilidades de un estado inicial y de una matriz de transición de la recombinación de haplotipos;
la construcción de una matriz de probabilidades de las observaciones;
la construcción de una matriz de probabilidades parciales y un cursor de inversión;
la determinación de un estado final y el retroceso por una ruta opcional; y
la obtención de un resultado para analizar la secuencia de ADN del genoma fetal.

15. Medio legible por ordenador, según la reivindicación 14, en el que la pluralidad de instrucciones está adaptada para determinar la información de bases de una región predeterminada utilizando el modelo oculto de Markov basándose en el algoritmo de Viterbi,
en el que en el algoritmo de Viterbi, se utiliza 0,25 como una distribución de probabilidades de un estado inicial, se utiliza re/N como las probabilidades de recombinación, siendo re 25~30 y siendo N la longitud de la región predeterminada,

$$a_{jk} = \Pr\left(q_i = k \middle| q_{i-1} = j\right) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad o \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases},$$

se utiliza como una matriz de transición de la recombinación, siendo p_r re/N.

16. Medio legible por ordenador, según la reivindicación 15, en el que re es 25.

17. Medio legible por ordenador, según la reivindicación 14, en el que la pluralidad de instrucciones están adaptadas para alinear el resultado de la secuenciación del feto con una secuencia de referencia para determinar el resultado de la secuenciación que deriva de la región predeterminada,
en el que la pluralidad de instrucciones está adaptada, además, para determinar una base que tiene la probabilidad más elevada basándose en la fórmula

$$P_{i,base} = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta\left(base, m_k\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, m_{x_i}\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, f_{y_i}\right)$$

en la que

$$\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}.$$

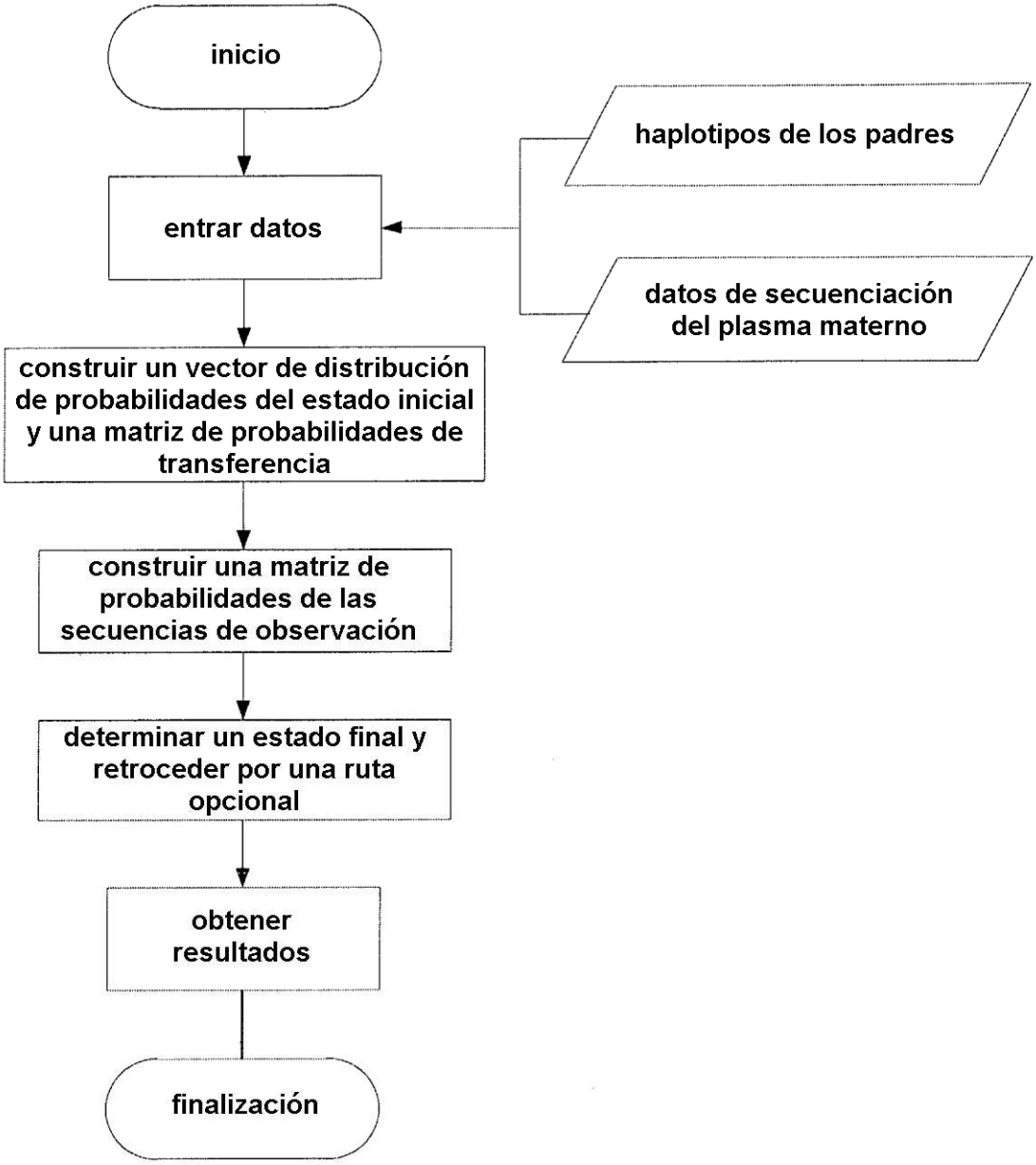
inicio
haplotipos de los padres
entrar datos
datos de secuenciación del plasma materno
construir un vector de distribución de probabilidades del estado inicial y una matriz de probabilidades de transferencia
construir una matriz de probabilidades de las secuencias de observación
determinar un estado final y retroceder por una ruta opcional
obtener resultados
finalización

Fig.1

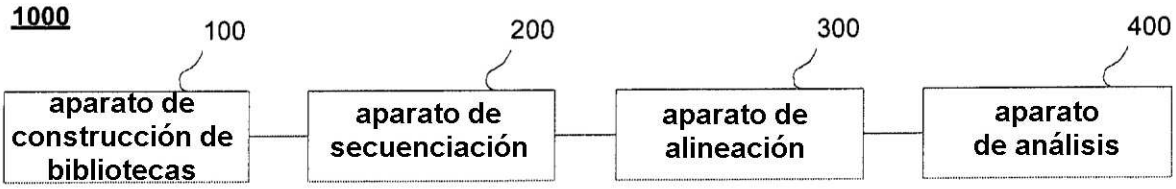
1000
100
200
300
400
aparato de construcción de bibliotecas
aparato de secuenciación
aparato de alineación
aparato de análisis

Fig.2