(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 572 912**

(51) Int. CI.:

C12Q 1/68 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **12.04.2012 E 12716939 (9)**

(97) Fecha y número de publicación de la concesión europea: **02.03.2016 EP 2697392**

(54) Título: **Resolución de fracciones de genoma mediante recuento de polimorfismos**

(30) Prioridad:

12.04.2011 US 201161474362 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
03.06.2016

(73) Titular/es:

VERINATA HEALTH, INC (100.0%)
800 Saginaw Drive
Redwood City CA 94063, US

(72) Inventor/es:

RAVA, RICHARD P.;
RHEES, BRIAN K. y
BURKE, JOHN P.

(74) Agente/Representante:

IZQUIERDO BLANCO, María Alicia

Resolución de fracciones de genoma mediante recuento de polimorfismos

Descripción

ANTECEDENTES

El descubrimiento de ADN fetal circulante (a veces denominado "ADN libre" o "cfADN") en la sangre materna permite la posibilidad de detectar anomalías cromosómicas, aneuploidía y aberraciones a partir de muestras de sangre. La abundancia fraccional de ADN fetal en el plasma sanguíneo materno no es constante y varía en función de diversos factores, incluidos la manipulación de la muestra y la edad gestacional.

Cuando se utiliza la secuenciación del ADN para identificar aberraciones cromosómicas o defectos genéticos, es importante conocer la abundancia relativa de ADN fetal en la población total de ADN. Por ejemplo, cuando se conoce la fracción fetal, la potencia estadística (la probabilidad de identificar casos anómalos, o la sensibilidad) puede calcularse mediante métodos de permutación o mediante integración de las combinaciones lineales o convoluciones de distribuciones F no centrales desde alfa hasta infinito, donde el punto crítico alfa para la significación (máxima probabilidad de asignar equivocadamente una anomalía) de la población de puntuaciones bajo la hipótesis nula de ninguna aberración.

En el documento US 7332277 se ilustra un método para detectar la presencia o ausencia de una anomalía cromosómica fetal cuantificando la relación de la cantidad relativa de alelos en un locus heterocigoto de interés.

Un inconveniente de los métodos existentes para la detección de la fracción fetal es que dependen de medidas de la abundancia de los cromosomas sexuales (que sólo puede utilizarse para medir de manera fiable la abundancia relativa de ADN embrionario de varón) o la secuencia de ARNm de genes conocidos que se expresan de manera diferencial entre el tejido de la embarazada y el embrionario (que está sujeto a la variabilidad de expresión debido a la edad gestacional u otros factores).

La estimación de la fracción fetal puede ser difícil debido a varios factores perturbadores, incluidos: parámetros de la genética de poblaciones diferenciales étnicos parentales y errores de secuenciación. Por lo tanto, es deseable disponer de métodos robustos en presencia de estos y otros factores de confusión que se producen comúnmente.

RESUMEN

La invención proporciona un método de estimación de la fracción de ADN fetal en el ADN obtenido a partir de un fluido corporal de una embarazada, comprendiendo el método:

(a) recibir una muestra del fluido corporal; (b) extraer ADN de la muestra en condiciones que extraen el ADN tanto de un genoma materno como de un genoma fetal presente en el fluido corporal; (c) secuenciar el ADN extraído con un secuenciador de ácidos nucleicos en condiciones que producen secuencias de segmentos de ADN que contienen uno o más polimorfismos; (d) mapear las secuencias de segmentos de ADN derivadas de la secuenciación del ADN en el fluido corporal contra uno o más polimorfismos designados en una secuencia de referencia, en el que el mapeo se realiza utilizando un aparato computacional programado para mapear secuencias de ácidos nucleicos contra el uno o más polimorfismos designados; (e) determinar las frecuencias alélicas de las secuencias de segmentos de ADN mapeadas para al menos uno de los polimorfismos designados; (f) clasificar el al menos un polimorfismo designado basándose en una combinación de la cigosidad de la embarazada y la cigosidad del feto; y (g) estimar la fracción de ADN fetal en el ADN obtenido de la embarazada utilizando las frecuencias alélicas determinadas en (e), junto con la clasificación de cigosidades de (f), en el que (e)-(g) se realizan en uno o más procesadores que ejecutan las instrucciones de un programa para la determinación, la clasificación y la estimación, en el que la secuenciación de la etapa (c) produce secuencias de segmentos de ADN que contienen una pluralidad de polimorfismos y las etapas (d)-(f) se realizan basándose en la pluralidad de polimorfismos; y en el que la clasificación en (f) clasifica cada una de la pluralidad de polimorfismos en una de las siguientes combinaciones: (i) la embarazada es homocigoto y el feto es homocigoto, (ii) la embarazada es homocigoto y el feto es heterocigoto, (iii) la embarazada es heterocigoto y el feto es homocigoto, y (iv) la embarazada es heterocigoto y el feto es heterocigoto.

Determinadas formas de realización descritas se refieren a métodos computacionales para medir de manera fiable la abundancia relativa de ADN fetal circulante mediante secuenciación de una muestra de sangre materna.

En formas de realización específicas, la invención proporciona métodos para estimar de manera fiable la fracción fetal a partir de polimorfismos tales como pequeñas variaciones de bases o inserciones-deleciones que son robustos con respecto a la etnicidad de los progenitores, el sexo del embrión, la edad gestacional y otros factores ambientales. Muchos ejemplos descritos en el presente documento emplean SNPs como polimorfismo pertinente. La invención puede aplicarse como parte de un estudio de resecuenciación prediseñado intencional dirigido contra

polimorfismos conocidos o puede utilizarse en un análisis retrospectivo de las variaciones encontradas por casualidad en secuencias solapantes generadas a partir de plasma materno (o cualquier otro entorno en el que haya una mezcla de ADN de varias personas).

En el presente documento se presentan técnicas para la estimación de la abundancia fraccional de ADN fetal en muestras de sangre materna. Determinadas técnicas descritas utilizan las frecuencias alélicas de SNPs observadas encontradas por casualidad o que se encuentran en paneles de SNPs previamente conocidos diseñados con el fin de estimar la fracción fetal.

Aunque gran parte de descripción tiene que ver con la estimación de la fracción de ácido nucleico fetal en una muestra, la descripción no se limita a ello. Las técnicas y los aparatos descritos en el presente documento pueden emplearse en muchos casos para estimar la fracción de ácido nucleico a partir de un genoma en una mezcla de dos genomas, que pueden estar relacionados, o no, como genomas de los progenitores y del niño/a.

Determinados aspectos de la descripción se refieren a métodos de estimación de la fracción de ADN fetal en el ADN obtenido a partir de un fluido corporal de una embarazada. Tales métodos pueden caracterizarse por las siguientes operaciones: (a) recibir una muestra del fluido corporal; (b) extraer ADN de la muestra en condiciones que extraen el ADN tanto de un genoma materno como de un genoma fetal presente en el fluido corporal; (c) secuenciar el ADN extraído con un secuenciador de ácidos nucleicos en condiciones que producen secuencias de segmentos de ADN que contienen uno o más polimorfismos; (d) mapear las secuencias de segmentos de ADN derivadas de la secuenciación del ADN en el fluido corporal contra uno o más polimorfismos designados en una secuencia de referencia; (e) determinar las frecuencias alélicas de las secuencias de segmentos de ADN mapeadas para al menos uno de los polimorfismos designados; (f) clasificar el al menos un polimorfismo designado basándose en una combinación de la cigosidad de la embarazada y la cigosidad del feto; y (g) estimar la fracción de ADN fetal en el ADN obtenido de la embarazada utilizando las frecuencias alélicas determinadas en (e) y la combinación de cigosidades de (f).

El mapeo puede realizarse utilizando un aparato computacional programado para mapear secuencias de ácidos nucleicos contra el uno o más polimorfismos designados. En general, cualquiera de las operaciones (d)-(g) puede realizarse en uno o más procesadores que ejecutan las instrucciones de un programa.

En determinadas formas de realización, el ADN obtenido a partir de un fluido corporal de una embarazada es el ADN libre obtenido del plasma de la embarazada. Por lo general, la secuenciación se lleva a cabo sin amplificar selectivamente ninguno del uno o más polimorfismos designados.

En determinadas formas de realización, el mapeo de los segmentos de ADN obtenidos a partir de la sangre de la portadora del feto comprende mapear computacionalmente los segmentos contra una base de datos de polimorfismos. En determinadas formas de realización, la clasificación en (f) clasifica el al menos un polimorfismo designado en una de las siguientes combinaciones: (i) la embarazada es homocigoto y el feto es homocigoto, (ii) la embarazada es homocigoto y el feto es heterocigoto, (iii) la embarazada es heterocigoto y el feto es homocigoto, y (iv) la embarazada es heterocigoto y el feto es heterocigoto.

Pueden emplearse diversas operaciones de filtrado. Estas incluyen, por ejemplo, no tener en cuenta ningún polimorfismo clasificado en la combinación (i) o en la combinación (iv). En otro ejemplo, los métodos incluyen adicionalmente filtrar el al menos un polimorfismo designado para no tener en cuenta ningún polimorfismo con una frecuencia del alelo minoritario superior a un umbral definido. En otro ejemplo, los métodos incluyen una operación de filtrado del al menos un polimorfismo designado para no tener en cuenta ningún polimorfismo con una frecuencia del alelo minoritario inferior a un umbral definido.

La operación de clasificación puede implementarse de diversas maneras. Por ejemplo, puede implicar la aplicación de un umbral a la frecuencia alélica determinada en (e). En otro ejemplo, la operación de clasificación implica la aplicación de los datos de frecuencia alélica de (e), obtenidos para una pluralidad de polimorfismos, a un modelo de mezcla. En una implementación, el modelo de mezcla emplea momentos factoriales.

La fracción fetal determinada como se describe en el presente documento puede utilizarse para diversas aplicaciones. En algunos ejemplos, los métodos descritos en el presente documento incluyen una operación de ejecución de instrucciones de programa en el uno o más procesadores que registren automáticamente la fracción de ADN fetal como se determina en (g) en un expediente clínico del paciente, almacenado en un medio legible por ordenador, para la embarazada. El expediente clínico del paciente puede mantenerse en un sitio web de expedientes clínicos personal, un laboratorio, consultorio médico, un hospital, una organización de mantenimiento de la salud o una compañía de seguros. En otra aplicación, la estimación de la fracción de ADN fetal se utiliza para prescribir, iniciar y/o modificar el tratamiento de una paciente humana de la que se obtuvo la muestra de ensayo materna. En otra aplicación, la estimación de la fracción de ADN fetal se utiliza para ordenar y/o realizar uno o más ensayos adicionales.

Otro aspecto de la descripción tiene que ver con un aparato para estimar la fracción de ADN fetal en el ADN obtenido a partir de un fluido corporal de una embarazada. Tal aparato puede caracterizarse por los siguientes elementos: (a) un secuenciador configurado para (i) recibir el ADN extraído de una muestra del fluido corporal que comprende ADN tanto de un genoma materno como de un genoma fetal, y (ii) secuenciar el ADN extraído en condiciones que producen secuencias de segmentos de ADN que contienen uno o más polimorfismos designados; y (b) un aparato computacional configurado para (por ejemplo, programado para) ordenar a uno o más procesadores que realicen diversas operaciones tales como las descritas, con dos o más de las operaciones del método descritas en el presente documento. En algunas formas de realización, el aparato computacional está configurado para (i) mapear secuencias de ácidos nucleicos contra el uno o más polimorfismos designados en una secuencia de referencia, (ii) determinar las frecuencias alélicas de las secuencias de segmentos de ADN mapeadas para al menos uno de los polimorfismos designados, (iii) clasificar el al menos un polimorfismo designado basándose en una combinación de la cigosidad de la embarazada y la cigosidad del feto, y (iv) estimar la fracción de ADN fetal en el ADN obtenido de la embarazada utilizando las frecuencias alélicas y la combinación de cigosidades.

En determinadas formas de realización, el aparato incluye también una herramienta para extraer ADN de la muestra en condiciones que extraen el ADN tanto del genoma materno como del genoma fetal. En algunas implementaciones, el aparato incluye un módulo configurado para extraer ADN libre obtenido a partir de plasma de la embarazada para la secuenciación en el secuenciador.

En algunos ejemplos, el aparato incluye una base de datos de polimorfismos. El aparato computacional puede estar configurado adicionalmente para que ordene al uno o más procesadores que mapeen los segmentos de ADN obtenidos a partir de la sangre de la portadora del feto mapeando computacionalmente los segmentos contra la base de datos de polimorfismos. Las secuencias en la base de datos son un ejemplo de secuencia de referencia. Más adelante se presentan otros ejemplos de secuencias de referencia.

En determinadas formas de realización, el aparato computacional está configurado adicionalmente para que ordene al uno o más procesadores que clasifiquen el al menos un polimorfismo designado en una de las siguientes combinaciones: (i) la embarazada es homocigoto y el feto es homocigoto, (ii) la embarazada es homocigoto y el feto es heterocigoto, (iii) la embarazada es heterocigoto y el feto es homocigoto, y (iv) la embarazada es heterocigoto y el feto es heterocigoto. En algunas formas de realización, el aparato computacional está configurado adicionalmente para que ordene al uno o más procesadores que no tengan en cuenta ningún polimorfismo clasificado en la combinación (i) o en la combinación (iv).

En determinadas formas de realización, el aparato computacional está configurado adicionalmente para que ordene al uno o más procesadores que no tengan en cuenta ningún polimorfismo con una frecuencia del alelo minoritario superior a un umbral definido. En algunas formas de realización, el aparato computacional está configurado adicionalmente para que ordene al uno o más procesadores que filtren el uno o más polimorfismos designados para que no tengan en cuenta ningún polimorfismo con una frecuencia del alelo minoritario inferior a un umbral definido. En determinadas formas de realización, el aparato computacional está configurado adicionalmente para que ordene al uno o más procesadores que clasifiquen el al menos un polimorfismo designado aplicando un umbral a la frecuencia alélica.

En determinadas formas de realización, el aparato computacional está configurado adicionalmente para que ordene al uno o más procesadores que clasifiquen el al menos un polimorfismo designado aplicando los datos de frecuencia alélica obtenidos para una pluralidad de polimorfismos, a un modelo de mezcla. El modelo de mezcla puede emplear momentos factoriales.

En determinadas formas de realización, el aparato computacional está configurado adicionalmente para que ordene al uno o más procesadores que registren automáticamente la fracción de ADN fetal en un expediente clínico del paciente, almacenado en un medio legible por ordenador, para la embarazada. El expediente clínico del paciente puede mantenerse en un sitio web de expedientes clínicos personal, un laboratorio, consultorio médico, un hospital, una organización de mantenimiento de la salud o una compañía de seguros.

Otro aspecto de la descripción tiene que ver con métodos de estimación de una fracción de ADN fetal en el ADN obtenido a partir de un fluido corporal de una embarazada según las siguientes operaciones: (a) mapear segmentos de ADN obtenido a partir del fluido corporal de la embarazada contra una pluralidad de secuencias de polimorfismos, en el que el ADN se secuenció en condiciones que identifican la pluralidad de secuencias de polimorfismos; (b) determinar una frecuencia alélica de los ácidos nucleicos mapeados para cada una de la pluralidad de secuencias de polimorfismos; y (c) aplicar las frecuencias alélicas a un modelo de mezcla para obtener una estimación de la fracción de ADN fetal en el ADN obtenido de la sangre de la portadora del feto. Puede realizarse una cualquiera o más de las operaciones (a)-(c) en uno o más procesadores que ejecutan las instrucciones de un programa. En determinadas formas de realización, la operación (c) implica ejecutar las instrucciones en el uno o más procesadores para resolver una serie de ecuaciones para los momentos factoriales de los datos de frecuencia alélica para cada una de la pluralidad de secuencias de polimorfismos. En algunas formas de realización, el modelo de mezcla tiene en cuenta el error de secuenciación.

En determinadas formas de realización, los métodos incluyen adicionalmente eliminar computacionalmente las frecuencias alélicas para los polimorfismos identificados como heterocigotos tanto en el feto como en la embarazada. En algunas implementaciones, antes de (c), los métodos incluyen una operación de eliminar computacionalmente las frecuencias alélicas para los polimorfismos identificados como homocigotos tanto en el feto como en la embarazada. En algunas implementaciones, antes de (c), los métodos incluyen una operación de eliminar computacionalmente las frecuencias alélicas para los polimorfismos identificados como heterocigotos en la embarazada.

El ADN obtenido a partir de un fluido corporal de una embarazada puede ser ADN libre obtenido del plasma de la embarazada. El mapeo de los ácidos nucleicos obtenidos a partir del fluido corporal puede implementarse mapeando los segmentos contra una base de datos de polimorfismos.

Los métodos de este aspecto de la descripción pueden incluir adicionalmente la secuenciación del ADN a partir del fluido corporal de una embarazada con un secuenciador de ácidos nucleicos en condiciones que producen secuencias de segmentos de ADN que contienen las secuencias de polimorfismos.

En algunas implementaciones, el mapeo en (a) comprende identificar una pluralidad de secuencias de polimorfismos bialélicos. En otras formas de realización, el mapeo en (a) comprende mapear los segmentos de ADN contra una pluralidad de secuencias de polimorfismos predefinidos.

En algunas formas de realización, los métodos de este aspecto incluyen adicionalmente ejecutar instrucciones de programa en el uno o más procesadores para que registren automáticamente la fracción de ADN fetal tal como se determina en (c) en un expediente clínico del paciente, almacenado en un medio legible por ordenador, para la embarazada. El expediente clínico del paciente puede mantenerse en un sitio web de expedientes clínicos personal, un laboratorio, consultorio médico, un hospital, una organización de mantenimiento de la salud o una compañía de seguros.

Basándose en la estimación de la fracción de ADN fetal, los métodos de este aspecto pueden incluir adicionalmente prescribir, iniciar y/o modificar el tratamiento de una paciente humana de la que se obtuvo la muestra de ensayo materna. Basándose en la estimación de la fracción de ADN fetal, los métodos de este aspecto pueden incluir adicionalmente ordenar y/o realizar uno o más ensayos adicionales.

Según aún otro aspecto de la descripción, se proporcionan métodos para estimar la fracción de ADN fetal en el ADN obtenido a partir de un fluido corporal de una embarazada mediante las siguientes operaciones: (a) recibir una muestra del fluido corporal; (b) extraer ADN de la muestra en condiciones que extraen el ADN tanto de un genoma materno como de un genoma fetal presente en el fluido corporal; (c) secuenciar el ADN extraído con un secuenciador de ácidos nucleicos en condiciones que producen secuencias de segmentos de ADN; (d) comparar las secuencias de segmentos de ADN derivadas del fluido corporal y, a partir de la comparación, identificar uno o más polimorfismos bialélicos; (e) determinar las frecuencias alélicas de las secuencias de segmentos de ADN para al menos uno de los polimorfismos identificados; (f) clasificar el al menos un polimorfismo identificado basándose en una combinación de la cigosidad de la embarazada y la cigosidad del feto; y (g) estimar la fracción de ADN fetal en el ADN obtenido de la embarazada utilizando las frecuencias alélicas determinadas en (e) y la combinación de cigosidades de (f).

El mapeo puede realizarse utilizando un aparato computacional programado para mapear secuencias de ácidos nucleicos contra el uno o más polimorfismos designados. En general, cualquiera de las operaciones (d)-(g) puede realizarse en uno o más procesadores que ejecutan las instrucciones de un programa.

En determinadas implementaciones de este aspecto, las secuencias de segmentos de ADN tienen una longitud de entre aproximadamente 20 pares de bases y aproximadamente 300 pares de bases.

En determinadas formas de realización de este aspecto, la clasificación en (f) clasifica el al menos un polimorfismo identificado en una de las siguientes combinaciones: (i) la embarazada es homocigoto y el feto es homocigoto, (ii) la embarazada es homocigoto y el feto es heterocigoto, (iii) la embarazada es heterocigoto y el feto es homocigoto, y (iv) la embarazada es heterocigoto y el feto es heterocigoto. Los métodos pueden incluir adicionalmente no tener en cuenta ningún polimorfismo clasificado en la combinación (i) o en la combinación (iv).

Según diversas formas de realización, los métodos de este aspecto pueden incluir el filtrado y/o las operaciones de clasificación que se describen en el presente documento en relación a otros aspectos. Por ejemplo, los métodos de este aspecto pueden incluir el filtrado del uno o más polimorfismos identificados para no tener en cuenta ningún polimorfismo con una frecuencia del alelo minoritario superior a un umbral definido. En algunos casos, la clasificación del al menos un polimorfismo identificado incluye aplicar un umbral a la frecuencia alélica determinada en (e). Tal como se describe en el presente documento, puede emplearse el uso de modelos de mezcla para clasificar los polimorfismos identificados.

Otro aspecto de la descripción tiene que ver con un aparato para estimar una fracción de ADN fetal y que incluye los siguientes elementos: (a) un secuenciador configurado para (i) recibir el ADN extraído de una muestra del fluido corporal que comprende ADN tanto de un genoma materno como de un genoma fetal, y (ii) secuenciar el ADN extraído para producir segmentos de secuencias de ADN; y (b) un aparato computacional configurado para ordenar a uno o más procesadores que (i) mapeen los segmentos de secuencias del ADN obtenido a partir del fluido corporal de la embarazada contra una pluralidad de secuencias de polimorfismos, (ii) determinen una frecuencia alélica para cada una de la pluralidad de secuencias de polimorfismos de los segmentos de secuencias de ADN mapeados, y (iii) apliquen las frecuencias alélicas a un modelo de mezcla para obtener una estimación de la fracción de ADN fetal en el ADN obtenido de la sangre de la portadora del feto.

Otro aparato para estimar la fracción de ADN fetal incluye los siguientes elementos: (a) un secuenciador configurado para (i) recibir el ADN extraído de una muestra del fluido corporal que comprende ADN tanto de un genoma materno como de un genoma fetal, y (ii) secuenciar el ADN extraído en condiciones que producen secuencias de segmentos de ADN; y (b) un aparato computacional configurado para ordenar a uno o más procesadores que (i) comparen las secuencias de segmentos de ADN derivadas del fluido corporal y, a partir de la comparación, identifiquen uno o más polimorfismos bialélicos, (ii) determinen las frecuencias alélicas de las secuencias de segmentos de ADN para al menos uno de los polimorfismos identificados, (iii) clasifiquen el al menos un polimorfismo identificado basándose en una combinación de la cigosidad de la embarazada y la cigosidad del feto, y (iii) estimen la fracción de ADN fetal en el ADN obtenido de la embarazada utilizando las frecuencias alélicas y la combinación de cigosidades.

Las instrucciones y/o el hardware empleados en los aspectos del aparato descritos en el presente documento pueden proporcionar la ejecución de una cualquiera o más de las operaciones algorítmicas o computacionales de los aspectos del método descritos en el presente documento, independientemente de si tales operaciones se han enumerado explícitamente anteriormente.

Estas y otras características y ventajas de las formas de realización descritas se describirán con más detalle más adelante con referencia a los dibujos asociados.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

La Figura 1 es un diagrama de bloques que representa la clasificación de los estados de cigosidad fetal y materna para una determinada posición genómica.

La Figura 2 es un flujo de procesos de ejemplo para implementar algunas de las formas de realización descritas.

La Figura 3 presenta las estimaciones de error mediante la posición de las bases secuenciadas sobre 30 calles de datos de Illumina GA2 alineados con el genoma humano HG18 utilizando Eland con los parámetros por defecto.

La figura 4 es un gráfico del recuento del alelo minoritario A frente a la cobertura D (suponiendo que no hay error) para los casos de heterocigosidad 1 a 4.

La figura 5 representa la transformación de los datos del caso 3 en caso 2.

La Figura 6 presenta los datos después de la rotación, en la que D1 se seleccionó de manera que el caso 1 y los casos 2, 3 no se solapasen. E1 representa un límite superior del intervalo de confianza superior del 99 por ciento de los datos del caso 1.

La Figura 7 muestra una comparación de los resultados utilizando un modelo de mezcla y la fracción fetal conocida y la fracción fetal estimada.

La Figura 8 muestra que el uso de la tasa de error de la máquina como parámetro conocido reduce en un punto el sesgo por exceso.

En la Figura 9 se muestra que los datos simulados utilizando la tasa de error de la máquina como parámetro conocido, que mejora los modelos de error del caso 1 y 2, reduce en gran medida el sesgo por exceso a menos de un punto para la fracción fetal por debajo de 0,2.

La Figura 10 es una representación esquemática de un sistema informático que, cuando está debidamente configurado (por ejemplo, programado) o diseñado, puede servir de aparato de análisis para las formas de realización descritas.

Las Figuras 11A y B muestran un histograma del número de observaciones de variantes (frecuencia) en el porcentaje del alelo minoritario (A/D) para el cromosoma cromosomas 1(A) y el cromosoma 7 tal como se produce en un ejemplo.

Las Figuras 12A y B muestran la distribución de la frecuencia alélica en los cromosomas 1 (A) y el cromosoma 7.

DESCRIPCIÓN DETALLADA

Introducción y visión de conjunto

Determinadas formas de realización descritas implican analizar el ADN obtenido de la sangre de una mujer embarazada y utilizar el análisis para estimar la fracción de ese ADN que proviene del feto. A continuación, puede utilizarse la fracción de ADN fetal para atribuir un cierto nivel de confianza a otra métrica o caracterización del feto basándose en el análisis independiente del ADN obtenido de la sangre de la madre. Por ejemplo, puede analizarse por separado una muestra de ADN fetal obtenida de la sangre materna para detectar aneuploidía en el feto que lleva mujer embarazada. La determinación de aneuploidía hecha mediante este análisis por separado puede ser proporcionada por un nivel de confianza fundamentado estadísticamente basándose en la cantidad fraccional de ADN fetal presente en el ADN obtenido de la sangre de la madre. Fracciones de ADN fetal relativamente bajas en el complemento total de ADN sugieren una baja confianza en cualquier caracterización basada en el ADN fetal.

Por lo general, aunque no necesariamente, el ADN analizado en la sangre de la madre es ADN libre, aunque en algunas formas de realización, puede ser ADN celular. El ADN libre se obtiene a partir del plasma de la madre. La cantidad de ADN fetal en el contenido de ADN libre obtenido de mujeres embarazadas varía mucho dependiendo de diversos factores, incluida la edad gestacional del feto. Para las mujeres humanas embarazadas típicas, actualmente se cree que aproximadamente el 5%-20% del ADN libre es ADN fetal. Sin embargo, no es infrecuente que la fracción fetal sea significativamente inferior (por ejemplo, aproximadamente un 1% o inferior). En tales casos, cualquier caracterización separada del ADN fetal puede ser intrínsecamente sospechosa. Por otro lado, algunos investigadores han informado acerca de muestras de ADN libre maternas con fracciones de ADN fetal tan altas como un 40% o un 50%.

En determinadas implementaciones descritas en el presente documento, la determinación de la fracción fetal de ADN materno se basa en múltiples lecturas de secuencias de ADN en los sitios de secuencia que se sabe albergan uno o más polimorfismos. Por lo general, aunque no necesariamente, tales polimorfismos son polimorfismos de un sólo nucleótido (SNP). Otros tipos de polimorfismos adecuados incluyen deleciones, STRs (repeticiones cortas en tándem), inserciones, indels (incluidos microindels), etc. Más adelante se presentan ejemplos adicionales. En determinadas formas de realización, los sitios de polimorfismo se encuentran en una "secuencia de referencia" como se describe más adelante. En algunas formas de realización, los sitios de polimorfismo se descubren mientras se alinean marcadores de secuencia entre sí y/o con una secuencia de referencia.

Determinados métodos descritos se valen del hecho de que las secuencias de ADN de un feto en los sitios de polimorfismo en cuestión pueden no corresponder a los de su madre. Por ejemplo, el ADN de la madre en el sitio de un SNP particular puede ser homocigoto, mientras que la versión del feto del SNP será heterocigoto. Por lo tanto, un grupo de muestras de secuencias obtenidas para el SNP en cuestión será heterogéneo con la mayoría de las secuencias que contienen el alelo mayoritario y la fracción restante que contiene el alelo minoritario. Las cantidades relativas de los alelos mayoritario y minoritario están determinadas por la fracción de ADN fetal en la muestra.

Cabe mencionar que en una muestra homocigoto ambas copias de un determinado SNP u otro polimorfismo contienen el mismo alelo, mientras que un SNP heterocigoto u otro polimorfismo contiene una copia del alelo mayoritario y una copia del alelo minoritario. Por lo tanto, se sabe que el ADN obtenido exclusivamente a partir de un individuo heterocigoto debe contener un 50% del alelo mayoritario y un 50% del alelo minoritario. Este conocimiento puede utilizarse para dilucidar la fracción de ADN fetal como se resume más adelante. Como se explica más detalladamente más adelante, los diversos métodos descritos en el presente documento consideran sólo los polimorfismos en los que sólo hay dos alelos en el ADN materno y fetal, de manera colectiva.

En algunas implementaciones, el ADN obtenido de sangre de la madre se lee muchas veces, considerándose el número total de lecturas que se mapean contra un sitio particular de un polimorfismo la "cobertura" del polimorfismo, y considerándose el número de lecturas que se mapean contra el alelo minoritario para ese polimorfismo el recuento del alelo minoritario. La relación entre el recuento del alelo minoritario y la cobertura es importante en varias implementaciones.

Algunos de los métodos descritos en el presente documento identifican y caracterizan cuatro casos de polimorfismos en las muestras de ADN que comprenden ADN tanto de la madre como del feto. La Figura 1 que se presenta más adelante representa estos cuatro casos. En concreto, en un primer caso, que es bastante poco interesante, tanto la madre como el feto son homocigotos en el polimorfismo particular que se está considerando. En tal caso, cada secuencia en la muestra de ADN que contiene el polimorfismo en cuestión contendrá el mismo alelo y no puede recogerse información sobre las cantidades relativas de ADN de la madre y del feto. Sin embargo, cabe señalar que este caso podría ser interesante en el sentido de que permite que el investigador o el técnico se haga una idea de la tasa de error relativo del aparato de secuenciación de ADN utilizado para generar los datos de secuencia en cuestión.

El segundo caso que encontrará el análisis es un polimorfismo para el que la mujer embarazada es homocigoto y el feto es heterocigoto. En este caso, una fracción relativamente pequeña, pero significativa, de las secuencias detectadas contendrá el alelo minoritario. En concreto, en este segundo caso, la frecuencia del alelo minoritario viene dada concretamente por la fracción de ADN fetal en el torrente sanguíneo de la madre dividido por dos.

En un tercer caso, el polimorfismo en cuestión es heterocigoto en el ADN de la madre y homocigoto en el ADN del feto. En esta situación, la frecuencia del alelo minoritario viene dada concretamente por 0,5 menos la mitad de la fracción de ADN fetal en la muestra de ADN.

Por último, en el cuarto caso, el polimorfismo en cuestión es heterocigoto tanto en la madre como en el feto. En este caso, se espera que la frecuencia de los alelos mayoritario y minoritario sea 0,5 para ambos. Al igual que en el primer caso, el cuarto caso es relativamente poco interesante para determinar la fracción de ADN fetal.

Si el investigador, el técnico o el software encargado de determinar la fracción de ADN fetal en una muestra supiera para un determinado polimorfismo a cuál de los cuatro casos pertenece ese polimorfismo, podría entonces estimarse directamente la fracción de ADN fetal, suponiendo que el polimorfismo en cuestión se encontrase dentro de cualquiera de los casos dos o tres. Sin embargo, en la práctica, nunca se tiene este conocimiento *a priori*. Por lo tanto, resulta necesario un aparato computacional para realizar las operaciones descritas en el presente documento.

En determinadas formas de realización, descritas en otra parte del presente documento, se emplea una técnica de umbralización para clasificar un solo polimorfismo en uno de los cuatro casos. Una vez así clasificado el polimorfismo, y descubierto que se encuentra en cualquiera de los casos 2 ó 3, puede estimarse la fracción fetal. En otras formas de realización, la técnica considera polimorfismos múltiples distribuidos a través de todo el genoma o una parte del mismo. Como se ilustra en los ejemplos específicos, pueden utilizarse múltiples SNPs diferentes en todo el genoma con este fin.

En formas de realización concretas, se determina la frecuencia alélica para varios polimorfismos diferentes en una muestra de ADN obtenida de la muestra de sangre de la madre. Para esta pluralidad de polimorfismos, una fracción corresponderá al caso de cigosidad 1, otra fracción corresponderá al caso 2, una tercera fracción corresponderá al caso 3, y una fracción final corresponderá al caso 4. La suma de estas fracciones dará un valor de 1. Puede emplearse un modelo de mezcla o una técnica relacionada para evidenciar una o más propiedades estadísticas de los polimorfismos en cada una de estas cuatro categorías. En concreto, puede emplearse un modelo de mezcla para determinar una media y, opcionalmente, la varianza para cada uno de los cuatro casos encontrados en una muestra de ADN obtenida de la sangre de una mujer embarazada. En formas de realización específicas, esta es la media y la varianza asociada con la frecuencia del alelo minoritario con relación al número total de recuentos para un polimorfismo en cuestión (cobertura). Como se desarrolla en otra parte del presente documento, los valores medios para cada una de estas cuatro categorías, o al menos las categorías segunda y tercera, están directamente relacionados con la fracción fetal en el ADN obtenido de la sangre de la madre.

En una implementación que emplea modelos de mezcla, se calculan uno o más momentos factoriales para cada posición en la que se está considerando un polimorfismo. Por ejemplo, se calcula un momento factorial (o un grupo de momentos factoriales) utilizando múltiples posiciones de SNP consideradas en la secuencia de ADN. Como se muestra en la ecuación 4 que se presenta más adelante, cada uno de los diversos momentos factoriales es un sumatorio sobre todas las diferentes posiciones de SNP en cuestión para la relación entre la frecuencia del alelo minoritario y la cobertura de una determinada posición. Como se muestra en la ecuación 5 que se presenta más adelante, estos momentos factoriales también están relacionados con los parámetros asociados con cada uno de los cuatro casos de cigosidad descritos anteriormente. En concreto, se relacionan con la probabilidad para cada uno de los casos, así como las cantidades relativas de cada uno de los cuatro casos en el grupo de polimorfismos en cuestión. Como se ha explicado, la probabilidad está en función de la fracción de ADN fetal en el ADN libre en la sangre de la madre. Como se explica más detalladamente más adelante, mediante el cálculo de un número suficiente de estos momentos factoriales (que se muestran en la ecuación 4), el método proporciona un número suficiente de expresiones para calcular el valor de todas las incógnitas. Las incógnitas en este caso serían las cantidades relativas de cada uno de los cuatro casos en la población de polimorfismos en cuestión, así como las probabilidades (y por tanto las fracciones de ADN fetal) asociadas con cada uno de estos cuatro casos. Véase la ecuación 5. Pueden obtenerse resultados similares utilizando otras versiones de modelos de mezcla tal como se representa en las ecuaciones 7-12 que se presentan más adelante. Estas versiones particulares se valen sólo de polimorfismos que se encuentran en los casos 1 y 2, filtrándose los polimorfismos de los casos 3 y 4 mediante una técnica de umbralización.

Por lo tanto, los momentos factoriales pueden utilizarse como parte de un modelo de mezcla para identificar las probabilidades de cualquier combinación de los cuatro casos de cigosidad. Y, como se ha mencionado, estas probabilidades, o al menos aquellas para los casos segundo y tercero, están directamente relacionadas con la fracción de ADN fetal en el ADN libre total en la sangre de la madre.

Debe mencionarse también que puede emplearse el error de secuenciación para reducir la complejidad del sistema de ecuaciones de los momentos factoriales que deben resolverse. En este sentido, debe reconocerse que el error de secuenciación puede tener en realidad uno cualquiera de entre cuatro resultados (que corresponden a cada una de las cuatro bases posibles en cualquier posición de polimorfismo determinada).

En determinadas formas de realización, los marcadores se alinean con un cromosoma o genoma de referencia, y se identifican los polimorfismos bialélicos. Estos polimorfismos no están predefinidos ni identificados de otra manera antes del alineamiento. Se identifican simplemente durante el alineamiento y, a continuación, se caracterizan basándose en sus cigosidades y recuentos de alelos minoritarios, como se describe en el presente documento. Esta información se utiliza para estimar las fracciones genómicas tal como se describe en el presente documento.

Las longitudes de los marcadores utilizadas en las formas de realización descritas en el presente documento se determinarán generalmente mediante el método de secuenciación empleado para generar los marcadores. Los métodos son robustos en un amplio intervalo de longitudes de marcador. En determinadas implementaciones, los marcadores tienen una longitud de aproximadamente 20 a 300 pares de bases (o una longitud de aproximadamente 30 a 100 pares de bases).

En la Figura 2 se muestra un flujo de procesos de ejemplo para implementar algunas de las formas de realización descritas. Como ahí se representa, el proceso comienza en 201 con la recogida de ADN (libre o celular) a partir de sangre materna u otro fluido corporal. De este ADN se mapean múltiples secuencias contra uno o más polimorfismos en una secuencia de referencia. Este mapeo proporciona una frecuencia alélica para cada uno de los polimorfismos. Véase el bloque 203.

Más concretamente, el proceso del bloque 203 puede implicar la lectura de las secuencias del ADN recogido en ubicaciones de polimorfismos múltiples. En algunos casos, éstos pueden generarse como parte del proceso para determinaciones de ploidía u otra determinación hecha con respecto al ADN fetal. Por lo tanto, en algunas formas de realización, no es necesario generar secuencias separadas. Las secuencias leídas se alinean con una secuencia de referencia para maximizar el alineamiento utilizando BLAST o una herramienta similar.

La secuencia de referencia puede proporcionarse como una base de datos de polimorfismos. En algunos casos, se trata de un conjunto de referencia de búsqueda de alelos producido a partir de una expansión combinatoria de todas las definiciones de polimorfismo (por ejemplo, cuando los polimorfismos son SNPs, todas las secuencias SNP). Véase el Anexo, por ejemplo. En un ejemplo específico, las secuencias tienen una longitud de aproximadamente 100 a 150 pares de bases.

Volviendo a la Figura 2, el método determina la combinación de cigosidad materna/fetal para uno o más de los polimorfismos considerados en la operación del bloque 203. Véase el bloque 205. En determinadas formas de realización puede emplearse un modelo de mezcla con este fin. Como se ha mencionado, las combinaciones son del siguiente modo: M y F homocigotos, M homocigoto y F heterocigoto, M heterocigoto y F homocigoto, y M y F heterocigotos.

Por último, como se ilustra en el bloque 207, el método utiliza la combinación de frecuencia alélica del caso de cigosidad en uno o más de los polimorfismos para estimar la cantidad fraccional de componente fetal en el ADN de la muestra materna.

Definiciones

El siguiente análisis se proporciona como ayuda para comprender determinados aspectos y ventajas de las formas de realización descritas.

El término "lectura" se refiere a una lectura de secuencia de una porción de una muestra de ácido nucleico. Por lo general, aunque no necesariamente, una lectura representa una secuencia corta de pares de bases contiguas de la muestra. La lectura puede representarse simbólicamente mediante la secuencia de pares de bases (en ATCG) de la porción de la muestra. Puede almacenarse en un dispositivo de memoria y procesarse según corresponda para determinar si coincide con una secuencia de referencia o cumple con otros criterios. Una lectura puede obtenerse directamente de un aparato de secuenciación o indirectamente de información almacenada de la secuencia acerca de la muestra.

El término "marcador" también se refiere a secuencias cortas de una muestra de ácido nucleico. Por lo general, un marcador contiene información asociada tal como la ubicación de la secuencia en el genoma. Para algunos fines, los términos “lectura” y “marcador” son intercambiables en el presente documento. Sin embargo, por lo general, las lecturas de secuencias se alinean con una secuencia de referencia, y las lecturas que se mapean en un solo sitio en el genoma de referencia se denominan marcadores. En el presente documento, a veces se utiliza "secuencia de segmento" de manera intercambiable con "marcador".

Con frecuencia en el presente documento las "lecturas" se describen como secuencias de ácidos nucleicos que tienen una longitud de 36 pares de bases (36-meros). Por supuesto, las formas de realización descritas no se limitan a este tamaño. En muchas aplicaciones resultan adecuadas lecturas menores y mayores. Para las aplicaciones que alinean lecturas con el genoma humano, se considera generalmente suficiente una lectura de un tamaño de 30 pares de bases o mayor para mapear una muestra contra un solo cromosoma. Para algunas aplicaciones resultan adecuados marcadores/lecturas mucho mayores. Con la secuenciación del genoma completo, pueden utilizarse lecturas del orden de 1.000 pares de bases o mayores. En determinadas formas de realización, una lectura puede tener una longitud de entre aproximadamente 20 y 10.000 pares de bases, o entre aproximadamente 30 y 1.000 pares de bases, o entre aproximadamente 30 y 50 pares de bases.

Una "secuencia de referencia" es una secuencia de una molécula biológica, que con frecuencia es un ácido nucleico tal como un cromosoma o genoma. Por lo general, múltiples lecturas son miembros de una determinada secuencia de referencia. En determinadas formas de realización, se compara una lectura o marcador con una secuencia de referencia para determinar si la secuencia de referencia contiene la secuencia leída. Este proceso se denomina a veces alineamiento.

En diversas formas de realización, la secuencia de referencia es significativamente mayor que las lecturas que se alinean con la misma. Por ejemplo, puede ser al menos aproximadamente 100 veces mayor, o al menos aproximadamente 1.000 veces mayor, o al menos aproximadamente 10.000 veces mayor, o al menos aproximadamente 10^5 veces mayor, o al menos aproximadamente 10^6 veces mayor, o al menos aproximadamente 10^7 veces mayor.

En un ejemplo, la secuencia de referencia es la de un genoma humano de longitud completa. Tales secuencias pueden denominarse secuencias de referencia genómicas. En otro ejemplo, la secuencia de referencia se limita a un cromosoma humano específico, tal como el cromosoma 13. Tales secuencias pueden denominarse secuencias de referencia cromosómicas. Otros ejemplos de secuencias de referencia incluyen genomas de otras especies, así como cromosomas, regiones subcromosómicas (por ejemplo, cadenas), etc., de cualquier especie.

En diversas formas de realización, la secuencia de referencia es una secuencia consenso u otra combinación procedente de múltiples individuos. Sin embargo, en determinadas aplicaciones, la secuencia de referencia puede obtenerse de un individuo concreto.

El término "alineamiento" se refiere al proceso de comparar una lectura o marcador con una secuencia de referencia y determinar de este modo si la secuencia de referencia contiene la secuencia leída. Si la secuencia de referencia contiene la lectura, la lectura puede mapearse contra la secuencia de referencia o, en determinadas formas de realización, contra una ubicación particular de la secuencia de referencia. En algunos casos, el alineamiento simplemente indica si una lectura es miembro, o no, de una secuencia de referencia particular (es decir, si la lectura está presente o ausente en la secuencia de referencia). Por ejemplo, el alineamiento de una lectura con la secuencia de referencia para el cromosoma 13 humano indicará si la lectura está presente en la secuencia de referencia para el cromosoma 13. Una herramienta que proporciona esta información puede denominarse verificador de pertenencia al conjunto. En algunos casos, un alineamiento indica además una ubicación en la secuencia de referencia contra la que se mapea la lectura o el marcador. Por ejemplo, si la secuencia de referencia es la secuencia del genoma humano completo, un alineamiento puede indicar que hay una lectura en el cromosoma 13, y puede indicar adicionalmente que la lectura está en una cadena concreta del cromosoma 13.

Un "sitio" es una posición única en una secuencia de referencia correspondiente a una lectura o marcador. En determinadas formas de realización, éste especifica la identidad de un cromosoma (por ejemplo, el cromosoma 13), una cadena del cromosoma, y una posición exacta en el cromosoma.

Un "sitio polimórfico" es un locus en el que se produce una divergencia de la secuencia nucleotídica. El locus puede tener tan solo un par de bases. Los marcadores ilustrativos tienen al menos dos alelos, dándose cada uno con una frecuencia superior al 1%, y, más generalmente, superior al 10% ó 20% de una población seleccionada. Un sitio polimórfico puede tener tan solo un par de bases. Las expresiones "locus polimórfico" y "sitio polimórfico" se utilizan indistintamente en el presente documento.

En el presente documento, "secuencia polimórfica" se refiere a una secuencia de ácido nucleico, por ejemplo, una secuencia de ADN, que comprende uno o más sitios polimórficos, por ejemplo un SNP o un SNP en tándem. Las secuencias polimórficas según la presente tecnología pueden utilizarse para diferenciar específicamente entre los alelos maternos y no maternos en la muestra materna que comprende una mezcla de ácidos nucleicos fetales y maternos.

Formas de realización detalladas

Por lo general, los procesos descritos en el presente documento emplean una secuencia de referencia que abarca uno o más polimorfismos y está asociada con el ADN que se está muestreando. Una secuencia de referencia puede ser, por ejemplo, el genoma humano, un cromosoma, o una región de un cromosoma. Pueden designarse

uno o más de los polimorfismos con el fin de estimar la fracción de ADN fetal. Los polimorfismos que se designan para su uso en la determinación de la fracción fetal son polimorfismos previamente conocidos. Por ejemplo, se ha recopilado un listado completo de referencias, hechos e información de secuencias sobre STRs previamente conocidas, y de datos de población relacionados en la STRBase, al que puede accederse a través de la web en ibm4.carb.nist.gov:8800/dna/home.htm. También puede accederse a la información de secuencias del GenBank® (http://www2.ncbi.nlm.nih.gov/cgibin/genbank) para los loci STR comúnmente utilizados a través de la STRBase. La información de SNPs previamente conocidos está disponible en bases de datos de acceso público, incluidas pero no limitadas a Human SNP Database en la dirección web wi.mit.edu, la página inicial de dbSNP del NCBI en la dirección web ncbi.nlm.nih.gov, la dirección web lifesciences.perkinelmer.com, Applied Biosystems by Life Technologies™ (Carlsbad, CA) en la dirección web appliedbiosystems.com, la base de datos de SNP de Celera Human en la dirección web celera.com, la base de datos de SNP del Genome Analysis Group (GAN) en la dirección web gan.iarc.fr. En una forma de realización, los SNPs designados para determinar la fracción fetal se seleccionan del grupo de 92 SNPs de identificación individuales (IISNPs) descrito por el Pakstis *et al*. (Pakstis *et al*. Hum. Genet. 127:315-324 [2010]), que han demostrado tener una variación de frecuencia muy pequeña entre las poblaciones (F_{st} < 0,06), y ser muy reveladores en todo el mundo con una heterocigosidad media ≥ 0,4. Los SNPs abarcados por el método de la invención incluyen SNPs unidos y no unidos. Para designar las secuencias SNP en tándem adecuadas, pueden buscarse en la base de datos del International HapMap Consortium (The International HapMap Project, Nature 426:789-796 [2003]). La base de datos está disponible en la web en hapmap.org.

Los polimorfismos así empleados pueden ser paneles de polimorfismos previamente conocidos designados para determinar la fracción de ADN fetal o pueden encontrarse por casualidad en un análisis de ADN materno para otros fines, tal como el mapeo de marcadores de ADN de la muestra contra los cromosomas.

En determinadas formas de realización, el método comprende secuenciar el ADN en una muestra utilizando una mezcla de genomas, por ejemplo, una muestra materna que comprende ADN libre fetal y materno, para proporcionar una pluralidad de marcadores de secuencia que se mapean contra secuencias que comprenden sitios polimórficos previamente conocidos en un genoma de referencia, y utilizar los marcadores mapeados en los sitios previamente conocidos para determinar la fracción fetal como se describe detalladamente más adelante. Como alternativa, después de la secuenciación del ADN, los marcadores de secuencia que se obtienen mediante la tecnología de secuenciación, por ejemplo, NGS, se mapean contra un genoma de referencia, por ejemplo, hg19, y los marcadores de secuencia que se mapean contra los sitios en los que los polimorfismos se producen por casualidad, es decir, no conocidos previamente, se utilizan para determinar la fracción fetal.

La secuencia de referencia contra la que se mapean los marcadores de secuencia a sitios polimórficos previamente conocidos, puede ser un genoma de referencia publicado o puede ser una base de datos artificial u otro grupo predefinido de secuencias para los polimorfismos en cuestión. Cada una de las secuencias de la base de datos abarcará el uno o más nucleótidos asociados con el polimorfismo. Como ejemplo, véase la lista de secuencias de polimorfismos que se presenta más adelante en el "Anexo 1".

En diversas formas de realización, el número de polimorfismos empleados para estimar la fracción de ADN fetal es de al menos 2 polimorfismos, y más particularmente para cada uno de al menos aproximadamente 10 polimorfismos, y más preferentemente para cada uno de al menos aproximadamente 100 polimorfismos.

En un ejemplo, la cobertura de SNP y la frecuencia alélica se determinan alineando las secuencias generadas con un genoma de referencia construido a partir de la expansión combinatoria de las definiciones de SNP. La base de datos de amplicones contiene información de variación bialélica rodeada, por ejemplo, por al menos aproximadamente 50 bases de secuencia flanqueante. Por ejemplo, un amplicón con una cadena de información de variación "[g/c]" (que representa los alelos alternos "g" y "c") puede parecerse a:

atcg.....accg[g/c]ccgt....

En algunos casos, el procedimiento para introducir la base de datos de amplicones y las secuencias generadas y devolver los recuentos de SNP/alelo es del siguiente modo.

1. Crear un conjunto de referencia de búsqueda de alelos a partir de la expansión combinatoria de las definiciones de SNP. Para cada secuencia en la base de datos de amplicones, para cada alelo en la cadena de información de variación, crear una secuencia alélica, sustituyéndose la cadena de información de variación por el alelo.

a. Por ejemplo, teniendo en cuenta la secuencia de amplicón del ejemplo anterior, se crearían dos secuencias: 1) atcg.....accgGccgt.... y 2) atcg.....accgCccgt....

b. Puede encontrarse un ejemplo de un conjunto de referencia de búsqueda de alelos completo en el Allele Search Database Sequence Listing.

2. Mapear las secuencias contra el conjunto de referencia de búsqueda de alelos manteniendo sólo los mapeos que coincidan con sólo una secuencia en el conjunto de búsqueda.

3. **El recuento de alelos** se determina contando el número de secuencias que coinciden con su secuencia alélica.

Los métodos descritos en el presente documento suponen un embarazo "normal", es decir, un embarazo en el que la madre lleva un solo feto, y no gemelos, trillizos, etc. Los expertos comprenderán las modificaciones que tienen en cuenta embarazos no normales, particularmente aquellos en los que se conoce el número de fetos.

Como se ha indicado, al determinar la fracción fetal, el método secuencia el ADN en la muestra de sangre materna y realiza el recuento de los marcadores de secuencia que se mapean contra cada secuencia de polimorfismo(s) en cuestión. Para cada polimorfismo, el método hace recuento del número total de lecturas que se mapean contra el mismo (la cobertura) y el número de marcadores de secuencia asociados a cada alelo (los recuentos de alelos). En un ejemplo sencillo, un polimorfismo con una cobertura de 5, pueden tener 3 lecturas del alelo B y 2 lecturas del alelo A. En este ejemplo, el alelo A se considera el alelo minoritario y el alelo B se considera el alelo mayoritario.

En algunas formas de realización, esta operación se vale de herramientas de secuenciación muy rápidas tales como las herramientas de secuenciación masiva en paralelo de ADN. Más adelante se describen con más detalle ejemplos de tales herramientas. En algunos casos, se leen para una sola muestra muchos miles o millones de secuencias marcadoras. Preferentemente, la secuenciación se realiza de manera que permita una asignación rápida y directa del ADN secuenciado a secuencias predefinidas particulares que albergan los polimorfismos en cuestión. En general, hay información suficiente para ello en los marcadores con un tamaño de 30 pares de bases o mayores. Los marcadores de este tamaño pueden mapearse de modo inequívoco contra las secuencias de interés. En una forma de realización específica, las secuencias marcadoras empleadas en el proceso tienen una longitud de 36 pares de bases.

Los marcadores se mapean contra un genoma de referencia o contra las secuencias de una base de datos de secuencias alélicas (por ejemplo, véase el Anexo 1 como se ha mencionado anteriormente) y se determina el número de marcadores así mapeados. Esto proporcionará tanto la cobertura como el recuento del alelo minoritario para cada polimorfismo en cuestión. En algunos casos, esto puede hacerse simultáneamente al mapeo de cada marcador contra uno de los 23 cromosomas humanos y la determinación del número de marcadores mapeados por cromosoma.

Como se ha mencionado, la cobertura es el número total de secuencias leídas que se mapean contra un determinado polimorfismo en una secuencia de referencia. El recuento de alelos es el número total de secuencias leídas que se mapean contra tal polimorfismo que tiene un alelo. La suma de todos los recuentos de alelos debe ser igual a la cobertura. El alelo con el recuento más alto es el alelo mayoritario, y el alelo con el recuento más bajo es el alelo minoritario. En determinadas formas de realización, la única información necesaria para estimar la fracción de ADN fetal es la cobertura y el recuento del alelo minoritario para cada uno de una pluralidad de polimorfismos. En algunas formas de realización, también se utiliza una tasa de error de asignación de bases del aparato de secuenciación del ADN.

Resulta útil tener en cuenta los fundamentos matemáticos o simbólicos de determinados métodos descritos en el presente documento. Como se ha mencionado, en diversos ejemplos, las secuencias generadas a partir de la sangre materna se alinean (se superponen de manera que se maximicen las bases idénticas) con un genoma de referencia u otra secuencia de ácido nucleico. Dada una posición genómica, j, y un conjunto de secuencias alineadas con la referencia, déjese que el número de apariciones de cada una de las cuatro bases del ADN ("a", "t", "g" y "c", también denominadas "alelos"), entre las secuencias alineadas sea w(j,1), w(j,2), w(j,3), y w(j,4), respectivamente. Para los fines de este análisis, puede suponerse sin pérdida de generalidad que todas las variaciones son bialélicas. Por lo tanto, pueden utilizarse las siguientes notaciones:

Recuento del alelo mayoritario en la posición genómica j como $B \equiv B_i \equiv \{b_j\} \equiv w_{j,i}^{(1)} = \max_{i\in\{1,2,3,4\}} \{w_{j,i}\}$

como estadístico de primer orden de los recuentos en la posición j. (El alelo mayoritario, b, es el argmax correspondiente. Los subíndices se utilizan cuando se está considerando más de un SNP).

Recuento del alelo minoritario en la posición j como $A \equiv A_i \equiv \{a_j\} = w_{j,i}^{(2)}$ como estadístico de segundo orden de los recuentos (es decir, el segundo recuento más alto de alelo) en la posición j,

Cobertura en la posición j como $D \equiv D_j = \{d_i\} = A_j + B_j$, y

Tasa de error de la máquina de secuenciación se indica como e.

Cuando el contexto está claro, por razones de conveniencia las notaciones se utilizan indistintamente; por ejemplo, pueden utilizarse indistintamente A, Ai, o {ai} para el alelo minoritario o el recuento del alelo minoritario. Pueden utilizarse subíndices, o no, dependiendo de si se está considerando más de un SNP. (Los SNPs se utilizan sólo a efectos de ejemplo. Pueden utilizarse otros tipos de polimorfismos como se analiza en otra parte del presente documento).

En la Figura 1, se representa la base para los cuatro estados de cigosidad para el polimorfismo. Como se ilustra, la madre puede ser homocigoto o heterocigoto en un determinado polimorfismo. Del mismo modo, el bebé puede ser heterocigoto u homocigoto en la misma posición. Como se ilustra, los casos 1 y 2 son los casos de polimorfismo en los que la madre es homocigoto. Si el bebé y la madre son ambos homocigotos, el polimorfismo es un polimorfismo de caso 1. Como se ha indicado anteriormente, esta situación no es por lo general particularmente interesante. Si la madre es homocigoto y el bebé es heterocigoto, la fracción fetal, *f*, viene dada concretamente por dos veces la relación entre el alelo minoritario y la cobertura. En el caso de polimorfismo en el que la madre es heterocigoto y el bebé es homocigoto (caso 3 de la Figura 1), la fracción fetal es concretamente uno menos dos veces la relación entre el alelo minoritario y la cobertura. Por último, en el caso en el que tanto la madre como el feto son heterocigotos, la fracción del alelo minoritario siempre debería ser 0,5, salvo error. La fracción fetal no puede obtenerse para polimorfismos que se encuentran en el caso 4.

A continuación se desarrollarán adicionalmente los cuatro casos.

Caso 1: Madre y bebé homocigotos

- En este caso, exceptuando errores de secuenciación o contaminación, no debería observarse ninguna diferencia.
- E (frecuencia del alelo min) = E(A) = 0.
- En la práctica, A ~ (se distribuye como) una distribución Binomial que se aproxima bien mediante la distribución de Poisson para una np baja. El parámetro tasa de distribución para la Binomial o de Poisson está relacionado con la tasa de error de secuenciación, e y la cobertura D. La Figura 3 muestra las frecuencias de desapareamiento de las secuencias 36-mero generadas alineadas con un genoma humano de referencia.
- Este caso no contiene información sobre la fracción fetal.

La Figura 3 presenta las estimaciones de error mediante la posición de la base secuenciada en 30 calles de datos de Illumina GA2 alineados con el genoma humano HG18 utilizando Eland con los parámetros por defecto.

Caso 2: Madre homocigoto y bebé heterocigoto

- En este caso, para una pequeña fracción fetal (*f*), las frecuencias alélicas observadas serán notablemente diferentes. Apareciendo el alelo mayoritario por lo general con una frecuencia varias veces mayor que el alelo minoritario.
- Salvo error, dada una sola posición de SNP (D,A), E(A) = Df/2 y una estimación no sesgada para f es 2A/D.
- Salvo error, A ~ Binomial (f/2,D). Media Df/2, Varianza (1-f/2)Df/2. [Dist. aproximadamente normal si D > 15].

Caso 3: Madre heterocigoto y bebé homocigoto

- En este caso, las frecuencias observadas para los alelos mayoritario y minoritario están próximas y A/D es ligeramente inferior a 0,5.
- Salvo error, E(A) = D(1-f)/2, y E(1 - (2A/D)) = f
- Salvo error, A ~ Binomial ((1-f)/2, D). Media D((1-f)/2). Varianza D/4(1-f^2).

Caso 4: Madre heterocigoto y bebé heterocigoto

Adviértase que, salvo error, hay dos subcasos para este.

Caso 4.1: El alelo del padre es diferente de los alelos de la madre. Esto introduciría un tercer alelo que sería el alelo minoritario, siendo E(A) = Df/2. Estos casos no deberían tener efecto sobre las estimaciones para f porque el procedimiento para asignar secuencias a los amplicones eliminará por filtración estos casos cuando los SNPs de referencia sean bialélicos.

Caso 4.2: El alelo del padre coincide con uno de los alelos de la madre

- En este caso, salvo error, los dos alelos aparecerían con una proporción 1:1 por lo que este caso no resulta útil para la estimación de la fracción fetal.

- Salvo error, E(A) = 0,5, y A ~ Binomial (0,5,D) truncado en 0,5.

La Figura 4 presenta un gráfico del recuento del alelo minoritario A frente a la cobertura D (suponiendo que no hay error) para los casos de heterocigosidad 1 a 4.

En diversas formas de realización, el método tiene que ver en términos generales con el análisis de la frecuencia alélica en uno o más SNPs (u otros polimorfismos) para clasificar los polimorfismos dentro del caso 2 y/o del caso 3. Utilizando la frecuencia alélica junto con la clasificación, el método puede estimar la fracción fetal.

En algunos casos, dado el recuento del alelo minoritario A y la cobertura D, es decir, un solo punto (D,A), para una posición de SNP individual permite que los métodos realicen una única estimación puntual. Por ejemplo, determinados métodos clasifican un SNP con el recuento de alelo (D,A) en un solo caso y obtienen una estimación de la fracción fetal del siguiente modo:

ES1.1 Umbrales simples para decidir el caso
Dada una posición individual (SNP),

1. Decidir sobre el caso 1 con una función de decisión como 2A/D < e o un valor crítico definido de la Binomial(e,D) o de Poisson(De). También puede utilizarse una distribución alternativa dentro del alcance de la presente invención. Sin estimación de la fracción fetal (f).

2. Decidir sobre el caso 4 si 2A/D > (0,5-e) o algún valor crítico de la Binomial(0,5,D), (u otra distribución de aproximación adecuada). No utilizar la posición para una estimación de f.

3. De lo contrario, decidir sobre el caso 2 si 2A/D < 0,25 (o algún otro umbral establecido manualmente o estimado automáticamente). La fracción fetal f se estimó como 2A/D.

4. De lo contrario, caso 3. Utilizar una estimación de la fracción fetal f = (1-2A/D).

Puede ganarse precisión combinando la información de recuento de alelos de varios SNPs para estimar la fracción fetal.

Método EM1: combinar múltiples SNPs calculando la media.

Tómense la media, la mediana, otra medición de centro (por ejemplo: biponderada de Turkey, estimadores M, etc.). También pueden utilizarse promedios ponderados. Para un ejemplo de cómo pueden definirse las ponderaciones, véase EM2.4 que se presenta más adelante. Además, pueden utilizarse medidas de centro robustas.

Método EM2: estimación simultánea a partir del caso 2 y del caso 3 mediante transformación

Para las ocasiones en las que f es inferior al X% de los puntos del caso 3 (D,A) puede transformarse para que coincida con los puntos del caso 2. A partir de esta línea, puede calcularse una pendiente común mediante regresión por el origen (véase la Figura 5).

Una desventaja teórica de los métodos basados en la transformación es que las distribuciones binomiales del caso 2 y 3 tendrán una forma diferente. A niveles típicos de fracción fetal (<10%) los datos del caso 2 tendrán una distribución próxima a Poisson sesgada hacia la derecha y el caso 3 tendrá una distribución cercana a la normal.

La Figura 5 representa la transformación de datos del caso 3 en caso 2. A continuación, una sola regresión puede estimar f a partir de ambos casos simultáneamente.

Método para calcular EM2.3:

Etapa 1: Desestimar los datos del caso 4.
Para cada punto de datos (D,A) si A > (0,5D-T1), excluir (D,A) del posterior análisis. T1(D,A) una función real.
Etapa 2: Transformar los datos del caso 3.
Véase la Figura 6. Para cada punto de datos (D,A) que no se dictamina que sea 4, si A > T2*D, transformar los puntos a nuevas coordenadas (D1,A1). T2(D,A) una función real.

$$\alpha = 2A / D$$

$$A1 = -1(0{,}5D - A)$$

$$D1 = D$$

Etapa 3: Establecer un umbral DT para reducir la contaminación debida a los datos del caso 1.
Ignorar todos los puntos de datos por debajo de T2(D,A) una función real.

Etapa 4: Realizar la estimación de la regresión para los restantes datos transformados del caso 2 y 3.
Aplicar la regresión por el origen hasta los puntos restantes. La estimación de la fracción fetal es dos veces la pendiente de la línea de regresión.

Adviértase que hay muchas clases de transformaciones que pueden construirse para lograr la misma coincidencia de los datos del caso 2 y 3. Los ejemplos incluyen la trigonométrica, la transformación o el uso de matrices de rotación. Se entiende que estas desviaciones quedan incluidas en el alcance de la presente descripción. Además, pueden utilizarse muchas clases de regresión (L2, L1,...) u optimización. Intercambiar el algoritmo de optimización es un cambio trivial y queda dentro del alcance de la presente descripción.

La Figura 6 presenta los datos después de la rotación, seleccionando D1 de manera que el caso 1 y los casos 2 y 3 no se solapen. E1 representa un límite superior del intervalo de confianza superior del 99 por ciento de los datos del caso 1.

Método EM3: mínimos cuadrados ponderados

El método de regresión de EM2.3 supone que todos los puntos de datos traducidos tienen igual varianza. Es más adecuado tener en cuenta la heterocedasticidad de las diferentes fuentes de datos e incluso de los puntos de un mismo patrón de heterocigosidad.

Las etapas 1 a 3 son idénticas a EM2.3.

Etapa 4: Regresión

En la regresión a partir de EM2.3, los puntos de los datos del caso 2 tendrán una varianza v2(f,D) = [0,5*Df-0,25*Df^2] y los puntos de los datos del caso 3 tendrán una varianza v3(f,D)=[0,25D(1 - f^2)]. Suponiendo que se da a cada punto una ponderación diferente, w, como en EM2.3, se busca minimizar

$$Q = \sum_{i=1}^{n} w_i (a_i - sd_i)^2$$

Ecuación 1

Ajuste de las primeras derivadas a cero y cálculo del valor de s:

$$\frac{\partial Q}{\partial s} = \sum_{i=1}^{n} 2w_i(d_i - sa_i)(-a_i) = 0$$

$$\sum_{i=1}^{n} sa_i^2 - \sum_{i=1}^{n} 2w_i a_i x_i = 0$$

y

$$s = \frac{\sum_{i=1}^{n} 2w_i d_i a_i}{\sum_{i=1}^{n} a_i^2}$$

donde d_i es la cobertura de SNP i y a_i es el recuento del alelo minoritario (transformado para el caso 3) de SNP i.

Ecuación 2

Este método se pondera con la inversa de la varianza de cada punto, que se estima como v2(2A/D,D), o v3(2A/D,D), según corresponda. La estimación de la fracción fetal es 2*s.

En determinadas formas de realización, puede emplearse un modelo de mezcla para clasificar un grupo de polimorfismos en dos o más de los casos de cigosidad y, al mismo tiempo, estimar la fracción de ADN fetal a partir de las frecuencias alélicas medias para cada uno de estos casos. En general, un modelo de mezcla supone que un conjunto de datos particular se compone de una mezcla de diferentes tipos de datos, cada uno de los cuales tiene su propia distribución esperada (por ejemplo, una distribución normal). El proceso intenta encontrar la media y, posiblemente, otras características para cada tipo de datos. En las formas de realización descritas en el presente documento, hay hasta cuatro tipos de datos diferentes (los casos de cigosidad) que constituyen los datos de frecuencia del alelo minoritario para los polimorfismos en cuestión.

En la siguiente sección se presenta una implementación de un modelo de mezcla. En esta forma de realización, la frecuencia del alelo minoritario A es una suma de cuatro términos como se muestra en la ecuación 3. Cada uno de los términos corresponde a uno de los cuatro casos de cigosidad. Cada término es el producto de una fracción de polimorfismo α y una distribución binomial de la frecuencia del alelo minoritario. Los α son las fracciones de los polimorfismos que se encuentran en cada uno de los cuatro casos. Cada distribución binomial tiene una probabilidad asociada, p, y una cobertura, *d*. La probabilidad del alelo minoritario para el caso 2, por ejemplo, viene dada por *f/2*.

Las formas de realización descritas se valen de "momentos factoriales" para los datos de frecuencia alélica en cuestión. Como es bien sabido, la media de la distribución es el primer momento. Es el valor esperado de la frecuencia del alelo minoritario. La varianza es el segundo momento. Se calcula a partir del valor esperado de la frecuencia alélica al cuadrado.

Los datos de frecuencia alélica en todos los polimorfismos pueden utilizarse para calcular los momentos factoriales (un primer momento factorial, un segundo momento factorial, etc.) como se muestra en la ecuación 4. Como indican estas ecuaciones, los momentos factoriales son los sumatorios de términos sobre *i*, los polimorfismos individuales en el conjunto de datos, en los que hay *n* de tales polimorfismos en el conjunto de datos. Los términos que se suman son funciones de los recuentos de alelos minoritarios, α_i, y las coberturas d_i.

De manera provechosa, los momentos factoriales tienen relaciones con los valores de α_i y p_i como se ilustra en la ecuación 5. A partir de las probabilidades, p_i, puede determinarse la fracción fetal, f. Por ejemplo, $p_2 = f/2$ y p_3 es 1 - *f/2*. Por lo tanto, la lógica responsable puede resolver un sistema de ecuaciones que relacionan los αs y *ps* desconocidos con las expresiones de momento factorial para las fracciones del alelo minoritario en los polimorfismos múltiples en cuestión. Por supuesto, hay otras técnicas para resolver los modelos de mezcla dentro del alcance de la presente invención.

Resulta útil tener en cuenta además los fundamentos matemáticos o simbólicos de las formas de realización de modelo de mezcla descritas en el presente documento. Los cuatro casos de heterocigosidad descritos anteriormente sugieren el siguiente modelo de mezcla Binomial para la distribución de ai en los puntos (ai,di):

$$A = \{a_i\} \sim \alpha_1 Bin(p_1, d_i) + \alpha_2 Bin(p_2, d_i) + \alpha_3 Bin(p_3, d_i) + \alpha_4 Bin(p_4, d_i)$$

donde

$$1 = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4$$
$$m = 4$$

Ecuación 3

Más adelante se describen varios modelos para relacionar el p_i con la fracción fetal y secuenciar las tasas de error. Los parámetros α_i se relacionan con los parámetros específicos de la población y la capacidad para dejar que estos valores "floten" da a estos métodos robustez adicional con respecto a factores como la etnia y la descendencia de los progenitores.

Para diversos casos de heterocigosidad puede resolverse la ecuación anterior para la fracción fetal. Quizás el método más fácil de calcular el valor de la fracción fetal es mediante el método de los momentos factoriales en el que los parámetros de la mezcla pueden expresarse en términos de momentos que puede estimarse fácilmente a partir de los datos observados.

Dadas n posiciones de SNP, los momentos factoriales se definen de la siguiente manera:

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i}{d_i}$$
$$F_2 = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i(a_i-1)}{d_i(d_i-1)}$$
$$\cdots$$
$$F_j = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i(a_i-1)\cdots(a_i-j+1)}{d_i(d_i-1)(d_i-j+1)}$$

Ecuación 4

Los momentos factoriales pueden relacionarse con el $\{\alpha_i, p_j\}$ con

$$F_1 \approx \sum_{i=1}^{m} \alpha_i p_i^1$$

$$F_2 \approx \sum_{i=1}^{m} \alpha_i p_i^2$$

...

$$F_j \approx \sum_{i=1}^{m} \alpha_i p_i^j$$

...

$$F_g \approx \sum_{i=1}^{m} \alpha_i p_i^g$$

Ecuación 5

Puede identificarse una solución calculando el valor de $\{\alpha_i, p_i\}$ en un sistema de ecuaciones derivado de la anterior relación Ecuación 5 cuando n > 2* (número de parámetros a estimar). Obviamente, el problema se vuelve mucho más difícil matemáticamente para una g, ya que es necesario estimar más $\{\alpha_i, p_i\}$.

Por lo general no es posible discriminar con precisión entre los datos del caso 1 y 2 (o del caso 3 y 4) mediante umbrales simples en fracciones fetales más bajas. Afortunadamente para el uso de modelos de casos reducidos, los datos de los casos 1/2 se separan fácilmente de los datos de los casos 3/4 mediante la discriminación en el punto (2A/D)=T. Se ha descubierto que el uso de T=0,5 se comporta satisfactoriamente.

Adviértase que el método de modelo de mezcla que emplea las ecuaciones 4 y 5 se vale de los datos para todos los polimorfismos pero no tiene en cuenta por separado el error de secuenciación. Los métodos apropiados que separan los datos para los casos primero y segundo de los datos para los casos tercero y cuarto pueden tener en cuenta el error de secuenciación.

En otros ejemplos, el conjunto de datos proporcionado a un modelo de mezcla contiene datos sólo para polimorfismos del caso 1 y del caso 2. Estos son los polimorfismos para los que la madre es homocigoto. Puede emplearse una técnica de umbralización para eliminar los polimorfismos del caso 3 y 4. Por ejemplo, los polimorfismos con frecuencias del alelo minoritario superiores a un umbral particular se eliminan antes de emplear el modelo de mezcla. Utilizando los datos debidamente filtrados y los momentos factoriales como reducidos a las ecuaciones 7 y 8, puede calcularse la fracción fetal, *f*, como se muestra en la ecuación 9. Obsérvese que la ecuación 7 es una reformulación de la ecuación 3 para esta implementación de un modelo de mezcla. Adviértase también que en este ejemplo concreto, no se conoce el error de secuenciación asociado con la lectura de la máquina. Como consecuencia, el sistema de ecuaciones debe calcular el valor del error por separado, *e*.

La Figura 7 muestra una comparación de los resultados utilizando este modelo de mezcla y la fracción fetal conocida (eje x) y la fracción fetal estimada. Si el modelo de mezcla predijo perfectamente la fracción fetal, los resultados representados gráficamente seguirían la línea discontinua. Sin embargo, las fracciones estimadas son muy buenas, sobre todo teniendo en cuenta que muchos de los datos se eliminaron antes de aplicar el modelo de mezcla.

Para dar más detalles, se dispone de otros varios métodos para la estimación de parámetros del modelo a partir de la Ecuación 3. En algunos casos puede encontrarse una solución manejable estableciendo las derivadas en cero del estadístico chi-cuadrado. En los casos en que no pueda encontrarse una solución fácil mediante diferenciación directa, puede resultar eficaz el desarrollo en serie de Taylor de la PDF binomial u otros polinomios de aproximación. Se sabe que los estimadores chi-cuadrado mínimo son eficaces.

$$\chi^2(\alpha_i, p_i) = \sum_{i=1}^{n} \frac{\left(P_i - \sum \alpha_i Binomial(p_i, di) \right)^2}{Binomial(n, p)}$$

Ecuación 6

donde Pi es el número de puntos del recuento i. Un método alternativo de Le Cam ["On the Asymptotic Theory of Estimation and Testing Hypotheses". Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, volumen 1, Berkeley CA: University of CA Press, 1956, págs. 129-156] utiliza la iteración de Ralph-Newton de la función de verosimilitud. Puede utilizarse el método de soluciones de momentos de la Ecuación 5 como punto de partida para la iteración.

En otra aplicación, se analiza un método de resolución de modelos de mezcla que implica métodos de maximización de la expectativa que operan en mezclas de distribuciones Beta de aproximación.

Casos (1+2) modelo, error de secuenciación desconocido

Considérese un modelo reducido que sólo tiene en cuenta los casos de heterocigosidad 1 y 2. En este caso, la distribución de mezcla puede escribirse como

$$A = \{a_i\} \sim \alpha_1 Bin(e, d_i) + \alpha_2 Bin(f/2, d_i)$$

donde

$$1 = \alpha_1 + \alpha_2$$
$$m = 4$$

Ecuación 7

Y el sistema

$$F_1 = \alpha_1 e + (1 - \alpha_1)(f/2)$$
$$F_2 = \alpha_1 e^2 + (1 - \alpha_1)(f/2)^2$$
$$F_3 = \alpha_1 e^3 + (1 - \alpha_1)(f/2)^3$$

Ecuación 8

se resuelve para el e (tasa de error de secuenciación), alfa (proporción de puntos del caso 1), y f (fracción fetal). Donde la Fi se define como en la Ecuación 4 anterior. Se elige que una solución de forma cerrada para la fracción fetal sea la verdadera solución de

$$F \approx \frac{(F1-1)F2 \pm \sqrt{F2}\sqrt{4F1^3 + F2 - 3F1(2+F1)F2 + 4F2^2}}{2(F1^2 - F2)}$$

Ecuación 9

que se encuentra entre 0 y 1.

Para medir el comportamiento de los estimadores se construyó un conjunto de datos simulado de puntos de Equilibrio de Hardy-Weinberg (ai,di) diseñándose la fracción fetal para que fuese {1%, 3%, 5%, 10%, 15%, 20 % y 25%} y una tasa de error de secuenciación constante del 1%. La tasa de error del 1% es la tasa actualmente aceptada para los protocolos y las máquinas de secuenciación que se están utilizando y es coherente con la gráfica de datos del analizador Illumina Genome Analyzer II que se muestran en la Figura 3 anterior. Se aplicó la Ecuación 9 a los datos y se encontró, a excepción de un sesgo por exceso de cuatro puntos, una concordancia general con la fracción fetal "conocida". Curiosamente, se estima que la tasa de error de secuenciación, e, es ligeramente superior al 1%.

En el siguiente ejemplo de modelo de mezcla, se emplea de nuevo la umbralización u otra técnica de filtrado para eliminar los datos de los polimorfismos que se encuentran en los casos 3 y 4. Sin embargo, en este caso, se conoce el error de secuenciación. Esto simplifica la expresión resultante para la fracción de ADN fetal, *f*, como se muestra en las ecuaciones 10. La Figura 8 muestra que esta versión de un modelo de mezcla proporcionaba mejores resultados en comparación con el enfoque empleado con la ecuación 9.

En las ecuaciones 11 y 12 se muestra un enfoque similar. Este enfoque reconoce que sólo algunos errores de secuenciación se suman al recuento del alelo minoritario. Más bien sólo uno de cada cuatro errores de secuenciación debería aumentar el recuento del alelo minoritario. La figura 9 muestra una concordancia muy buena entre las fracciones fetales real y estimada utilizando esta técnica.

Casos (1+2) modelo, error de secuenciación conocido

Dado que se conoce en gran medida la tasa de error de secuenciación de las máquinas utilizadas, puede reducirse el sesgo y la complejidad de los cálculos eliminando e como variable a resolver. Por lo tanto, se obtiene el sistema de ecuaciones

$$F_1 = \alpha_1 e + (1-\alpha_1)(f/2)$$
$$F_2 = \alpha_1 e^2 + (1-\alpha_1)(f/2)^2$$

Ecuación 10

para la fracción fetal f, para obtener la solución:

$$F \approx \frac{2(eF1 - F2)}{(e - F1)}$$

La Figura 8 muestra que el uso de la tasa de error de la máquina como parámetro conocido reduce en un punto el sesgo por exceso.

Casos (1+2) modelo, error de secuenciación conocido, modelos de error mejorados

Para mejorar el sesgo en el modelo se amplió el modelo de error de las ecuaciones anteriores para tener en cuenta el hecho de que no todos los eventos de error de secuenciación se sumarán al recuento del alelo minoritario A=ai en el caso de heterocigosidad 1. Además, se tienen en cuenta el hecho de que los eventos de error de secuenciación puedan contribuir a los recuentos del caso de heterocigosidad 2. Por lo tanto, se determinó la fracción fetal F resolviendo el siguiente sistema de relaciones de momentos factoriales:

$$F_1 = \alpha_1 e/4 + (1-\alpha_1)(e + f/2)$$
$$F_2 = \alpha_1 (\frac{e}{4})^2 + (1-\alpha_1)(e + f/2)^2$$

Ecuación 11

que da la solución

$$F \approx \frac{-2(e^2 - 5eF1 + 4F2)}{(e - 4F1)}.$$

Ecuación 12

En la Figura 9 se muestra que los datos simulados utilizando la tasa de error de la máquina como parámetro conocido, que mejora los modelos de error del caso 1 y 2, reduce en gran medida el sesgo por exceso a menos de un punto para la fracción fetal por debajo de 0,2.

Opciones de implementación

MUESTRAS

Las muestras que se utilizan en las formas de realización descritas en el presente documento comprenden ADN genómico celular o libre. El ADN celular se deriva de células enteras mediante la extracción manual o mecánica del ADN genómico a partir de células enteras de la misma o de diferentes composiciones genéticas. El ADN celular puede derivarse, por ejemplo, de células enteras de la misma composición genética procedentes de un sujeto, de una mezcla de células enteras de diferentes sujetos, o de una mezcla de células enteras que difieren en composición genética que proceden de un sujeto. Los métodos para extraer ADN genómico de células enteras son conocidos en la técnica, y difieren en función de la naturaleza de la fuente.

En algunos casos, puede ser ventajoso fragmentar el ADN genómico celular. La fragmentación puede ser aleatoria, o puede ser específica, tal como se consigue, por ejemplo, mediante digestión con endonucleasas de restricción. Los métodos para la fragmentación aleatoria son conocidos en la técnica, e incluyen, por ejemplo, la digestión limitada con ADNasa, el tratamiento alcalino y el fraccionamiento físico. En determinadas formas de realización, los ácidos nucleicos de la muestra se someten a fragmentación en fragmentos de aproximadamente 500 o más pares de bases, y a los que pueden aplicarse fácilmente los métodos de secuenciación de nueva generación (NGS). En una forma de realización, los ácidos nucleicos de la muestra se obtienen a partir de cfADN, que no se somete a fragmentación.

El ADN libre es el ADN genómico que se produce de forma natural como una mezcla de fragmentos genómicos que se encuentran por lo general en los fluidos biológicos, por ejemplo la sangre, de un sujeto. La mezcla genómica puede derivarse de células que se rompen de forma natural para liberar su contenido genómico mediante procesos biológicos, por ejemplo, la apoptosis. Una muestra de cfADN puede comprender cfADN derivado de una mezcla de células de diferentes sujetos de la misma especie, de una mezcla de células de un sujeto que difieren en composición genética, o de una mezcla de células de diferentes especies, por ejemplo, un sujeto.

Los ácidos nucleicos libres, incluido el ADN libre, pueden obtenerse mediante diversos métodos conocidos en la técnica a partir de muestras biológicas, incluidas pero no limitadas a plasma, suero y orina (Fan *et al*., Proc. Natl. Acad. Sci. 105:16266-16271 [2008]; Koide *et al*., Prenatal Diagnosis 25:604-607 [2005]; Chen *et al*., Nature Med. 2: 1033-1035 [1996]; Lo *et al*., Lancet 350: 485-487 [1997.]; Botezatu *et al*., Clin. Chem. 46: 1078-1084, 2000; y Su *et al*., J. Mol. Diagn. 6: 101-107 [2004]). Para separar el cfADN de las células, pueden utilizarse el fraccionamiento, la centrifugación (por ejemplo, centrifugación en gradiente de densidad), la precipitación específica de ADN, o la separación de células de alto rendimiento y/o métodos de separación. Se dispone en el mercado de kits para la separación manual y automatizada de cfADN (Roche Diagnostics, Indianapolis, IN, Qiagen, Valencia, CA, Macherey-Nagel, Düren, Alemania).

La muestra que comprende la mezcla de ácidos nucleicos a la que se aplican los métodos descritos en el presente documento puede ser una muestra biológica tal como una muestra de tejido, una muestra de fluido biológico o una muestra de células. En algunas formas de realización, la mezcla de ácidos nucleicos se purifica o aísla a partir de la muestra biológica mediante cualquiera de los métodos conocidos. Una muestra puede ser un polinucleótido purificado o aislado. Un fluido biológico incluye, como ejemplos no limitativos, sangre, plasma, suero, sudor, lágrimas, esputo, orina, secreción del oído, linfa, saliva, líquido cefalorraquídeo, lavados, suspensión de médula ósea, flujo vaginal, lavado transcervical, líquido cerebral, líquido ascítico, leche, secreciones de los tractos respiratorio, intestinal y genitourinario, líquido amniótico y muestras de leucoforesis. En algunas formas de realización, la muestra es una muestra que se obtiene fácilmente mediante procedimientos no invasivos, por ejemplo sangre, plasma, suero, sudor, lágrimas, esputo, orina, secreción del oído, saliva o heces. Preferentemente, la muestra biológica es una muestra de sangre periférica, o las fracciones de plasma y suero. En otras formas de realización, la muestra biológica es un hisopo o un frotis, una muestra de biopsia, o un cultivo celular. En otra forma de realización, la muestra es una mezcla de dos o más muestras biológicas, por ejemplo, una muestra biológica puede comprender dos o más de entre una muestra de fluido biológico, una muestra de tejido, y una muestra de cultivo celular. Tal como se utilizan en el presente documento, los términos "sangre", "plasma" y "suero" incluyen

expresamente fracciones o porciones procesadas de los mismos. Del mismo modo, cuando se toma una muestra de una biopsia, un hisopo, un frotis, etc., el término "muestra" incluye expresamente una fracción procesada o porción obtenida de la biopsia, el hisopo, el frotis, etc.

En algunas formas de realización, las muestras pueden obtenerse de fuentes, incluidas, pero no limitadas a, muestras de diferentes individuos, diferentes etapas de desarrollo del mismo o de diferentes individuos, diferentes individuos enfermos (por ejemplo, individuos con cáncer o que se sospecha tienen un trastorno genético), individuos normales, muestras obtenidas en diferentes etapas de una enfermedad en un individuo, muestras obtenidas de un individuo sometido a diferentes tratamientos para una enfermedad, muestras de individuos sometidos a diferentes factores ambientales, o individuos con predisposición a una patología, o individuos con exposición a un agente de una enfermedad infecciosa (por ejemplo, VIH).

En una forma de realización, la muestra es una muestra materna que se obtiene de una hembra embarazada, por ejemplo, una mujer embarazada. En este caso, la muestra puede analizarse utilizando los métodos descritos en el presente documento para proporcionar un diagnóstico prenatal de posibles anomalías cromosómicas en el feto. La muestra materna puede ser una muestra de tejido, una muestra de fluido biológico o una muestra de células. Un fluido biológico incluye, como ejemplos no limitativos, sangre, plasma, suero, sudor, lágrimas, esputo, orina, secreción del oído, linfa, saliva, líquido cefalorraquídeo, lavados, suspensión de médula ósea, flujo vaginal, lavado transcervical, líquido cerebral, líquido ascítico, leche, secreciones de los tractos respiratorio, intestinal y genitourinario, y muestras de leucoforesis. En otra forma de realización, la muestra materna es una mezcla de dos o más muestras biológicas, por ejemplo, una muestra biológica puede comprender dos o más de entre una muestra de fluido biológico, una muestra de tejido, y una muestra de cultivo celular. En algunas formas de realización, la muestra es una muestra que puede obtenerse fácilmente mediante procedimientos no invasivos, por ejemplo sangre, plasma, suero, sudor, lágrimas, esputo, orina, secreción del oído, saliva y heces. En algunas formas de realización, la muestra biológica es una muestra de sangre periférica, o las fracciones de plasma y suero. En otras formas de realización, la muestra biológica es un hisopo o un frotis, una muestra de biopsia, o un cultivo celular.

Las muestras también pueden obtenerse a partir de células, tejidos cultivados *in vitro*, u otras fuentes que contienen polinucleótidos. Las muestras cultivadas pueden tomarse de fuentes que incluyen, pero no se limitan a, cultivos (por ejemplo, tejido o células) que se mantienen en diferentes medios y condiciones (por ejemplo, pH, presión o temperatura), cultivos (por ejemplo, tejido o células) que se mantienen durante diferentes períodos de duración, cultivos (por ejemplo, tejido o células) tratados con diferentes factores o reactivos (por ejemplo, un fármaco potencial o un modulador), o cultivos de diferentes tipos de tejido o células. Los métodos de aislamiento de ácidos nucleicos a partir de fuentes biológicas son conocidos y diferirán dependiendo de la naturaleza de la fuente como se ha explicado anteriormente.

POLIMORFISMOS PARA SU USO EN LA IDENTIFICACIÓN DE LA FRACCIÓN GENÓMICA

Como se ha explicado, los polimorfismos pueden utilizarse para evaluar la fracción fetal. En la evaluación se utiliza la fracción alélica y la cigosidad de uno o más polimorfismos. Los ejemplos de polimorfismos útiles incluyen, sin limitación, polimorfismos de un sólo nucleótido (SNP), SNPs en tándem, deleciones o inserciones de múltiples bases a pequeña escala, denominadas indels (también denominadas polimorfismos inserción/deleción o DIPs), polimorfismos de múltiples nucleótidos (MNP), repeticiones cortas en tándem (STR), polimorfismos en la longitud de los fragmentos de restricción (RFLPs), deleciones, incluidas microdeleciones, inserciones, incluidas microinserciones, duplicaciones, inversiones, translocaciones, multiplicaciones, variantes complejas multisitio, variaciones en el número de copias (CNV), y polimorfismos que comprenden cualquier otro cambio de secuencia en un cromosoma.

En algunas formas de realización, los polimorfismos que se utilizan en los métodos descritos incluyen SNPs y/o STRs. Los polimorfismos SNP pueden ser SNP únicos, SNPs en tándem. Los SNPs únicos incluyen SNPs individuales y SNPs marcadores, es decir SNPs presentes en un haplotipo, y/o un bloque de haplotipos. En algunas formas de realización, se utilizan combinaciones de polimorfismos. Por ejemplo, pueden detectarse diferencias en el número de copias comparando una combinación de secuencias polimórficas que comprenden uno o más SNPs y una o más STRs.

En general, puede utilizarse cualquier sitio polimórfico que pueda quedar abarcado por las lecturas generadas mediante los métodos de secuenciación descritos en el presente documento para identificar la fracción genómica en las muestras que comprenden ADN de diferentes genomas. Las secuencias polimórficas útiles para poner en práctica los métodos de la invención están disponibles en diversas bases de datos de acceso público, que están en continua expansión. Por ejemplo, las bases de datos útiles incluyen, sin limitación, la Human SNP Database en la dirección web wi.mit.edu, la página inicial de dbSNP del NCBI en la dirección web ncbi.nlm.nih.gov, la dirección web lifesciences.perkinelmer.com, la base de datos de SNP de Celera Human en la dirección web celera.com, la base de datos de SNP del Genome Analysis Group (GAN) en la dirección web gan.iarc.fr., la base de datos de repeticiones cortas en tándem (STR) de la ATCC en la dirección web atcc.org, y la base de datos HapMap en la dirección web hapmap.org.

El número de polimorfismos que pueden utilizarse en una evaluación de la fracción fetal puede ser al menos 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1.000 o más. Por ejemplo, se estima que el genoma humano comprende al menos aproximadamente 10 millones de SNPs. Por lo tanto, el número de polimorfismos disponibles que pueden genotiparse en una muestra de un paciente humano puede ser al menos aproximadamente 10 millones de SNPs, así como muchos otros tipos de polimorfismos que están presentes en cualquier genoma humano. En algunas formas de realización, la identificación de uno o más polimorfismos en un primer genoma de una muestra que comprende una mezcla de ADN, por ejemplo, cfADN, de un primer y un segundo genoma se realiza mediante secuenciación del genoma completo utilizando un método de NGS como se describe en el presente documento. En algunas formas de realización, el método de secuenciación del genoma completo es un método de NGS que identifica las secuencias polimórficas mediante secuenciación masiva en paralelo de moléculas de ácido nucleico amplificadas por clonación o mediante secuenciación masiva en paralelo de moléculas individuales de ácido nucleico, es decir, secuenciación de una sola molécula.

APLICACIONES

La fracción de ácido nucleico procedente de cada una de las dos fuentes genómicas distintas en una muestra puede utilizarse con diversos fines. En diversas formas de realización descritas en el presente documento, se utiliza la fracción de ADN fetal en el ADN libre de una muestra de material para facilitar diagnósticos prenatales y para ayudar en la toma de decisiones sobre el tratamiento de los embarazos. En otras formas de realización, los genomas en cuestión no son maternos ni fetales. Más adelante se presentan diversos ejemplos de fuentes genómicas para determinar la presencia del genoma fraccional.

Puede utilizarse ARN y ADN fetal libre que circula en la sangre materna para el diagnóstico prenatal no invasivo (NIPD) precoz de un número cada vez mayor de afecciones genéticas, tanto para la gestión del embarazo como para facilitar la toma de decisiones sobre la reproducción. Durante el embarazo, hay pequeñas cantidades de ADN fetal circulante en el torrente sanguíneo materno (Lo *et al*., Lancet 350: 485-487 [1997]). Se ha demostrado que el ADN fetal libre, que se cree se deriva de células moribundas de la placenta, consiste en fragmentos cortos por lo general de menos de 200 pb de longitud (Chan *et al*., Clin. Chem. 50:88-92 [2004]), que pueden distinguirse ya a las 4 semanas de gestación (Illanes *et al*., Early Human Dev. 83:563-566 [2007]), y se sabe se eliminan de la circulación materna pocas horas después del parto (lo *et al*., Am. J. Hum. Genet. 64:218-224 [1999]). Además del cfADN, también pueden distinguirse fragmentos de ARN fetal libre (cfRNA) en el torrente sanguíneo materno, procedentes de genes que se transcriben en el feto o la placenta. La extracción y el análisis posterior de estos elementos genéticos fetales a partir de una muestra de sangre materna ofrecen nuevas oportunidades para el NIPD.

Como se ha explicado, los métodos descritos determinan la fracción de un segundo genoma en una muestra biológica. Los métodos determinan opcionalmente la presencia o ausencia de varios trastornos en una muestra de sangre que comprende una mezcla de ADN (tal como cfADN) de un primer y un segundo genoma. En algunas formas de realización, la determinación de la fracción fetal puede comprender (a) secuenciar el genoma de al menos una porción de la mezcla de cfADN para obtener una pluralidad de marcadores de secuencia; (b) determinar en la pluralidad de marcadores de secuencia la presencia o ausencia de polimorfismos múltiples, y (c) asociar los polimorfismos múltiples con el primer y/o el segundo genoma en la mezcla. En formas de realización preferentes, la mezcla no está enriquecida en polimorfismos múltiples. La identificación de los polimorfismos múltiples en la mezcla de ADN se realiza comparando la secuencia de los marcadores mapeados obtenidos mediante el método de secuenciación del genoma completo con polimorfismos múltiples de referencia, como se describe en el presente documento.

En la forma de realización descrita anteriormente, el primer genoma es un genoma fetal, y un segundo genoma es un genoma materno. En otra forma de realización, el primer genoma es un genoma de una célula no afectada y el segundo genoma es un genoma de una célula afectada, por ejemplo, una célula cancerosa. En algunas formas de realización, las células afectadas y no afectadas proceden del mismo sujeto. Por ejemplo, la célula afectada puede ser una célula cuyo genoma ha sido modificado por un trastorno. En algunas formas de realización, el trastorno es un trastorno monogénico. En otras formas de realización, el trastorno es un trastorno poligénico. Los trastornos pueden identificarse mediante un solo polimorfismo, por ejemplo, un SNP marcador, o mediante polimorfismos múltiples presentes en un haplotipo. En algunas formas de realización, los polimorfismos múltiples identificados según el método de la presente invención están presentes en un bloque de haplotipos.

Los trastornos que pueden identificarse con ayuda del método de la presente invención son trastornos genéticos, que son enfermedades debidas al menos en parte a anomalías en los genes o cromosomas. El conocimiento de una fracción fetal en una muestra puede ayudar a identificar tales trastornos en un contexto prenatal. Los trastornos identificados mediante el método de la presente invención incluyen trastornos monogénicos, es decir, de un solo gen, y trastornos poligénicos, es decir, complejos. Los trastornos de un solo gen incluyen autosómicos dominantes, autosómicos recesivos, dominantes ligados al cromosoma X, recesivos ligados al cromosoma X, y ligados al cromosoma Y.

En los trastornos autosómicos dominantes, será necesaria sólo una copia mutada del gen para que una persona se vea afectada por el trastorno. Por lo general, un sujeto afectado tiene un padre afectado, y hay una probabilidad del 50% de que la descendencia herede el gen mutado. Las afecciones que son autosómicas dominantes tienen a veces una penetrancia reducida, lo que significa que aunque sólo se necesita una copia mutada, no todos los individuos que heredan la mutación llegan a desarrollar la enfermedad. Los ejemplos de trastornos autosómicos dominantes que pueden identificarse mediante el método de la presente invención incluyen, sin limitación, la hipercolesterolemia familiar, la esferocitosis hereditaria, el síndrome de Marfan, la neurofibromatosis de tipo 1, el cáncer colorrectal hereditario sin poliposis, y la exostosis múltiple hereditaria, y la enfermedad de Huntington.

Los trastornos autosómicos recesivos detectados utilizando el método de la presente invención incluyen la drepanocitemia, la fibrosis quística, la enfermedad de Tay-Sachs, la mucopolisacaridosis, las enfermedades de almacenamiento de glucógeno, y la galactosemia. Los trastornos ligados al cromosoma X detectados mediante el método de la presente invención incluyen la distrofia muscular de Duchenne y la hemofilia. En los trastornos autosómicos recesivos, deben estar mutadas dos copias del gen para que un sujeto que se vea afectado por un trastorno autosómico recesivo. Un sujeto afectado suele tener progenitores no afectados que portan cada uno una sola copia del gen mutado (y se conocen como portadores). Dos personas no afectadas que portan cada una copia del gen mutado tienen una probabilidad del 25% en cada embarazo de tener un/a hijo/a afectado/a por el trastorno. Los ejemplos de este tipo de trastorno que pueden identificarse mediante el método de la presente invención incluyen la fibrosis quística, la drepanocitemia, la enfermedad de Tay-Sachs, la enfermedad de Niemann-Pick, la atrofia muscular espinal, y el síndrome de Roberts. Otros determinados fenotipos, tal como el cerumen húmedo frente al seco, también se determinan de forma autosómica recesiva. Los trastornos dominantes ligados al cromosoma X se deben a mutaciones en los genes del cromosoma X. Sólo unos pocos trastornos tienen este patrón de herencia, siendo un exponente el raquitismo hipofosfatémico ligado al cromosoma X. Tanto varones como mujeres se ven afectados en estos trastornos, siendo por lo general los varones más gravemente afectados que las mujeres. Algunas afecciones dominantes ligadas al cromosoma X, tal como el síndrome de Rett, la incontinencia pigmentaria de tipo 2 y el síndrome de Aicardi suelen ser mortales en los varones, y por lo tanto se observa de manera predominante en las mujeres. Las excepciones a este hallazgo son casos extremadamente raros en los que los niños con síndrome de Klinefelter (47, XXY) también heredan una afección dominante ligada al cromosoma X y presentan síntomas más similares a los de una mujer en términos de gravedad de la enfermedad. La probabilidad de transmitir un trastorno dominante ligado al cromosoma X difiere entre hombres y mujeres. Ninguno de los hijos de un hombre con un trastorno dominante ligado al cromosoma X se verá afectado (ya que reciben el cromosoma Y de su padre), y todas sus hijas heredarán la afección. Una mujer con un trastorno dominante ligado al cromosoma X tiene una probabilidad del 50% de tener un feto afectado en cada embarazo, aunque cabe señalar que en casos como el de la incontinencia pigmentaria sólo la descendencia femenina es generalmente viable. Además, aunque estas afecciones no modifican la fertilidad *per se*, los individuos con el síndrome de Rett o el síndrome de Aicardi rara vez se reproducen.

El método de la presente invención también puede facilitar la identificación de polimorfismos asociados con trastornos ligados al cromosoma X. Las afecciones recesivas ligadas al cromosoma X también se deben a mutaciones en genes del cromosoma X. Los varones se ven más frecuentemente afectados que las mujeres, y la probabilidad de transmitir el trastorno difiere entre hombres y mujeres. Los hijos de un hombre con un trastorno recesivo ligado al cromosoma X no se verán afectados, y sus hijas portarán una copia del gen mutado. Una mujer que es portadora de un trastorno recesivo ligado al cromosoma X (X^RX^r) tiene una probabilidad del 50% de tener hijos que se vean afectados y una probabilidad del 50% de tener hijas que porten una copia del gen mutado y que, por lo tanto, sean portadoras. Las afecciones recesivas ligadas al cromosoma X incluyen, sin limitación, las graves enfermedades hemofilia A, distrofia muscular de Duchenne y síndrome de Lesch-Nyhan, así como afecciones comunes y menos graves tales como la calvicie de patrón masculino y la ceguera para el rojo-verde. Las afecciones recesivas ligadas al cromosoma X pueden manifestarse a veces en las mujeres debido a una inactivación sesgada del cromosoma X o una monosomía X (síndrome de Turner).

Los trastornos ligados al cromosoma Y se deben a mutaciones en el cromosoma Y. Debido a que los varones heredan un cromosoma Y del padre, *cada* hijo de un padre afectado se verá afectado. Debido a que las mujeres heredan un cromosoma X de su padre, la descendencia femenina de padres afectados *nunca* se ve afectada. Puesto que el cromosoma Y es relativamente pequeño y contiene muy pocos genes, existen relativamente pocos trastornos ligados al cromosoma Y. Con frecuencia, los síntomas incluyen la infertilidad, que puede salvarse con ayuda de algunos tratamientos de fertilidad. Los ejemplos son la infertilidad masculina y la hipertricosis auricular.

Como se ha explicado, los métodos descritos para detectar fracciones genómicas en una muestra pueden utilizarse para facilitar la detección de una aneuploidía a partir de muestras de material. En algunas formas de realización, la aneuploidía es una monosomía o trisomía cromosómica completa, o una monosomía o trisomía parcial. Las aneuploidías parciales se deben a la pérdida o ganancia de una parte de un cromosoma, y abarcan desequilibrios cromosómicos resultado de translocaciones desequilibradas, inversiones desequilibradas, deleciones e inserciones. Con diferencia, la aneuploidía conocida más común compatible con la vida es la trisomía 21 es decir, el síndrome de Down (DS), que se debe a la presencia de parte o la totalidad del cromosoma 21. En raras ocasiones, el DS puede deberse a un defecto hereditario o esporádico mediante el cual una copia adicional de parte

o la totalidad del cromosoma 21 se une a otro cromosoma (por lo general el cromosoma 14) para formar un único cromosoma aberrante. El DS se asocia con la deficiencia intelectual, dificultades de aprendizaje graves y el exceso de mortalidad debida a problemas de salud a largo plazo tales como las cardiopatías. Otras aneuploidías con importancia clínica conocida incluyen el síndrome de Edward (trisomía 18) y el síndrome de Patau (trisomía 13), que resultan frecuentemente fatales en los primeros meses de vida. También se conocen anomalías asociadas con el número de cromosomas sexuales e incluyen la monosomía X, por ejemplo, el síndrome de Turner (XO), y el síndrome de triple X (XXX) en los nacimientos de mujeres y el síndrome de Kleinefelter (XXY) y el síndrome XYY en nacimientos de varones, todos los cuales están asociados con diversos fenotipos, incluidos la esterilidad y la reducción de las habilidades intelectuales. La monosomía X [45,X] es una causa común de aborto espontáneo precoz que representa aproximadamente el 7% de los abortos espontáneos. Basándose en la frecuencia en nacidos vivos de 45,X (también llamado síndrome de Turner) de 1-2/10.000, se estima que menos del 1% de conceptos 45,X llegará a término. Aproximadamente el 30% de los pacientes con síndrome de Turner son mosaicos, con una estirpe celular 45,X y una estirpe celular 46,XX o una que contiene un cromosoma X reordenado (Hook y Warburton, 1983). El fenotipo en un recién nacido vivo es relativamente leve teniendo en cuenta la alta letalidad embrionaria, y se ha planteado la hipótesis de que, posiblemente, todas las mujeres nacidas vivas con síndrome de Turner portan una estirpe celular que contiene dos cromosomas sexuales. La monosomía X puede darse en las mujeres como 45,X o como 45,X/46,XX, y en los varones como 45,X/46,XY. Generalmente se ha señalado que las monosomías autosómicas en seres humanos son incompatibles con la vida; sin embargo, existe un número considerable de informes citogenéticos que describen la monosomía completa de un cromosoma 21 en bebés nacidos vivos (VosranovaIet *et al*., Molecular Cytogen. 1:13 [2008]; Joosten *et al*., Prenatal Diagn. 17:271-5. [1997]. El método de la invención puede utilizarse para diagnosticar estas y otras anomalías cromosómicas en período prenatal.

Según algunas formas de realización, la fracción fetal puede ser útil para determinar la presencia o ausencia de trisomías cromosómicas de cualquiera de los cromosomas 1-22, X e Y. Los ejemplos de trisomías cromosómicas que pueden detectarse según el método de la presente invención incluyen, sin limitación, la trisomía 21 (T21; síndrome de Down), la trisomía 18 (T18, síndrome de Edward), la trisomía 16 (T16), la trisomía 20 (T20), la trisomía 22 (T22, síndrome del ojo de gato), la trisomía 15 (T15; síndrome de Prader-Willi), la trisomía 13 (T13, síndrome de Patau), la trisomía 8 (T8; síndrome de Warkany), la trisomía 9, y las trisomías XXY (síndrome de Kleinefelter), XYY o XXX. Las trisomías completas de los otros autosomas existentes en un estado no mosaico son letales, pero pueden ser compatibles con la vida cuando están presentes en un estado de mosaico. Se entenderá que diversas trisomías completas, existan en un estado de mosaico o no mosaico, y las trisomías parciales pueden determinarse en el cfADN fetal según las enseñanzas de la presente invención.

Los ejemplos no limitativos de trisomías parciales que pueden determinarse mediante el método de la presente invención incluyen, pero no se limitan a, trisomía parcial 1q32-44, trisomía 9p, mosaicismo de trisomía 4, trisomía 17p, trisomía parcial 4q26-qter, trisomía parcial 2p, trisomía parcial 1q, y/o trisomía parcial 6p/monosomía 6q.

Los métodos descritos en el presente documento también pueden utilizarse para ayudar a determinar una monosomía del cromosoma X, una monosomía del cromosoma 21, y monosomías parciales tales como, la monosomía 13, monosomía 15, monosomía 16, monosomía 21 y monosomía 22, que se sabe están involucradas en el aborto espontáneo. La monosomía parcial de cromosomas implicados por lo general en la aneuploidía completa también puede determinarse mediante el método de la invención. Los ejemplos no limitativos de síndromes de deleción que pueden determinarse según el método de la presente invención incluyen síndromes debidos a deleciones parciales de cromosomas. Los ejemplos de deleciones parciales que pueden determinarse según el método de la invención incluyen, sin limitación, deleciones parciales de los cromosomas 1, 4, 5, 7, 11, 18, 15, 13, 17, 22 y 10, que se describen a continuación.

El síndrome de deleción 1q21.1 o microdeleción 1q21.1 (recurrente) es una aberración poco frecuente del cromosoma 1. Junto al síndrome de deleción, también hay un síndrome de duplicación 1q21.1. Aunque hay una parte del ADN que falta con el síndrome de deleción en un sitio concreto, hay dos o tres copias de una parte similar del ADN en el mismo sitio con el síndrome de la duplicación. La literatura se refiere a la deleción y a la duplicación como las variaciones en el número de copias (CNV) 1q21.1. La deleción 1q21.1 puede estar asociada con el síndrome de TAR (trombocitopenia con aplasia de radio).

El síndrome de Wolf-Hirschhorn (WHS) (OMIN #194190) es un síndrome de deleción de genes contiguos asociado con una deleción hemicigota del cromosoma 4p16.3. El síndrome de Wolf-Hirschhorn es un síndrome de malformación congénita caracterizado por la deficiencia del crecimiento pre y postnatal, discapacidad del desarrollo de grado variable, rasgos craneofaciales característicos (aspecto de la nariz de “casco de guerrero griego”, frente alta, glabela prominente, hipertelorismo, cejas muy arqueadas, ojos saltones, pliegues epicánticos, surco nasolabial corto, boca definida con comisuras hacia abajo, y micrognatia), y un trastorno convulsivo.

La deleción parcial del cromosoma 5, también conocida como 5p- o 5p menos, y denominado síndrome del maullido del gato (OMIN#123450), se debe a una deleción del brazo corto (brazo p) del cromosoma 5 (Sp15.3-p15.2). Los bebés con esta afección tienen con frecuencia un llanto agudo que suena como el de un gato. El trastorno se caracteriza por la discapacidad intelectual y el retraso del desarrollo, cabeza pequeña (microcefalia),

bajo peso al nacer y tono muscular débil (hipotonía) en la infancia, rasgos faciales característicos y, posiblemente, defectos cardíacos.

El síndrome de Williams-Beuren, también conocido como síndrome de deleción del cromosoma 7q11.23 (OMIN 194050), es un síndrome de deleción de genes contiguos que da como resultado un trastorno multisistémico debido a la deleción hemicigota de 1,5 Mb a 1,8 Mb en el cromosoma 7q11.23, que contiene aproximadamente 28 genes.

El síndrome de Jacobsen, también conocido como trastorno de deleción 11q, es un trastorno congénito poco frecuente resultado de la deleción de una región terminal del cromosoma 11 que incluye la banda 11q24.1. Puede producir discapacidades intelectuales, un aspecto facial característico, y diversos problemas físicos, incluidos los defectos cardíacos y un trastorno de la coagulación.

La monosomía parcial del cromosoma 18, conocida como monosomía 18p, es un trastorno cromosómico poco frecuente en el que parte o la totalidad del brazo corto (p) del cromosoma 18 está delecionado (monosómico). El trastorno se caracteriza por lo general por una baja estatura, grados variables de retraso mental, retraso en el habla, malformaciones del cráneo y la región facial (craneofacial), y/u otras anomalías físicas adicionales. Los defectos craneofaciales asociados pueden variar mucho en alcance y gravedad de un caso a otro.

Las afecciones debidas a cambios en la estructura o el número de copias del cromosoma 15 incluyen el síndrome de Angelman y el síndrome de Prader-Willi, que implican una pérdida de actividad génica en la misma parte del cromosoma 15, la región 15q11-q13. Se entenderá que varias translocaciones y microdeleciones pueden ser asintomáticas en el progenitor portador, y que sin embargo, pueden generar una enfermedad genética importante en la descendencia. Por ejemplo, una madre sana portadora de la microdeleción 15q11-q13 puede dar a luz a un bebé con síndrome de Angelman, un trastorno neurodegenerativo grave. Por lo tanto, la presente invención puede utilizarse para identificar una deleción parcial de este tipo y otras deleciones en el feto.

La monosomía parcial 13q es un trastorno cromosómico poco frecuente que se produce cuando falta una parte del brazo largo (q) del cromosoma 13 (monosómico). Los bebés que nacen con monosomía parcial 13q pueden presentar bajo peso al nacer, malformaciones de la cabeza y la cara (región craneofacial), anomalías esqueléticas (especialmente de las manos y los pies), y otras anomalías físicas. El retraso mental es característico de esta afección. La tasa de mortalidad durante la infancia es alta entre los individuos que nacen con este trastorno. Casi todos los casos de monosomía parcial 13q se producen al azar, sin razón aparente (esporádicos).

El síndrome de Smith-Magenis (SMS - OMIM#182290) se debe a una deleción, o pérdida de material genético, en una copia del cromosoma 17. Este síndrome bien conocido se asocia con retraso del desarrollo, retraso mental, anomalías congénitas tales como defectos cardíacos y renales, y anomalías neuroconductuales tales como trastornos del sueño graves y conducta autolesiva. El síndrome de Smith-Magenis (SMS) se debe, en la mayoría de los casos (90%), a una deleción intersticial de 3,7 Mb en el cromosoma 17p11.2.

El síndrome de deleción 22q11.2, también conocido como síndrome de DiGeorge, es un síndrome debido a la deleción de una pequeña parte del cromosoma 22. La deleción (22q11.2) se produce cerca de la mitad del cromosoma en el brazo largo de uno del par de cromosomas. Las características de este síndrome varían mucho, incluso entre miembros de una misma familia, y afectan a muchas partes del cuerpo. Los signos y síntomas característicos pueden incluir defectos de nacimiento tales como una cardiopatía congénita, defectos en el paladar, más comúnmente relacionados con problemas neuromusculares con la oclusión (insuficiencia velofaríngea), problemas de aprendizaje, leves diferencias en los rasgos faciales, e infecciones recurrentes. Las microdeleciones en la región cromosómica 22q11.2 están asociadas con un riesgo de esquizofrenia 20 a 30 veces mayor.

Las deleciones en el brazo corto del cromosoma 10 están asociadas con un fenotipo similar al síndrome de DiGeorge. La monosomía parcial del cromosoma 10p es poco frecuente, pero se ha observado en una parte de los pacientes que muestran características del síndrome de DiGeorge.

En una forma de realización, se utiliza el método de la invención para determinar las monosomías parciales, incluidas pero no limitadas a la monosomía parcial de los cromosomas 1, 4, 5, 7, 11, 18, 15, 13, 17, 22 y 10, por ejemplo, la monosomía parcial 1q21.11, la monosomía parcial 4p16.3, la monosomía parcial 5p15,3-p15.2, la monosomía parcial 7q11.23, la monosomía parcial 11q24.1, la monosomía parcial 18p, la monosomía parcial del cromosoma 15 (15q11-q13), la monosomía parcial 13q, la monosomía parcial 17p 11.2, la monosomía parcial del cromosoma 22 (22q11.2), y la monosomía parcial 10p también pueden determinarse mediante el método.

Otras monosomías parciales que pueden determinarse según el método de la invención incluyen la translocación desequilibrada t(8;11)(p23.2;p15.5); la microdeleción 11q23; la deleción 17p11.2; la deleción 22q13.3; la microdeleción Xp22.3; la deleción 10p14; la microdeleción 20p, las deleciones [del(22)(q11.2q11.23)], 7q11.23 y 7q36; la deleción 1p36; la microdeleción 2p; la neurofibromatosis de tipo 1 (microdeleción 17q11.2), la deleción Yq; la microdeleción 4p16.3; la microdeleción 1p36.2; la deleción 11q14; la microdeleción 19q13.2; Rubinstein-Taybi (microdeleción 16p13.3); la microdeleción 7p21; el síndrome de Miller-Dicker (17p 13.3); y la microdeleción 2q37.

Las deleciones parciales pueden ser pequeñas deleciones de parte de un cromosoma, o pueden ser microdeleciones de un cromosoma en el que puede producirse la deleción de un solo gen.

Se han identificado varios síndromes de duplicación debidos a la duplicación de parte de los brazos del cromosoma (véase OMIN [Online Mendelian Inheritance in Man visto en línea en ncbi.nlm.nih.gov/omim]). En una forma de realización, puede utilizarse el método de la presente invención para determinar la presencia o ausencia de duplicaciones y/o multiplicaciones de segmentos de cualquiera de los cromosomas 1-22, X e Y. Los ejemplos no limitativos de síndromes de duplicaciones que pueden determinarse según el método de la presente invención incluyen duplicaciones de parte de los cromosomas 8, 15, 12 y 17, que se describen a continuación.

El síndrome de duplicación 8p23.1 es un trastorno genético poco frecuente debido a una duplicación de una región del cromosoma humano 8. Este síndrome de duplicación tiene una prevalencia estimada de 1 de cada 64.000 nacimientos y es el recíproco del síndrome de deleción 8p23.1. La duplicación 8p23.1 está asociada con un fenotipo variable que incluye uno o más de entre retraso en el habla, retraso del desarrollo, dismorfismo leve, con frente prominente y cejas arqueadas, y la cardiopatía congénita (CHD).

El síndrome de duplicación del cromosoma 15q (Dup15q) es un síndrome clínicamente identificable resultado de duplicaciones del cromosoma 15q11-13.1. Los bebés con Dup15q suelen tener hipotonía (escaso tono muscular), retraso del crecimiento; pueden nacer con labio leporino y/o paladar hendido o malformaciones del corazón, los riñones u otros órganos; muestran cierto grado de retraso/discapacidad cognitiva (retraso mental), retraso en el habla y el lenguaje, y trastornos de procesamiento sensorial.

El síndrome de Pallister-Killian es el resultado de material cromosómico adicional del cromosoma 12. Suele haber una mezcla de células (mosaicismo), algunas con material adicional del cromosoma 12, y algunas que son normales (46 cromosomas sin el material adicional del cromosoma 12). Los bebés con este síndrome tienen muchos problemas, incluidos retraso mental grave, escaso tono muscular, rasgos faciales “toscos”, y una frente prominente. Tienden a tener un labio superior muy fino con un labio inferior más grueso y una nariz corta. Otros problemas de salud incluyen convulsiones, falta de apetito, rigidez en las articulaciones, cataratas en la edad adulta, pérdida de audición, y defectos cardíacos. Las personas con síndrome de Pallister-Killian tienen un menor tiempo de vida.

Los individuos con la afección genética denominada dup(17)(p11.211.2) o dup 17p portan información genética adicional (conocida como duplicación) en el brazo corto del cromosoma 17. La duplicación del cromosoma 17p11.2 subyace al síndrome de Potocki-Lupski (PTLS), que es una afección genética recientemente reconocida con sólo unas pocas docenas de casos descritos en la literatura médica. Los pacientes que tienen esta duplicación tienen con frecuencia bajo tono muscular, falta de apetito, y retraso estaturo-ponderal durante la infancia, y también presentan retraso del desarrollo de los hitos motores y del lenguaje. Muchos individuos con PTLS tienen dificultades con la articulación y el procesamiento del lenguaje. Además, los pacientes pueden tener características de comportamiento similares a las observadas en personas con autismo o trastornos del espectro autista. Los individuos con PTLS pueden tener defectos cardíacos y apnea del sueño. Se sabe que una duplicación de una gran región en el cromosoma 17p12 que incluye el gen PMP22 provoca la enfermedad de Charcot-Marie-Tooth.

Las CNV se han asociado con la mortinatalidad. Sin embargo, debido a las limitaciones inherentes de la citogenética convencional, se cree que la contribución de las CNV a la mortinatalidad está poco representada (Harris *et al*., Prenatal Diagn. 31:932-944 [2011]). Los métodos de la presente invención son útiles para ayudar a determinar la presencia de aneuploidías parciales, por ejemplo, deleciones y multiplicaciones de segmentos cromosómicos, y pueden utilizarse para ayudar a identificar y determinar la presencia o ausencia de CNV que están asociadas con la mortinatalidad.

El método de la presente invención también puede ayudar a identificar polimorfismos asociados con trastornos genéticos que son complejos, multifactoriales o poligénicos, lo que significa que es probable que estén asociados con los efectos de múltiples genes en combinación con el estilo de vida y los factores ambientales. Los trastornos multifactoriales incluyen, por ejemplo, la cardiopatía y la diabetes. Aunque los trastornos complejos se agrupan con frecuencia en familias, no tienen un patrón de herencia bien definido. En una genealogía, las enfermedades poligénicas tienden a "ser hereditarias", pero la herencia no es tan simple como en las enfermedades mendelianas. Hay fuertes componentes ambientales asociados con muchos trastornos complejos, por ejemplo, la presión arterial. El método de la presente invención puede utilizarse para identificar polimorfismos que están asociados con trastornos poligénicos, incluidos pero no limitados al asma, enfermedades autoinmunitarias tales como la esclerosis múltiple, los cánceres, las ciliopatías, el paladar hendido, la diabetes, la cardiopatía, la hipertensión, la enfermedad inflamatoria intestinal, el retraso mental, los trastornos del estado de ánimo, la obesidad, el error de refracción, y la infertilidad. En algunas formas de realización, los polimorfismos son SNPs. En otras formas de realización, los polimorfismos son STRs. En otras formas de realización más, los polimorfismos son una combinación de SNPs y STRs.

En una forma de realización, la identificación de las secuencias polimórficas asociadas con los trastornos comprende la secuenciación de al menos una porción del genoma celular que corresponde al segundo genoma en la mezcla de cfADN. La identificación de las secuencias polimórficas aportadas por un primer genoma se realiza

determinando la secuencia en múltiples sitios polimórficos en una primera muestra que contiene moléculas de ADN derivadas básicamente de sólo un segundo genoma, determinando la secuencia en los correspondientes sitios polimórficos múltiples en una segunda muestra que contiene una mezcla de moléculas de ADN derivadas de un primer y un segundo genoma, y comparando las secuencias polimórficas determinadas en ambas muestras, identificando así polimorfismos múltiples en un primer genoma de una muestra que comprende una mezcla de dos genomas. Por ejemplo, la identificación de secuencias polimórficas aportadas por un genoma fetal, es decir, un primer genoma, se realiza determinando la secuencia en múltiples sitios polimórficos en una muestra de la capa leucocitaria materna, es decir, una muestra que contiene moléculas de ADN derivadas básicamente de sólo un segundo genoma, determinando la secuencia en los correspondientes sitios polimórficos múltiples en una muestra de plasma purificada, es decir, una segunda muestra que contiene una mezcla de moléculas de cfADN derivadas de los genomas materno y fetal, y comparando las secuencias polimórficas determinadas en ambas muestras para identificar polimorfismos fetales múltiples. En una forma de realización, el primer genoma es un genoma fetal, y un segundo genoma es un genoma materno. En otra forma de realización, el primer genoma es un genoma de una célula no afectada y el segundo genoma es un genoma de una célula afectada. En algunas formas de realización, las células afectadas y no afectadas proceden del mismo sujeto. Por ejemplo, la célula afectada puede ser una célula cuyo genoma ha sido modificado por un trastorno.

En una forma de realización, los métodos de estimación de la fracción genómica descritos ayudan a detectar un cáncer en un paciente. En diversos ejemplos, un cáncer se detecta mediante un método que comprende: proporcionar una muestra de un paciente que comprende una mezcla de genomas derivados de células normales, es decir, no afectadas, y de células cancerosas, es decir, afectadas; e identificar los polimorfismos múltiples asociados con el cáncer. En algunas formas de realización, la muestra se selecciona a partir de sangre, plasma, suero y orina. En algunas formas de realización, la muestra es una muestra de plasma. En otras formas de realización, la muestra es una muestra de orina.

En una forma de realización, la identificación de polimorfismos múltiples asociados con el cáncer comprende enriquecer el ADN de la muestra en secuencias diana polimórficas. En otras formas de realización, no se realiza el enriquecimiento de la muestra en secuencias diana polimórficas. En algunas formas de realización, la identificación de polimorfismos múltiples asociados con el cáncer comprende cuantificar el número de copias de la secuencia polimórfica.

Los cánceres que pueden identificarse y/o vigilarse según el método de la descripción incluyen tumores sólidos, así como tumores hematológicos y/o tumores malignos. Los diversos tipos de cáncer a tratar incluyen sarcomas, carcinomas y adenocarcinomas no limitados al cáncer de mama, cáncer de pulmón, cáncer colorrectal, cáncer de páncreas, cáncer de ovario, cáncer de próstata, carcinoma renal, hepatoma, cáncer cerebral, melanoma, mieloma múltiple, linfoma, linfoma de Hodgkin, linfoma no Hodgkin, linfomas infantiles, y linfomas de linfocitos y de origen cutáneo, leucemia, leucemia infantil, leucemia de células pilosas, leucemia linfocítica aguda, leucemia mieloide aguda, leucemia linfocítica crónica, leucemia mieloide crónica, leucemia mielógena crónica, y leucemia de mastocitos, neoplasias mieloides, neoplasias de mastocitos, tumor hematológico y tumor linfoide, incluidas las lesiones metastásicas en otros tejidos u órganos distantes del sitio del tumor primario.

Los métodos de la presente invención son útiles, por ejemplo, para diagnosticar o determinar un pronóstico de un estado patológico que se sabe está asociado con un(os) haplotipo(s) específico(s), para determinar nuevos haplotipos, y para detectar asociaciones de haplotipos con capacidad de respuesta a productos farmacéuticos. La asociación de múltiples secuencias polimórficas con múltiples trastornos puede determinarse a partir de la identidad de una sola secuencia polimórfica para cada uno de los múltiples trastornos. Como alternativa, la asociación de múltiples secuencias polimórficas con múltiples trastornos puede determinarse a partir de la identidad de múltiples secuencias polimórficas para cada uno de los múltiples trastornos.

Las técnicas convencionales de genotipado se han limitado a la identificación de polimorfismos en regiones genómicas cortas de unas pocas kilobases, y la identificación de haplotipos se ha basado en los datos familiares y la estimación estadística utilizando algoritmos computacionales. La secuenciación del genoma completo posibilita la identificación de haplotipos identificando directamente los polimorfismos en un genoma. La identificación de los haplotipos según las diversas formas de realización no está limitada por la distancia que media entre los polimorfismos. En algunas formas de realización, un método comprende la secuenciación del genoma completo del ADN celular materno. El ADN celular materno puede obtenerse de una muestra biológica desprovista de ADN genómico fetal. Por ejemplo, el ADN materno puede obtenerse de la capa leucocitaria de la sangre materna. Pueden determinarse los haplotipos que comprenden una pluralidad de secuencias polimórficas que abarcan la totalidad de los cromosomas. En una forma de realización, los haplotipos fetales se comparan con los haplotipos asociados a trastornos conocidos, y sobre la base de una coincidencia del haplotipo fetal con cualquiera de los haplotipos asociados a trastornos conocidos, indica que el feto tiene el trastorno o que el feto presenta predisposición al trastorno. Los haplotipos fetales también pueden compararse con los haplotipos asociados con una capacidad de respuesta o falta de respuesta al tratamiento del polimorfismo específico. La comparación de los haplotipos fetales identificados con bases de datos de haplotipos conocidos permite el diagnóstico y/o pronóstico de un trastorno. Puede utilizarse cualquier muestra biológica que comprenda una mezcla de cfADN fetal y materno para determinar la presencia o ausencia del trastorno fetal. Preferentemente, la muestra biológica se selecciona de entre sangre, o

fracciones de la misma, incluido el plasma, u orina. En una forma de realización, la muestra biológica es una muestra de sangre. En otra forma de realización, la muestra biológica es una muestra de plasma. En otra forma de realización más, la muestra biológica es una muestra de orina.

En una forma de realización, la invención proporciona un método para determinar la presencia o ausencia de trastornos fetales múltiples, que comprende (a) obtener una muestra de sangre materna que comprende una mezcla de ADN fetal y materno libre, (b) secuenciar el genoma completo de al menos una porción de la mezcla de ADN fetal y materno libre, obteniendo así una pluralidad de marcadores de secuencia; (c) determinar los polimorfismos fetales múltiples en los marcadores de secuencia, y (d) determinar la presencia o ausencia de trastornos fetales múltiples. Los ejemplos de trastornos fetales múltiples que pueden identificarse según el método de la presente invención incluyen los trastornos monogénicos y poligénicos descritos en el presente documento.

En una forma de realización, la invención proporciona un método para determinar la presencia o ausencia de trastornos fetales múltiples que comprende identificar los polimorfismos fetales múltiples asociados con los haplotipos relacionados con trastornos múltiples. En algunas formas de realización, cada uno de los haplotipos comprende al menos dos, al menos tres, al menos cuatro, al menos cinco, al menos diez o al menos quince polimorfismos marcadores diferentes. Los polimorfismos marcadores presentes en el haplotipo pueden ser del mismo tipo de polimorfismo, por ejemplo, todos ellos polimorfismos SNP marcadores, o pueden ser una combinación de polimorfismos, por ejemplo, SNPs marcadores y deleciones marcadoras. En una forma de realización, los polimorfismos son SNPs marcadores. En otra forma de realización, los polimorfismos son STRs marcadoras. En otra forma de realización más, los polimorfismos son una combinación de SNPs marcadores y STRs marcadoras. Los polimorfismos marcadores pueden estar en las regiones codificantes y/o no codificantes del genoma. La identificación de los polimorfismos se realiza mediante secuenciación del genoma completo utilizando tecnologías de NGS como se describe en el presente documento.

La invención proporciona un método para identificar las variaciones en el número de copias (CNV) como polimorfismos de una secuencia de interés en una muestra de ensayo que comprende una mezcla de ácidos nucleicos derivados de dos genomas diferentes, y que se sabe o se sospecha difieren en la cantidad de una o más secuencias de interés. Las variaciones en el número de copias determinadas mediante el método de la invención incluyen ganancias o pérdidas de cromosomas enteros, modificaciones que implican segmentos cromosómicos muy grandes que son visibles al microscopio, y una gran cantidad de variación submicroscópica en el número de copias de segmentos de ADN con un tamaño que va de las kilobases (kb) a las megabases (Mb).

La CNV en el genoma humano influye significativamente en la diversidad humana y la predisposición a la enfermedad (Redon *et al*., Nature. 23:444-454 [2006], Shaikh *et al*., Genome Res. 19:1682-1690 [2009]). Se sabe que las CNVs contribuyen a la enfermedad genética a través de diferentes mecanismos, lo que da como resultado el desequilibrio de la dosis génica o la interrupción génica en la mayoría de los casos. Además de su correlación directa con los trastornos genéticos, se sabe que las CNVs intervienen en cambios fenotípicos que pueden ser perjudiciales. Recientemente, varios estudios han informado acerca de un aumento de la carga de CNVs *de novo* o poco frecuentes en trastornos complejos tales como el autismo, el TDAH, y la esquizofrenia en comparación con los controles normales, destacando la posible patogenicidad de CNVs únicas o poco frecuentes (Sebat *et al*., 316:445-449 [2007]; Walsh *et al*., Science 320:539-543 [2008]). Las CNV surgen de reordenamientos genómicos, debidos principalmente a eventos de deleción, duplicación, inserción y translocación desequilibrada.

Las formas de realización de la invención proporcionan un método para evaluar la variación en el número de copias de una secuencia de interés, por ejemplo, una secuencia clínicamente pertinente, en una muestra de ensayo que comprende una mezcla de ácidos nucleicos derivados de dos genomas diferentes, y que se sabe o se sospecha difieren en la cantidad de una o más secuencias de interés. La mezcla de ácidos nucleicos se deriva de dos o más tipos de células. En una forma de realización, la mezcla de ácidos nucleicos se deriva de células normales y cancerosas procedentes de un sujeto que padece una enfermedad, por ejemplo, un cáncer.

Se cree que muchos tumores sólidos, tal como el cáncer de mama, evolucionan desde el inicio hasta la metástasis mediante la acumulación de varias aberraciones genéticas. [Sato *et al*., Cancer Res., 50:7184-7189 [1990]; Jongsma *et al*., J Clin. Pathol.: Mol. Path. 55:305-309 [2002])]. Tales aberraciones genéticas, a medida que se acumulan, pueden conferir ventajas proliferativas, inestabilidad genética y la consiguiente capacidad de desarrollar rápidamente farmacorresistencia, y potenciación de la angiogénesis, proteólisis y metástasis. Las aberraciones genéticas pueden afectar tanto a "genes supresores de tumores" recesivos como a oncogenes que actúan de manera dominante. Se cree que las deleciones y la recombinación que conducen a la pérdida de heterocigosidad (LOH) desempeñan un papel importante en la progresión tumoral al dejar al descubierto alelos supresores de tumores mutados.

Se ha encontrado cfADN en la circulación de pacientes con diagnóstico de tumores malignos, incluidos pero no limitados al cáncer de pulmón (Pathak *et al*., Clin. Chem. 52:1833-1842 [2006]), el cáncer de próstata (Schwartzenbach *et al*., Clin. Cancer Res. 15:1032-8 [2009]), y el cáncer de mama (Schwartzenbach *et al*., disponible en línea en breast-cancer-research.com/content/11/5/R71 [2009]). La identificación de inestabilidades genómicas asociadas con cánceres que pueden determinarse en el cfADN circulante en pacientes con cáncer es

una posible herramienta de diagnóstico y de pronóstico. En una forma de realización, el método de la invención evalúa la CNV de una secuencia de interés en una muestra que comprende una mezcla de ácidos nucleicos procedente de un sujeto que se sospecha o se sabe tiene cáncer, por ejemplo, un carcinoma, sarcoma, linfoma, leucemia, tumores de células germinales y un blastoma. En una forma de realización, la muestra es una muestra de plasma derivada (procesada) de la sangre periférica y que comprende una mezcla de cfADN derivado de células normales y cancerosas. En otra forma de realización, la muestra biológica que se necesita para determinar si una CNV está presente se deriva de una mezcla de células cancerosas y no cancerosas de otros fluidos biológicos, incluidos pero no limitados a suero, sudor, lágrimas, esputo, orina, secreción del oído, linfa, saliva, líquido cefalorraquídeo, lavados, suspensión de médula ósea, flujo vaginal, lavado transcervical, líquido cerebral, líquido ascítico, leche, secreciones de los tractos respiratorio, intestinal y genitourinario, y muestras de leucoforesis, o en biopsias de tejido, hisopos o frotis.

La secuencia de interés es una secuencia de ácido nucleico que se sabe o se sospecha es importante para el desarrollo y/o la evolución del cáncer. Los ejemplos de una secuencia de interés incluyen secuencias de ácidos nucleicos que están amplificadas o delecionadas en las células cancerosas como se describe a continuación.

Los genes que actúan de manera dominante asociados con tumores sólidos humanos ejercen por lo general su efecto mediante la sobreexpresión o la expresión modificada. La amplificación génica es un mecanismo común que conduce a la regulación positiva de la expresión génica. Hay estudios citogenéticos que indican que se produce una amplificación significativa en más del 50% de los cánceres de mama humanos. Muy en particular, la amplificación del protooncogén receptor del factor de crecimiento epidérmico 2 (HER2) humano que se encuentra en el cromosoma 17 (17(17q21-q22)), da como resultado la sobreexpresión de los receptores HER2 en la superficie celular que conduce a una señalización excesiva y mal regulada en el cáncer de mama y otros tumores malignos (Park *et al*., Clinical Breast Cancer 8:392-401 [2008]). Se ha descubierto que diversos oncogenes están amplificados en otros tumores malignos humanos. Los ejemplos de la amplificación de oncogenes celulares en tumores humanos incluyen amplificaciones de: c-myc en la estirpe celular de leucemia promielocítica HL60, y estirpes celulares de carcinoma pulmonar de células pequeñas, N-myc en neuroblastomas primarios (fases III y IV), estirpes celulares de neuroblastoma, tumores primarios y estirpe celular de retinoblastoma, y tumores y estirpes celulares de carcinoma pulmonar de células pequeñas, L-myc en tumores y estirpes celulares de carcinoma pulmonar de células pequeñas, c-myb en la leucemia mieloide aguda y en estirpes celulares de carcinoma de colon, c-erbb en estirpes celulares de carcinoma epidermoide, y gliomas primarios, c-K-ras-2 en carcinomas primarios de pulmón, colon, vejiga y recto, N-ras en la estirpe celular de carcinoma de mama (Varmus H., Ann. Rev. Genetics 18:553-612 (1984) [citado en Watson *et al*., Molecular Biology of the Gene (4ª ed.; Benjamin/Cummings Publishing Co. 1987)].

Las deleciones cromosómicas que implican genes supresores de tumores pueden desempeñar un papel importante en el desarrollo y la progresión de los tumores sólidos. El gen supresor del tumor retinoblastoma (Rb-1), que se encuentra en el cromosoma 13q14, es el gen supresor de tumores más ampliamente caracterizado. El producto del gen Rb-1, una fosfoproteína nuclear de 105 kDa, desempeña al parecer un papel importante en la regulación del ciclo celular (Howe *et al*., Proc. Natl. Acad. Sci. (EE.UU.) 87:5883-5887 [1990]). La expresión modificada o perdida de la proteína Rb se debe a la inactivación de ambos alelos del gen, ya sea por una mutación puntual o una deleción cromosómica. Se ha descubierto que hay modificaciones del gen Rb-1 no sólo en los retinoblastomas, sino también en otros tumores malignos tales como los osteosarcomas, el cáncer de pulmón de células pequeñas (Rygaard *et al*., Cancer Res. 50: 5312-5317 [1990)]) y el cáncer de mama. Los estudios de polimorfismo en la longitud de los fragmentos de restricción (RFLP) han indicado que tales tipos de tumores han perdido con frecuencia heterocigosidad en 13q, lo que sugiere que uno de los alelos del gen Rb-1 se ha perdido debido a una gran deleción cromosómica (Bowcock *et al*., Am. J. Hum. Genet., 46: 12 [1990]). Las anomalías en el cromosoma 1, incluidas las duplicaciones, deleciones y translocaciones desequilibradas que implican al cromosoma 6 y otros cromosomas compañeros ("partner") indican que las regiones del cromosoma 1, en particular 1q21-1q32 y 1p11-13, podrían albergar oncogenes o genes supresores de tumores que son patogenéticamente pertinentes para las fases crónica y avanzada de las neoplasias mieloproliferativas (Caramazza *et al*., Eur. J. Hemato. 184:191-200 [2010]). Las neoplasias mieloproliferativas también están asociadas con deleciones del cromosoma 5. La pérdida completa o las deleciones intersticiales del cromosoma 5 son la anomalía del cariotipo más común en los síndromes mielodisplásicos (MDSs). Los pacientes con MDS con del(5q) aislada/5q- tienen un pronóstico más favorable que aquellos con defectos del cariotipo adicionales, que tienden a desarrollar neoplasias mieloproliferativas (MPNs) y leucemia mieloide aguda. La frecuencia de deleciones del cromosoma 5 desequilibradas ha llevado a la idea de que 5q alberga uno o más genes supresores de tumores que tienen funciones fundamentales en el control del crecimiento de las células madre/progenitoras hematopoyéticas (HSCs/HPC). El mapeo citogenético de las regiones comúnmente delecionadas (CDRs) centrado en 5q31 y 5q32 identificó genes supresores de tumores candidatos, incluida la subunidad ribosómica RPS14, el factor de transcripción Egr1/Krox20 y la proteína remodeladora del citoesqueleto, alfa-catenina (Eisenmann *et al*., Oncogene 28:3429-3441 [2009]). Los estudios citogenéticos y de alelotipado de tumores frescos y estirpes celulares de tumor han demostrado que la pérdida alélica de varias regiones distintas en el cromosoma 3p, incluidas 3p25, 3p21-22, 3p21.3, 3p12-13 y 3p14, son las anomalías genómicas más precoces y frecuentes implicadas en un amplio espectro de los principales cánceres epiteliales de pulmón, mama, riñón, cabeza y cuello, ovario, cuello del útero, colon, páncreas, esófago, vejiga y otros órganos. Se han mapeado varios genes supresores de tumores en la región del cromosoma 3p, y se cree que las deleciones

intersticiales o la hipermetilación del promotor preceden a la pérdida del 3p o de todo el cromosoma 3 en el desarrollo de carcinomas (Angeloni D., Briefings Functional Genomics 6:19-39 [2007]).

Los recién nacidos y los/las niños/as con síndrome de Down (DS) presentan con frecuencia leucemia transitoria congénita y tienen un mayor riesgo de leucemia mieloide aguda y leucemia linfoblástica aguda. El cromosoma 21, que alberga aproximadamente 300 genes, puede estar implicado en numerosas aberraciones estructurales, por ejemplo, translocaciones, deleciones y amplificaciones, en leucemias, linfomas y tumores sólidos. Además, se han identificado genes que se encuentran en el cromosoma 21 que desempeñan un papel importante en la tumorigénesis. Las aberraciones del cromosoma 21 numéricas somáticas, así como las estructurales, están asociadas con las leucemias, y son importantes en la tumorigénesis genes específicos, incluidos RUNX1, TMPRSS2 y TFF, que se encuentran en 21q (Fonatsch C Gene Chromosomes Cancer 49:497-508 [2010]).

En una forma de realización, la descripción proporciona un medio para evaluar la asociación entre la amplificación del gen y el grado de evolución del tumor. La correlación entre la amplificación y/o deleción y la fase o grado de un cáncer puede ser importante para el pronóstico ya que tal información puede contribuir a definir el grado de diferenciación de un tumor basado en la genética que podría predecir mejor el futuro curso de la enfermedad, teniendo los tumores más avanzados el peor pronóstico. Además, la información sobre eventos de amplificación y/o deleción tempranos puede ser útil a la hora de asociar estos eventos como factores predictivos de la progresión de la enfermedad subsiguiente. Las deleciones y la amplificación de genes identificadas mediante el método pueden asociarse con otros parámetros conocidos tales como el grado de diferenciación del tumor, la histología, el índice de marcado con Brd/Urd, el estado hormonal, la afectación ganglionar, el tamaño del tumor, la duración de la supervivencia y otras propiedades tumorales disponibles a partir de estudios epidemiológicos y bioestadísticos. Por ejemplo, el ADN tumoral a ensayar mediante el método podría incluir la hiperplasia atípica, el carcinoma ductal *in situ*, el cáncer en fase I-III y ganglios linfáticos metastásicos con el fin de permitir la identificación de asociaciones entre las amplificaciones y deleciones y la fase. Las asociaciones realizadas pueden hacer posible una intervención terapéutica eficaz. Por ejemplo, las regiones sistemáticamente amplificadas pueden contener un gen sobreexpresado, cuyo producto puede combatirse terapéuticamente (por ejemplo, tirosina quinasa del receptor del factor de crecimiento, $p185^{HER2}$).

El método puede utilizarse para identificar eventos de amplificación y/o deleción que están asociados con la farmacorresistencia, determinando la variación en el número de copias de los ácidos nucleicos a partir de cánceres primarios frente a la de las células que han metastatizado a otros sitios. Si la amplificación y/o deleción de genes es una manifestación de la inestabilidad cariotípica que permite un rápido desarrollo de la farmacorresistencia, se esperaría una mayor amplificación y/o deleción en los tumores primarios de pacientes quimiorresistentes que en los tumores de pacientes sensibles a la quimioterapia. Por ejemplo, si la amplificación de genes específicos es responsable del desarrollo de farmacorresistencia, se esperaría que las regiones que rodean a esos genes estuvieran sistemáticamente amplificadas en las células tumorales de derrames pleurales de pacientes quimiorresistentes pero no en los tumores primarios. El descubrimiento de las asociaciones entre la amplificación y/o deleción de genes y el desarrollo de farmacorresistencia puede permitir identificar a los pacientes que se beneficiarán o no de un tratamiento complementario.

En otras formas de realización, el método de la presente invención puede utilizarse para identificar polimorfismos asociados con trastornos por repetición de trinucleótidos, que son un conjunto de trastornos genéticos debidos a la expansión de repeticiones de trinucleótidos. Las expansiones de trinucleótidos son un subconjunto de repeticiones de microsatélites inestables que se producen a lo largo de todas las secuencias genómicas. Si la repetición está presente en un gen sano, una mutación dinámica puede aumentar el número de repeticiones y dar como resultado un gen defectuoso. En una forma de realización, el método puede utilizarse para identificar repeticiones de trinucleótidos asociadas con el síndrome de X frágil. El brazo largo del cromosoma X de los pacientes que padecen el síndrome de X frágil puede contener de 230 a 4.000 CGG, en comparación con las 60 a 230 repeticiones en los portadores y las 5 a 54 repeticiones en individuos no afectados. La inestabilidad cromosómica resultado de esta expansión de trinucleótidos se presenta clínicamente como retraso mental, rasgos faciales característicos y macroorquidismo en los hombres. La segunda enfermedad de repetición de tripletes de ADN relacionada, el síndrome de X frágil-E, también se identificó en el cromosoma X, pero se descubrió que era el resultado de una repetición expandida de CCG. El método de la presente invención puede identificar repeticiones de trinucleótidos asociadas con otros trastornos por expansión de repeticiones, incluidas las categorías I, II y III. Los trastornos de la categoría I incluyen la enfermedad de Huntington (HD) y las ataxias espinocerebelosas que se deben a una expansión de repeticiones CAG en porciones codificantes de proteínas de genes específicos. Las expansiones de la categoría II tienden a ser más diversas fenotípicamente, con expansiones heterogéneas que son generalmente de pequeña magnitud, pero que también se encuentran en los exones de los genes. La categoría III incluye el síndrome de X frágil, la distrofia miotónica, dos de las ataxias espinocerebelosas, la epilepsia mioclónica juvenil y la ataxia de Friereich. Estas enfermedades se caracterizan por expansiones de repeticiones por lo general mucho mayores que los dos primeros grupos, y las repeticiones se encuentran fuera de las regiones codificantes de proteínas de los genes.

En otras formas de realización, el método de la presente invención puede identificar repeticiones del trinucleótido CAG asociadas con al menos diez trastornos neurológicos que se sabe se deben a un aumento del

número de repeticiones de CAG, por lo general en regiones codificantes de proteínas por lo demás no relacionadas. Durante la síntesis de proteínas, las repeticiones expandidas de CAG se traducen en una serie de residuos de glutamina ininterrumpidos que forman lo que se conoce como tramos de poliglutamina ("polyQ"). Tales tramos de poliglutamina pueden estar sujetos a una mayor agregación. Estos trastornos se caracterizan por un modo de herencia autosómica dominante (a excepción de la atrofia muscular espinobulbar que muestra una herencia ligada al cromosoma X), el inicio de la madurez, un curso progresivo, y una correlación del número de repeticiones de CAG con la gravedad de la enfermedad y la edad de inicio. Los genes causales se expresan ampliamente en todas las enfermedades por poliglutamina conocidas. Un síntoma común de las enfermedades polyQ se caracteriza por una degeneración progresiva de las células nerviosas que por lo general afecta a personas más adelante en la vida. Aunque estas enfermedades comparten el mismo codón repetido (CAG) y algunos de los síntomas, las repeticiones para las diferentes enfermedades por poliglutamina se producen en diferentes cromosomas. Los ejemplos de trastornos polyQ que pueden identificarse mediante el método de la presente invención incluyen, sin limitación, la DRPLA (atrofia dentatorubro palidoluisiana), la HD (enfermedad de Huntington), la SBMA (atrofia muscular espinobulbar o enfermedad de Kennedy), la SCA1 (ataxia espinocerebelosa tipo 1), la SCA2 (ataxia espinocerebelosa tipo 2), la SCA3 (ataxia espinocerebelosa tipo 3 o enfermedad de Machado-Joseph), la SCA6 (ataxia espinocerebelosa tipo 6), la SCA7 (ataxia espinocerebelosa tipo 7), la SCA17 (ataxia espinocerebelosa tipo 17). Los ejemplos de trastornos no polyQ que pueden identificarse mediante el método de la presente invención incluyen el FRAXA (síndrome de X frágil), el FXTAS (síndrome de temblor/ataxia asociado a X frágil), el FRAXE (retraso mental por X frágil tipo E), la FRDA (ataxia de Friedreich), la DM (distrofia miotónica), la SCA8 (ataxia espinocerebelosa tipo 8), la SCA12 (ataxia espinocerebelosa tipo 12).

Además de la función de la CNV en el cáncer, las CNVs se han asociado con un número creciente de enfermedades complejas comunes, incluido el virus de la inmunodeficiencia humana (VIH), enfermedades autoinmunitarias y un espectro de trastornos neuropsiquiátricos.

Hasta la fecha, varios estudios han informado acerca de la asociación entre la CNV en los genes implicados en la inflamación y la respuesta inmunitaria y el VIH, el asma, la enfermedad de Crohn y otros trastornos autoinmunitarios (Fanciulli *et al*., Clin. Genet. 77:201-213 [2010]). Por ejemplo, se ha atribuido a la CNV en *CCL3L1* la susceptibilidad al VIH/SIDA (*CCL3L1*, deleción 17q11.2), la artritis reumatoide (*CCL3L1*, deleción 17q11.2), y la enfermedad de Kawasaki (*CCL3L1*, duplicación 17q11.2); se ha informado que la CNV en *HBD-2* predispone a la enfermedad de Crohn colónica (*HDB-2*, deleción 8p23.1) y a la psoriasis (*HDB-2*, deleción 8p23.1); se demostró que la CNV en *FCGR3B* predisponía a la glomerulonefritis en el lupus eritematoso sistémico (*FCGR3B*, deleción 1q23, duplicación 1q23), la vasculitis asociada a anticuerpos anticitoplasma de neutrófilos (ANCA) (*FCGR3B*, deleción 1q23), y el aumento del riesgo de desarrollar artritis reumatoide. Hay al menos dos enfermedades inflamatorias o autoinmunitarias que han demostrado estar asociadas con la CNV en diferentes loci de los genes. Por ejemplo, la enfermedad de Crohn se asocia con un bajo número de copias en *HDB-2*, pero también con un polimorfismo por deleción común aguas arriba del gen *IGRM* que codifica un miembro de la familia de GTPasas relacionada con la inmunidad p47. Además de la asociación con el número de copias de *FCGR3B*, también se ha informado que la susceptibilidad al SLE está significativamente aumentada entre los sujetos con un menor número de copias del componente del complemento C4.

Se ha informado acerca de asociaciones entre las deleciones genómicas en los loci *GSTM1* (*GSTM1*, deleción 1q23) y *GSTT1* (*GSTT1*, deleción 22q11.2) y un mayor riesgo de asma atópica en varios estudios independientes. En algunas formas de realización, el método de la presente invención puede utilizarse para determinar la presencia o ausencia de una CNV asociada con la inflamación y/o enfermedades autoinmunitarias. Por ejemplo, el método de la presente invención puede utilizarse para determinar la presencia de una CNV en un paciente que se sospecha padece VIH, asma, o enfermedad de Crohn. Los ejemplos de CNV asociada con tales enfermedades incluyen, sin limitación, deleciones en 17q11.2, 8p23.1, 1q23 y 22q11.2, y duplicaciones en 17q11.2 y 1q23. En algunas formas de realización, el método de la presente invención puede utilizarse para determinar la presencia de CNV en los genes, incluidos pero no limitados a *CCL3L1*, *HBD-2*, *FCGR3B*, *GSTM*, *GSTT1*, *C4* s *IRGM*.

Se ha informado acerca de asociaciones entre CNV *de novo* y heredadas, y varias enfermedades neurológicas y psiquiátricas comunes en el autismo, la esquizofrenia y la epilepsia, y algunos casos de enfermedades neurodegenerativas tales como la enfermedad de Parkinson, la esclerosis lateral amiotrófica (ALS) y la enfermedad de Alzheimer autosómica dominante (Fanciulli *et al*., Clin. Genet. 77:201-213 [2010]). Se han observado anomalías citogenéticas en pacientes con autismo y trastornos del espectro autista (ASD) con duplicaciones en 15q11-q13. Según el Autism Genome Project Consortium, 154 CNV incluidas varias CNV recurrentes, ya sea en el cromosoma 15q11-q13 o en nuevas localizaciones genómicas que incluyen el cromosoma 2p16, 1q21 y en 17p12 en una región asociada con el síndrome de Smith-Magenis que se solapa con los ASD. Las microdeleciones o microduplicaciones recurrentes en el cromosoma 16p 11.2 han destacado la observación de que las CNVs *de novo* se detectan en los loci de genes tales como *SHANK3* (deleción 22q13.3), neurexina 1 (*NRXN1*, deleción 2p16.3) y las neuroglinas (*NLGN4*, deleción Xp22.33) que se sabe regulan la diferenciación sináptica y regulan la liberación del neurotransmisor glutaminérgico. La esquizofrenia también se ha asociado con múltiples CNVs *de novo*. Las microdeleciones y microduplicaciones asociadas con la esquizofrenia contienen una sobrerrepresentación de genes que pertenecen a las vías del desarrollo neurológico y glutaminérgica, lo que sugiere

que múltiples CNVs que afectan a estos genes pueden contribuir directamente a la patogénesis de la esquizofrenia, por ejemplo, *ERBB4*, deleción 2q34, *SLC1A3*, deleción 5p13.3; *RAPEGF4*, deleción 2q31.1; *CIT*, deleción 12.24; y múltiples genes con CNV *de novo*. Las CNVs también se han asociado con otros trastornos neurológicos incluidas la epilepsia (*CHRNA7*, deleción 15q13.3), la enfermedad de Parkinson (*SNCA*, duplicación 4q22) y la ALS (*SMN1*, deleción 5q12.2.-q13.3; y deleción *SMN2*). En algunas formas de realización, el método de la presente invención puede utilizarse para determinar la presencia o ausencia de una CNV asociada con enfermedades del sistema nervioso. Por ejemplo, el método de la presente invención puede utilizarse para determinar la presencia de una CNV en un paciente que se sospecha padece autismo, esquizofrenia, epilepsia, enfermedades neurodegenerativas tales como la enfermedad de Parkinson, la esclerosis lateral amiotrófica (ALS) o la enfermedad de Alzheimer autosómica dominante. El método de la presente invención puede utilizarse para determinar la CNV de genes asociados con enfermedades del sistema nervioso, incluidas sin limitación, cualquiera de entre trastornos del espectro autista (ASD), esquizofrenia y epilepsia, y la CNV de genes asociados con trastornos neurodegenerativos tal como la enfermedad de Parkinson. Los ejemplos de CNV asociada con tales enfermedades incluyen, sin limitación, duplicaciones en 15q11-q13, 2p16, 1q21, 17p12, 16pa11.2 y 4q22, y deleciones en 22q13.3, 2p16.3, Xp22.33, 2q34, 5p13.3, 2q31.1, 12,24, 15q13.3 y 5q12.2. En algunas formas de realización, el método de la presente invención puede utilizarse para determinar la presencia de CNV en genes que incluyen, pero no se limitan a, *SHANK3*, *NLGN4*, *NRXN1*, *ERBB4*, *SLC1A3*, *RAPGEF4*, *CIT*, *CHRNA7*, *SNCA*, *SMN1* y *SMN2*.

Se ha informado acerca de la asociación entre rasgos metabólicos y cardiovasculares, tales como la hipercolesterolemia familiar (FH), la aterosclerosis y la enfermedad arterial coronaria, y las CNVs en varios estudios (Fanciulli *et al*., Clin. Genet. 77:201-213 [2010]). Por ejemplo, se han observado reordenamientos de la línea germinal, principalmente deleciones, en el gen *LDLR* (*LDLR*, deleción/duplicación 19p13.2) en algunos pacientes con FH que no portan ninguna otra mutación *LDLR*. Otro ejemplo es el gen *LPA* que codifica la apolipoproteína (a) (apo(a)) cuya concentración plasmática se asocia con el riesgo de enfermedad arterial coronaria, infarto de miocardio (MI) e ictus. Las concentraciones plasmáticas de lipoproteína Lp(a) que contiene la apo(a) pueden variar en más de 1.000 veces entre individuos y el 90% de esta variabilidad está determinada genéticamente en el locus *LPA*, siendo proporcionales la concentración plasmática y el tamaño de la isoforma de Lp(a) a un número muy variable de secuencias de repetición “kringle 4” (intervalo de 5-50). Estos datos indican que la CNV en al menos dos genes puede estar asociada con riesgo cardiovascular. El método de la presente invención puede utilizarse en estudios grandes para buscar específicamente las asociaciones de la CNV con trastornos cardiovasculares. En algunas formas de realización, el método de la presente invención puede utilizarse para determinar la presencia o ausencia de una CNV asociada con una enfermedad metabólica o cardiovascular. Por ejemplo, el método de la presente invención puede utilizarse para determinar la presencia de una CNV en un paciente que se sospecha padece hipercolesterolemia familiar. El método de la presente invención puede utilizarse para determinar la CNV de los genes asociados con una enfermedad metabólica o cardiovascular, por ejemplo, la hipercolesterolemia. Los ejemplos de CNV asociada con tales enfermedades incluyen, sin limitación, la deleción/duplicación 19p13.2 del gen LDLR, y multiplicaciones en el gen LPA.

SECUENCIACIÓN

En diversas formas de realización, el método descrito en el presente documento emplea la tecnología de secuenciación de nueva generación (NGS) en la que se secuencian moléculas de ADN individuales o moldes de ADN amplificados por clonación de forma masiva en paralelo dentro de una célula de flujo (por ejemplo, como se describe en Volkerding *et al*., Clin. Chem. 55:641-658 [2009]; Metzker M., Nature Rev. 11:31-46 [2010]). Además de información de secuencias de alto rendimiento, la NGS proporciona información digital cuantitativa, en el sentido de que cada lectura de secuencia es un "marcador de secuencia" contable que representa un molde de ADN clonal individual o una molécula de ADN individual. Las tecnologías de secuenciación de NGS incluyen pirosecuenciación, secuenciación por síntesis con terminadores con colorante reversibles, secuenciación por ligación de sondas oligonucleotídicas y secuenciación en tiempo real.

En diversas formas de realización, pueden analizarse las muestras que no están amplificadas, o que están amplificadas sólo parcialmente (amplificación de la diana). En algunos casos, los métodos de determinación de la fracción fetal pueden lograrse sin necesidad de ningún tipo de amplificación selectiva.

La amplificación del genoma completo que se produce como parte del proceso de secuenciación proporciona suficientes copias que pueden cubrirse aumentando el número de ciclos de secuenciación para proporcionar una cobertura cada vez mejor.

En formas de realización preferentes, la muestra que comprende la mezcla de moléculas de ADN derivadas de dos genomas diferentes está enriquecida no específicamente en las secuencias del genoma completo antes de la secuenciación del genoma completo, es decir, la amplificación del genoma completo se realiza antes de la secuenciación.

Enriquecimiento no específico del ADN de la muestra puede referirse a la amplificación del genoma completo de los fragmentos de ADN genómico de la muestra que pueden utilizarse para aumentar el nivel del ADN de la muestra antes de identificar los polimorfismos mediante la secuenciación. El enriquecimiento no específico

puede ser un enriquecimiento selectivo de uno de los dos genomas presentes en la muestra. Por ejemplo, el enriquecimiento no específico puede ser selectivo del genoma fetal en una muestra materna, que puede obtenerse mediante métodos conocidos para aumentar la proporción relativa de ADN fetal respecto al materno en una muestra. Como alternativa, el enriquecimiento no específico puede ser la amplificación no selectiva de ambos genomas presentes en la muestra. Por ejemplo, la amplificación no específica puede ser del ADN fetal y materno en una muestra que comprende una mezcla de ADN de los genomas fetal y materno. En la técnica se conocen métodos para amplificar el genoma completo. La PCR con oligonucleótidos degenerados (DOP), la técnica de PCR con extensión de cebador (PEP) y la amplificación por desplazamiento múltiple (MDA), son ejemplos de métodos de amplificación del genoma completo. En algunas formas de realización, la muestra que comprende la mezcla de cfADN de diferentes genomas no está enriquecida en el cfADN de los genomas presentes en la mezcla. En otras formas de realización, la muestra que comprende la mezcla de cfADN de diferentes genomas está enriquecida no específicamente en cualquiera de los genomas presentes en la muestra.

En otras formas de realización, el cfADN de la muestra está enriquecido específicamente. Enriquecimiento específico se refiere al enriquecimiento de una muestra genómica en secuencias específicas, por ejemplo, una secuencia diana polimórfica, que se seleccionan para la amplificación antes de la secuenciación de la muestra de ADN. Sin embargo, una ventaja de las formas de realización descritas es que no es necesaria la amplificación selectiva.

Algunas de las tecnologías de secuenciación están disponibles en el mercado, tales como la plataforma de secuenciación por hibridación de Affymetrix Inc. (Sunnyvale, CA) y las plataformas de secuenciación por síntesis de 454 Life Sciences (Bradford, CT), Illumina/Solexa (Hayward, CA) y Helicos Biosciences (Cambridge, MA), y la plataforma de secuenciación por ligación de Applied Biosystems (Foster City, CA), como se describe más adelante. Además de la secuenciación de una sola molécula realizada mediante la secuenciación por síntesis de Helicos Biosciences, quedan abarcadas por el método descrito otras tecnologías de secuenciación de una sola molécula e incluyen la tecnología SMRT™ de Pacific Biosciences, la tecnología Ion Torrent™, y la secuenciación por nanoporos que está desarrollando por ejemplo, Oxford Nanopore Technologies.

Aunque el método de Sanger automatizado se considera una tecnología de "primera generación", el método descrito también puede emplear la secuenciación de Sanger, incluida la secuenciación de Sanger automatizada. Los métodos de secuenciación adicionales que comprenden el uso de tecnologías en desarrollo de formación de imágenes de ácidos nucleicos, por ejemplo, la microscopía de fuerza atómica (AFM) o la microscopía electrónica de transmisión (TEM), también quedan abarcadas por el método descrito. Más adelante se describen tecnologías de secuenciación ejemplares.

En una forma de realización, la tecnología de secuenciación de ADN que se utiliza en los métodos descritos es la Helicos True Single Molecule Sequencing (tSMS) (por ejemplo, como se describe en Harris T.D. *et al*., Science 320:106-109 [2008]). En la técnica tSMS, una muestra de ADN se escinde en cadenas de aproximadamente 100 a 200 nucleótidos, y se añade una secuencia poliA al extremo 3' de cada cadena de ADN. Cada cadena se marca añadiendo un nucleótido adenosina marcado con fluorescencia. A continuación, las cadenas de ADN se hibridan con una célula de flujo, que contiene millones de sitios de captura oligo-T que están inmovilizados en la superficie de la célula de flujo. Los moldes pueden estar a una densidad de aproximadamente 100 millones de moldes/cm^2. A continuación, se carga la célula de flujo en un instrumento, por ejemplo, un secuenciador HeliScope™, y un láser ilumina la superficie de la célula de flujo, lo que pone de manifiesto la posición de cada molde. Una cámara CCD puede mapear la posición de los moldes en la superficie de la célula de flujo. A continuación, se escinde y se quita el marcador fluorescente del molde. La reacción de secuenciación se inicia introduciendo una ADN polimerasa y un nucleótido marcado con fluorescencia. El ácido nucleico oligo-T hace de cebador. La polimerasa incorpora los nucleótidos marcados al cebador de manera dirigida por molde. La polimerasa y los nucleótidos no incorporados se eliminan. Los moldes que han dirigido la incorporación del nucleótido marcado con fluorescencia se distinguen mediante formación de imágenes de la superficie de la célula de flujo. Después de la formación de imágenes, una etapa de escisión elimina el marcador fluorescente, y el proceso se repite con otros nucleótidos marcados con fluorescencia hasta que se consigue la longitud de lectura deseada. Se recoge información de secuencias con cada etapa de adición de nucleótidos. La secuenciación del genoma completo mediante tecnologías de secuenciación de una sola molécula excluye la amplificación basada en PCR en la preparación de las bibliotecas de secuenciación, y la manera directa de preparación de muestras permite la medición directa de la muestra, más que la medición de copias de esa muestra.

En una forma de realización, la tecnología de secuenciación de ADN que se utiliza en los métodos descritos es la secuenciación 454 (Roche) (por ejemplo, como se describe en Margulies, M. *et al*., Nature 437:376-380 (2005)). La secuenciación 454 implica dos etapas. En la primera etapa, el ADN se corta en fragmentos de aproximadamente 300-800 pares de bases, y se producen fragmentos de extremos romos. A continuación, se ligan adaptadores oligonucleotídicos a los extremos de los fragmentos. Los adaptadores hacen de cebadores para la amplificación y secuenciación de los fragmentos. Los fragmentos pueden fijarse a perlas de captura de ADN, por ejemplo, perlas recubiertas con estreptavidina utilizando, por ejemplo, Adaptador B, que contiene el marcador 5'-biotina. Los fragmentos fijados a las perlas se amplifican por PCR dentro de gotitas de una emulsión de aceite-agua. El resultado es varias copias de fragmentos de ADN amplificados por clonación en cada perla. En la segunda etapa,

las perlas se capturan en pocillos (del tamaño del picolitro). Se realiza la pirosecuenciación en cada fragmento de ADN en paralelo. La adición de uno o más nucleótidos genera una señal luminosa que es registrada por una cámara CCD en un instrumento de secuenciación. La intensidad de la señal es proporcional al número de nucleótidos incorporados. La pirosecuenciación se vale de pirofosfato (PPi) que se libera tras la adición de nucleótidos. El PPi es convertido en ATP por la ATP sulfurilasa en presencia de adenosina 5' fosfosulfato. La luciferasa utiliza ATP para convertir la luciferina en oxiluciferina, y esta reacción genera luz que se distingue y se analiza.

En una forma de realización, la tecnología de secuenciación de ADN que se utiliza en los métodos descritos es la tecnología SOLiD™ (Applied Biosystems). En la secuenciación por ligación SOLiD™, el ADN genómico se corta en fragmentos, y se fijan adaptadores a los extremos 5' y 3' de los fragmentos para generar una biblioteca de fragmentos. Como alternativa, pueden introducirse adaptadores internos ligando los adaptadores a los extremos 5' y 3' de los fragmentos, circularizando los fragmentos, digiriendo el fragmento circularizado para generar un adaptador interno, y fijando los adaptadores a los extremos 5' y 3' de los fragmentos resultantes para generar una biblioteca de pares emparejados. A continuación, se preparan poblaciones de perlas clonales en microrreactores que contienen las perlas, los cebadores, el molde y los componentes de la PCR. Después de la PCR, los moldes se desnaturalizan y las perlas se enriquecen para separar las perlas con moldes extendidos. Los moldes en las perlas seleccionadas se someten a una modificación 3' que permite la unión a un portaobjetos de vidrio. La secuencia puede determinarse mediante hibridación y ligación secuencial de oligonucleótidos parcialmente aleatorios con una base (o un par de bases) determinada central que se identifica mediante un fluoróforo específico. Después de registrarse un color, el oligonucleótido ligado se escinde y elimina y, a continuación, se repite el proceso.

En una forma de realización, la tecnología de secuenciación de ADN que se utiliza en los métodos descritos es la tecnología de secuenciación de una sola molécula en tiempo real (SMRT™) de Pacific Biosciences. En la secuenciación SMRT, se forman imágenes de la continua incorporación de nucleótidos marcados con colorante durante la síntesis de ADN. Se fijan moléculas de ADN polimerasa individuales a la superficie inferior de identificadores de guía de ondas en modo cero (identificadores ZMW) individuales que obtienen información de secuencias, mientras se están incorporando nucleótidos marcados en la cadena de cebador en crecimiento. Un ZMW es una estructura de confinamiento que permite observar la incorporación de un solo nucleótido por la ADN polimerasa contra el fondo de nucleótidos fluorescentes que se difunden rápidamente dentro y fuera del ZMW (en microsegundos). Se necesitan varios milisegundos para incorporar un nucleótido en una cadena en crecimiento. Durante este tiempo, el marcador fluorescente es excitado y produce una señal fluorescente, y el marcador fluorescente se escinde. La identificación de la correspondiente fluorescencia del colorante indica qué base se incorporó. El proceso se repite.

En una forma de realización, la tecnología de secuenciación de ADN que se utiliza en los métodos descritos es la secuenciación por nanoporos (por ejemplo, como se describe en Soni GV y Meller A. Clin. Chem. 53: 1996-2001 [2007]). Varias compañías, incluida Oxford Nanopore Technologies (Oxford, Reino Unido), están desarrollando industrialmente técnicas de análisis de ADN de secuenciación por nanoporos. La secuenciación por nanoporos es una tecnología de secuenciación de una sola molécula mediante la que una sola molécula de ADN se secuencia directamente a medida que pasa por un nanoporo. Un nanoporo es un pequeño agujero, del orden de 1 nanómetro de diámetro. La inmersión de un nanoporo en un fluido conductor y la aplicación de un potencial (voltaje) a través del mismo dan como resultado una ligera corriente eléctrica debida a la conducción de iones a través del nanoporo. La cantidad de corriente que fluye es sensible al tamaño y la forma del nanoporo. A medida que una molécula de ADN pasa por un nanoporo, cada nucleótido de la molécula de ADN obstruye el nanoporo en un grado diferente, cambiando la magnitud de la corriente a través del nanoporo en diferentes grados. Por lo tanto, este cambio en la corriente a medida que la molécula de ADN pasa por el nanoporo representa una lectura de la secuencia de ADN.

En una forma de realización, la tecnología de secuenciación de ADN que se utiliza en los métodos descritos es la matriz de transistor de efecto de campo sensible a químicos (chemFET) (por ejemplo, como se describe en la publicación de patente de EE.UU. nº 2009/0026082 presentada el 17 de diciembre de 2007). En un ejemplo de la técnica, las moléculas de ADN pueden colocarse en cámaras de reacción, y las moléculas de molde pueden hibridarse con un cebador de secuenciación unido a una polimerasa. La incorporación de uno o más trifosfatos en una nueva cadena de ácido nucleico en el extremo 3' del cebador de secuenciación puede distinguirse por un cambio en la corriente mediante un chemFET. Una matriz puede tener múltiples sensores chemFET. En otro ejemplo, pueden fijarse ácidos nucleicos individuales a perlas, y los ácidos nucleicos pueden amplificarse en la perla, y las perlas individuales pueden transferirse a cámaras de reacción individuales en una matriz e chemFET, teniendo cada cámara un sensor chemFET, y pueden secuenciarse los ácidos nucleicos.

En una forma de realización, la tecnología de secuenciación de ADN que se utiliza en los métodos descritos es el método de Halcyon Molecular que utiliza microscopía electrónica de transmisión (TEM). El método, denominado Individual Molecule Placement Rapid Nano Transfer (IMPRNT), comprende utilizar la formación de imágenes de microscopio electrónico de transmisión de resolución de un solo átomo de ADN de alto peso molecular (150 kb o superior) marcado selectivamente con marcadores de átomos pesados y disponer estas moléculas en películas ultrafinas en matrices paralelas ultradensas (3 nm entre las cadenas) con una separación constante entre las bases. Se utiliza el microscopio electrónico para formar imágenes de las moléculas en las películas para determinar la posición de los marcadores de átomos pesados y para extraer información de secuencias de bases del

ADN. El método se describe adicionalmente en la publicación de patente PCT WO 2009/046445. El método permite secuenciar genomas humanos completos en menos de diez minutos.

En una forma de realización, la tecnología de secuenciación de ADN es la secuenciación de una sola molécula Ion Torrent, que empareja la tecnología de semiconductores con una química de secuenciación simple para traducir directamente la información codificada químicamente (A, C, G, T) a información digital (0, 1) en un chip semiconductor. En la naturaleza, cuando un nucleótido es incorporado en una cadena de ADN por una polimerasa, se libera como subproducto un ion de hidrógeno. Ion Torrent utiliza una matriz de alta densidad de pocillos micromecanizados para realizar este proceso bioquímico de manera masiva en paralelo. Cada pocillo contiene una molécula de ADN diferente. Por debajo de los pocillos hay una capa sensible a los iones y por debajo de eso, un detector iónico. Cuando se añade un nucleótido, por ejemplo una C, a un molde de ADN y a continuación se incorpora en una cadena de ADN, se liberará un ion de hidrógeno. La carga de ese ion cambiará el pH de la solución, que puede ser identificado por el detector iónico del Ion Torrent. El secuenciador - esencialmente el medidor de pH de estado sólido más pequeño del mundo - asigna la base, pasando directamente de la información química a la información digital. A continuación, el secuenciador Ion Personal Genome Machine (PGM™) satura el chip secuencialmente con un nucleótido tras otro. Si el siguiente nucleótido que satura el chip no es una coincidencia, no se registrará ningún cambio de voltaje y no se asignará ninguna base. Si hay dos bases idénticas en la cadena de ADN, el voltaje será doble, y el chip registrará dos bases idénticas asignadas. La identificación directa permite registrar la incorporación de nucleótidos en cuestión de segundos.

En algunas formas de realización, los métodos emplean la PCR o una técnica relacionada para amplificar secuencias nucleotídicas de la muestra antes de identificarlas o mapearlas. Sin embargo, las técnicas algorítmicas descritas en el presente documento no requieren generalmente la amplificación, en particular la amplificación selectiva de polimorfismos utilizados para estimar la fracción del genoma.

Determinadas formas de realización emplean la PCR digital y la secuenciación por hibridación. Puede utilizarse la reacción en cadena de la polimerasa digital (PCR digital o dPCR) para identificar y cuantificar directamente los ácidos nucleicos en una muestra. La PCR digital puede realizarse en una emulsión. Se separan ácidos nucleicos individuales, por ejemplo, en un dispositivo de cámara microfluídica, y cada ácido nucleico se amplifica individualmente mediante PCR. Los ácidos nucleicos pueden separarse de manera que haya un promedio de aproximadamente 0,5 ácidos nucleicos/pocillo, o no más de un ácido nucleico/pocillo. Pueden utilizarse diferentes sondas para distinguir los alelos fetales y los alelos maternos. Los alelos pueden enumerarse para determinar el número de copias. En la secuenciación por hibridación, la hibridación comprende poner en contacto la pluralidad de secuencias polinucleotídicas con una pluralidad de sondas polinucleotídicas, en la que cada una de la pluralidad de sondas polinucleotídicas puede estar opcionalmente anclada a un sustrato. El sustrato puede ser una superficie plana que comprende una matriz de secuencias nucleotídicas conocidas. El patrón de hibridación a la matriz puede utilizarse para determinar las secuencias polinucleotídicas presentes en la muestra. En otras formas de realización, cada sonda está anclada a una perla, por ejemplo, una perla magnética o similar. Puede identificarse y utilizarse la hibridación a las perlas para identificar la pluralidad de secuencias polinucleotídicas dentro de la muestra.

En una forma de realización, el método emplea la secuenciación masiva en paralelo de millones de fragmentos de ADN utilizando la secuenciación por síntesis de Illumina y la química de secuenciación basada en terminadores reversibles (por ejemplo, como se describe en Bentley *et al*., Nature 6:53-59 [2009]). El ADN molde puede ser ADN genómico, por ejemplo, cfADN. En algunas formas de realización, se utiliza como molde ADN genómico de células aisladas, y se fragmenta en longitudes de varios cientos de pares de bases. En otras formas de realización, se utiliza como molde cfADN, y no resulta necesaria la fragmentación ya que el cfADN existe en forma de fragmentos cortos. Por ejemplo, el cfADN fetal circula en el torrente sanguíneo en forma de fragmentos de < 300 pb, y se ha estimado que el cfADN materno circula en forma de fragmentos de entre aproximadamente 0,5 kb y 1 kb (Li *et al*., Clin. Chem., 50: 1002-1011 (2004)). La tecnología de secuenciación de Illumina se basa en la fijación del ADN genómico fragmentado a una superficie plana ópticamente transparente sobre la que están unidos los anclajes oligonucleotídicos. El ADN molde se somete a reparación de extremos para generar extremos romos fosforilados en 5', y se utiliza la actividad polimerasa del fragmento Klenow para añadir una sola base A al extremo 3' de los fragmentos de ADN fosforilados romos. Esta adición prepara los fragmentos de ADN para la ligación a adaptadores oligonucleotídicos, que tienen una protuberancia de una sola base T en su extremo 3' para aumentar la eficacia de ligación. Los oligonucleótidos del adaptador son complementarios a los anclajes de las células de flujo. En condiciones de dilución limitante, se añade ADN molde monocatenario modificado con adaptador a la célula de flujo y se inmoviliza mediante hibridación a los anclajes. Los fragmentos de ADN fijados se prolongan y someten a amplificación en puente para crear una célula de flujo de secuenciación de densidad ultra alta con cientos de millones de clústeres, conteniendo cada uno ~ 1.000 copias del mismo molde. En una forma de realización, el ADN genómico fragmentado al azar, por ejemplo, cfADN, se amplifica utilizando PCR antes de someterlo a amplificación de clústeres. Como alternativa, se utiliza una preparación de biblioteca genómica sin amplificación, y el ADN genómico fragmentado al azar, por ejemplo, cfADN se enriquece mediante la amplificación de clústeres solo (Kozarewa *et al*., Nature Methods 6:291-295 [2009]). Los moldes se secuenciaron utilizando una tecnología robusta de secuenciación por síntesis de ADN de cuatro colores que emplea terminadores reversibles con colorantes fluorescentes eliminables. La identificación de fluorescencia de alta sensibilidad se consigue utilizando la óptica de reflexión interna total y excitación láser. Se alinean lecturas de secuencias cortas de aproximadamente 20 pb-40 pb,

por ejemplo, 36 pb, contra de un genoma de referencia con enmascaramiento de repeticiones y las diferencias genéticas se asignan utilizando un software de pipelines de análisis de datos especialmente desarrollado. Una vez terminada la primera lectura, los moldes pueden regenerarse *in situ* para posibilitar una segunda lectura desde el extremo opuesto de los fragmentos. Por lo tanto, se utiliza la secuenciación de un único extremo o de extremos emparejados de los fragmentos de ADN según el método. Se realiza la secuenciación parcial de fragmentos de ADN presentes en la muestra, y se realiza el recuento de los marcadores de secuencia que comprenden lecturas de una longitud predeterminada, por ejemplo, 36 pb, que se mapean contra un genoma de referencia conocido.

La longitud de la lectura de secuencia está asociada con la tecnología de secuenciación concreta. Los métodos NGS proporcionan lecturas de secuencias que varían en tamaño desde decenas a cientos de pares de bases. En algunas formas de realización del método descrito en el presente documento, las lecturas de secuencias tienen aproximadamente 20 pb, aproximadamente 25 pb, aproximadamente 30 pb, aproximadamente 35 pb, aproximadamente 40 pb, aproximadamente 45 pb, aproximadamente 50 pb, aproximadamente 55 pb, aproximadamente 60 pb, aproximadamente 65 pb, aproximadamente 70 pb, aproximadamente 75 pb, aproximadamente 80 pb, aproximadamente 85 pb, aproximadamente 90 pb, aproximadamente 95 pb, aproximadamente 100 pb, aproximadamente 110 pb, aproximadamente 120 pb, aproximadamente 130 pb, aproximadamente 140 pb, aproximadamente 150 pb, aproximadamente 200 pb, aproximadamente 250 pb, aproximadamente 300 pb, aproximadamente 350 pb, aproximadamente 400 pb, aproximadamente 450 pb, o aproximadamente 500 pb. Se espera que los avances tecnológicos posibiliten lecturas de un único extremo de más de 500 pb que posibiliten lecturas de más de aproximadamente 1.000 pb cuando se generen lecturas de extremos emparejados. En una forma de realización, las lecturas de secuencias tienen 36 pb. Otros métodos de secuenciación que pueden emplearse mediante los métodos descritos incluyen los métodos de secuenciación de una sola molécula que pueden secuenciar moléculas de ácidos nucleicos > 5.000 pb. La cantidad masiva de salida de secuencias se transfiere mediante un pipeline de análisis que transforma la salida de imágenes primarias del secuenciador en cadenas de bases. Un paquete de algoritmos integrados realiza las etapas principales de transformación de datos primarios: análisis de imágenes, puntuación de la intensidad, asignación de bases y alineamiento.

MAPEO

Pueden utilizarse diversos métodos computacionales para mapear cada secuencia identificada contra un bin, por ejemplo, identificando todas las secuencias en la muestra que se mapean contra un gen, cromosoma, alelo concretos, u otra estructura. Existen varios algoritmos informáticos para alinear secuencias, incluidos sin limitación BLAST (Altschul *et al*., 1990), BLITZ (MPsrch) (Sturrock y Collins, 1993), FASTA (Person y Lipman, 1988), BOWTIE (Langmead *et al*., Genome Biology 10:R25.1-R25.10 [2009]), o ELAND (Illumina, Inc., San Diego, CA, EE.UU.). En algunas formas de realización, las secuencias de los bins se encuentran en bases de datos de ácidos nucleicos conocidas por los expertos en la materia, incluidas sin limitación GenBank, dbEST, dbSTS, EMBL (el Laboratorio Europeo de Biología Molecular), y el DDBJ (el Banco de Datos de ADN de Japón). Puede utilizarse BLAST o herramientas similares para buscar las secuencias identificadas contra las bases de datos de secuencias, y puede utilizarse la búsqueda de aciertos para clasificar las secuencias identificadas en los bins apropiados.

APARATO

El análisis de los datos de secuenciación y los diagnósticos derivados de los mismos se realizan por lo general utilizando un hardware informático, que opera según algoritmos y programas definidos. Por lo tanto, determinadas formas de realización emplean procesos que implican los datos almacenados en o transferidos a través de uno o más sistemas informáticos u otros sistemas de procesamiento. Las formas de realización de la invención también se refieren a un aparato para realizar estas operaciones. Este aparato puede construirse especialmente para los fines necesarios, o puede ser un ordenador (o un grupo de ordenadores) de uso general selectivamente activado o reconfigurado por un programa informático y/o estructura de datos almacenada en el ordenador. En algunas formas de realización, un grupo de procesadores realiza algunas o todas las operaciones analíticas mencionadas conjuntamente (por ejemplo, a través de una red o computación en nube) y/o en paralelo. Un procesador o grupo de procesadores para realizar los métodos descritos en el presente documento puede ser de diversos tipos, incluidos microcontroladores y microprocesadores tales como dispositivos programables (por ejemplo, CPLDs y FPGAs) y otros dispositivos tales como ASIC basados en matriz de puertas, procesadores de señales digitales, y/o microprocesadores de uso general.

Además, determinadas formas de realización se refieren a productos de programa informático o medios legibles por ordenador tangibles y/o no transitorios que incluyen datos y/o instrucciones de programa (incluidas estructuras de datos) para realizar diversas operaciones implementadas por ordenador. Los ejemplos de medios legibles por ordenador incluyen, pero no se limitan a, dispositivos de memoria semiconductora, medios magnéticos tales como unidades de disco, cinta magnética, medios ópticos tales como CDs, medios magneto-ópticos, y dispositivos de hardware que están especialmente configurados para almacenar y ejecutar instrucciones de programa, tal como dispositivos de memoria de sólo lectura (ROM) y memoria de acceso aleatorio (RAM). Los medios legibles por ordenador pueden ser controlados directamente por un usuario final o los medios pueden ser controlados indirectamente por el usuario final. Los ejemplos de medios directamente controlados incluyen los medios ubicados en un equipo del usuario y/o medios que no son compartidos con otras entidades. Los ejemplos de

medios controlados indirectamente incluyen medios que son accesibles para el usuario indirectamente a través de una red externa y/o a través de un servicio que proporciona recursos compartidos tales como la "nube". Los ejemplos de instrucciones de programa incluyen tanto código máquina, tal como es producido por un compilador y archivos que contienen un código de nivel superior que puede ser ejecutado por el ordenador utilizando un intérprete.

En una forma de realización, se describe un producto de programa informático para generar una salida que indica la fracción de ácido nucleico derivada de un genoma definido (tal como el de un feto) y, opcionalmente, otra información tal como la presencia o ausencia de una aneuploidía fetal en una muestra de ensayo. El producto informático puede contener instrucciones para realizar uno cualquiera o más de los métodos anteriormente descritos para determinar una fracción de ácidos nucleicos a partir de un organismo concreto. Como se ha explicado, el producto informático puede incluir un medio legible por ordenador no transitorio y/o tangible que tiene una lógica compilable o ejecutable por ordenador (por ejemplo, instrucciones) grabada en el mismo para permitir que un procesador determine la fracción del genoma y, en algunos casos, si está presente o ausente en el genoma una aneuploidía u otra afección. En un ejemplo, el producto informático comprende un medio legible por ordenador que tiene una lógica compilable o ejecutable por ordenador (por ejemplo, instrucciones) grabada en el mismo para permitir que un procesador determine la fracción fetal y diagnostique una aneuploidía fetal que comprende: un procedimiento de recepción para recibir datos de secuenciación de al menos una porción de las moléculas de ácido nucleico de una muestra biológica materna, en el que dichos datos de secuenciación comprenden secuencias en los loci de uno o más polimorfismos; lógica asistida por ordenador para analizar secuencias para determinar los recuentos de alelos para el uno o más polimorfismos, y determinar la fracción fetal de los ácidos nucleicos en la muestra biológica materna; y un procedimiento de salida para generar una salida que indica la fracción fetal de ácidos nucleicos en la muestra.

La información de secuencias de la muestra en cuestión puede mapearse contra las secuencias de referencia de polimorfismo como se ha descrito. Además, la información de secuencias mapeada puede utilizarse para generar recuentos de alelos y/o determinar los casos de cigosidad para los polimorfismos. Tal información puede utilizarse para determinar la fracción fetal. En diversas formas de realización, las secuencias de referencia de polimorfismo se almacenan en una base de datos tal como una base de datos relacional u orientada a objetos, por ejemplo. Debe entenderse que no es práctico, ni incluso posible en la mayoría de los casos, que un ser humano sin ayuda realice cualquiera de estas operaciones computacionales o todas ellas. Por ejemplo, el mapeo de una única lectura de 30 pb de una muestra contra una base de datos de secuencias de referencia de polimorfismo llevaría un período prohibitivamente largo sin la ayuda de un aparato computacional. Por supuesto, el problema se agrava porque las asignaciones fiables con frecuencia requieren mapear miles (por ejemplo, al menos aproximadamente 10.000) o incluso millones de lecturas contra uno o más cromosomas.

En determinadas formas de realización, los métodos descritos se valen de una lista almacenada u otro conjunto de datos organizado acerca de los polimorfismos de referencia para el organismo que produce las secuencias de ácidos nucleicos a analizar. Como se ha explicado anteriormente, las secuencias de la muestra en cuestión pueden alinearse o mapearse de otro modo contra los polimorfismos almacenados. Los polimorfismos individuales son por lo general secuencias de una longitud suficiente para mapearse de modo inequívoco contra las secuencias identificadas a partir de la muestra de ácido nucleico. Por lo general, los polimorfismos vienen en grupos, uno para cada alelo. En diversas formas de realización, los polimorfismos de referencia se almacenan en una base de datos que contiene las características de los polimorfismos, además de sus secuencias. Esta recopilación de información acerca de los polimorfismos puede almacenarse en una base de datos relacional u orientada a objetos, por ejemplo.

La Figura 10 ilustra un sistema informático típico que, cuando está debidamente configurado o diseñado, puede hacer de aparato de análisis de la presente invención. El sistema informático 200 incluye cualquier número de procesadores 202 (también denominados unidades centrales de procesamiento, o CPUs) que están acoplados a dispositivos de almacenamiento, incluido el almacenamiento primario 206 (por lo general una memoria de acceso aleatorio, o RAM), un almacenamiento primario 204 (por lo general una memoria de sólo lectura, o ROM). La CPU 202 puede ser de diversos tipos, incluidos microcontroladores y microprocesadores tales como dispositivos programables (por ejemplo, CPLDs y FPGAs) y dispositivos no programables tal como ASIC basados en matriz de puertas o microprocesadores de uso general. Como es conocido en la técnica, el almacenamiento primario 204 funciona para transferir datos e instrucciones a la CPU y el almacenamiento primario 206 se utiliza por lo general para transferir datos e instrucciones de manera bidireccional. Ambos de estos dispositivos de almacenamiento primario pueden incluir cualquier medio legible por ordenador adecuado tal como los descritos anteriormente. También hay un dispositivo de almacenamiento masivo 208 acoplado bidireccionalmente a la CPU 202 y proporciona capacidad de almacenamiento de datos adicional y puede incluir cualquiera de los medios legibles por ordenador descritos anteriormente. El dispositivo de almacenamiento masivo 208 puede utilizarse para almacenar programas, datos y similares, y es por lo general un medio de almacenamiento secundario tal como un disco duro. Se entenderá que la información guardada en del dispositivo de almacenamiento masivo 208, puede, en los casos apropiados, incorporarse de manera convencional como parte del almacenamiento primario 206 como memoria virtual. Un dispositivo de almacenamiento masivo específico tal como un CD-ROM 214 también puede pasar datos unidireccionalmente a la CPU.

La CPU 202 también está acoplada a una interfaz 210 que se conecta a uno o más dispositivos de entrada/salida tales como monitores de vídeo, ratones, teclados, micrófonos, pantallas sensibles al tacto, lectores de tarjetas transductores, tabletas, lápices digitales, reconocedores de voz o de escritura manuscrita, u otros dispositivos de entrada conocidos tales como, por supuesto, otros ordenadores. Por último, la CPU 202 puede estar acoplada opcionalmente a un dispositivo externo tal como una base de datos o una red de ordenadores o de telecomunicaciones utilizando una conexión externa como se muestra en general en 212. Con una conexión de este tipo, se contempla que la CPU pueda recibir información de la red, o pueda devolver información a la red durante la realización de las etapas del método descritas en el presente documento.

Un usuario puede introducir una secuencia u otros datos en un ordenador, ya sea directa o indirectamente. En una forma de realización, el sistema informático 200 está directamente acoplado a una herramienta de secuenciación que lee y/o analiza secuencias de ácidos nucleicos amplificados. Las secuencias u otra información de tales herramientas se proporcionan a través de la interfaz 212 para su análisis por el sistema 200. Como alternativa, las secuencias procesadas por el sistema 200 se proporcionan a partir de una fuente de almacenamiento de secuencias tal como una base de datos u otro depósito. Una vez en el aparato de procesamiento 200, un dispositivo de memoria tal como el almacenamiento primario 206 o el almacenamiento masivo 208 almacena en búfer o almacena, al menos temporalmente, secuencias de ácidos nucleicos. Además, el dispositivo de memoria puede almacenar números de marcadores para diversos cromosomas o genes, recuentos de copias calculados, etc. La memoria también puede almacenar diversas rutinas y/o programas para analizar la presentación de la secuencia o los datos mapeados. Tales programas/rutinas pueden incluir programas para realizar análisis estadísticos, etc.

En un ejemplo, un usuario proporciona una muestra a un aparato de secuenciación. Los datos son recogidos y/o analizados por el aparato de secuenciación que está conectado a un ordenador. El software del ordenador permite la recogida y/o el análisis de los datos. Los datos pueden almacenarse, presentarse (a través de un monitor u otro dispositivo similar) y/o enviarse a otro lugar. Como se ha indicado, el ordenador puede estar conectado a Internet, que se utiliza para transmitir los datos a un dispositivo portátil utilizado por un usuario remoto (por ejemplo, un médico, científico o analista). Se entiende que los datos pueden almacenarse y/o analizarse antes de la transmisión. En algunas formas de realización, los datos sin procesar se recogen y envían a un usuario remoto (o aparato) que analizará y/o almacenará los datos. La transmisión puede producirse vía Internet, pero también puede producirse vía satélite u otra conexión. Como alternativa, los datos pueden almacenarse en un medio legible por ordenador (por ejemplo, un CD o un dispositivo de almacenamiento de memoria semiconductora) y el medio puede enviarse a un usuario final (por ejemplo, por correo). El usuario remoto puede estar en la misma ubicación geográfica o en una diferente, incluida pero no limitada a un edificio, ciudad, estado, país o continente.

En algunas formas de realización, los métodos de la invención comprenden adicionalmente recoger datos referentes a una pluralidad de secuencias polinucleotídicas y enviar los datos a un ordenador. Por ejemplo, el ordenador puede estar conectado a equipos de laboratorio, por ejemplo, un aparato de recogida de muestras, un aparato de amplificación de nucleótidos, un aparato de secuenciación de nucleótidos, o un aparato de hibridación. A continuación, el ordenador puede recoger los datos aplicables recopilados por el dispositivo de laboratorio. Los datos pueden almacenarse en un ordenador en cualquier etapa, por ejemplo, mientras se recogen en tiempo real, antes del envío, durante o juntamente con el envío, o después del envío. Los datos pueden almacenarse en un medio legible por ordenador que puede extraerse del equipo. Los datos recogidos o almacenados pueden transmitirse desde el ordenador a un lugar remoto, por ejemplo, a través de una red local o una red de área amplia tal como Internet.

En un aspecto, la descripción proporciona adicionalmente un sistema capaz de realizar el análisis cuantitativo de la secuenciación de nucleótidos con una precisión de al menos un 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, o al menos un 99%. La secuenciación de nucleótidos puede comprender la secuenciación de Sanger, la secuenciación masiva en paralelo, la hibridación u otras técnicas como se han descrito en el presente documento. El aparato puede comprender diversos componentes, por ejemplo, equipos de laboratorio y sistemas informáticos, y puede configurarse para llevar a cabo los métodos de la invención descritos en el presente documento.

En algunas formas de realización, el aparato y/o las instrucciones de programación pueden incluir adicionalmente instrucciones para registrar automáticamente la información pertinente al método tal como la fracción de ADN fetal y, opcionalmente, la presencia o ausencia de una aneuploidía cromosómica fetal en el expediente clínico de un paciente para una paciente humana que proporciona la muestra de ensayo materna. El expediente clínico del paciente puede mantenerse en un sitio web de expedientes clínicos personal, un laboratorio, consultorio médico, un hospital, una organización de mantenimiento de la salud o una compañía de seguros. Además, basándose en los resultados del análisis implementado por procesador, el método puede implicar adicionalmente prescribir, iniciar y/o modificar el tratamiento de una paciente humana de la que se obtuvo la muestra de ensayo materna. Esto puede implicar realizar uno o más ensayos o análisis adicionales de otras muestras tomadas del sujeto.

Ejemplo

Fracción fetal predicha a partir de las variaciones secuenciadas: Caso 2

Para demostrar que el método de la presente invención puede utilizarse para estimar con fiabilidad la fracción fetal en una muestra materna, se creó una muestra “materna” artificial, y se identificaron las variaciones de bases en todos los loci de los cromosomas 1 y 7 para predecir la fracción de genoma de contribución minoritaria.

El cfADN que se aísla de una mujer embarazada es una mezcla de cfADN materno y fetal, correspondiendo el nivel de cfADN fetal a una mediana de ~ 10% del cfADN total (Lo *et al*., 2010, "Maternal Plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus", Prenatal Diagnosis, 2, 1-12). Para crear la muestra materna artificial, se utilizó ADN genómico (ADNg) obtenido de una madre y su hijo (ADNs de madre e hijo NA10924 y NA10925; The Coriell Institute for Medical Research, Camden, NJ) para crear la muestra de genomas mixtos. Se cortaron cinco microgramos del ADNg de la madre y del ADNg del hijo, en fragmentos de aproximadamente 200 pb, y se determinó la concentración de cada uno. Se creó una muestra artificial que contenía un 10% de ADN del hijo y un 90% de ADN de la madre para imitar una muestra de sangre materna, que se cree contiene por lo general del 2%-40% de cfADN fetal, dependiendo de la edad gestacional [Lun *et al*., 2008 "Microfluidics digital PCR reveals a higher than expected fraction of fetal DNA in maternal plasma", Clinical Chemistry, 54, 1664-1672]. Se preparó una biblioteca de secuenciación a partir del ADN de la muestra artificial, y se sometió a 50 ciclos de secuenciación en 4 calles de la célula de flujo utilizando el IlluminaHiSeq 2000. Se generaron aproximadamente 800 millones de lecturas de secuencias 49-mero.

Los ~ 800 millones de lecturas se alinearon contra el Genoma de Referencia Humano con Enmascaramiento de Repeticiones (construcción hg19) utilizando el algoritmo GSNAP (http://research-pub.gene.com/gmap/), lo que permite un desapareamiento y ninguna inserción y ninguna deleción. Se ignoraron las secuencias que se mapearon contra múltiples ubicaciones en el genoma. Todas las demás lecturas mapeadas se contaron como marcadores de secuencia, y sólo los loci contra los que se mapearon 40 y 100 marcadores de secuencia se tuvieron en cuenta para su posterior análisis, es decir, sólo se tuvieron en cuenta las bases con una cobertura de 40 y 100 marcadores.

Para cada locus de bases, se hizo el recuento del número de marcadores que se mapearon contra cada una de las cuatro bases. Se eliminaron los loci con más de dos bases posibles, y sólo los marcadores que se mapearon contra loci monoalélicos y bialélicos se utilizaron para predecir la fracción fetal artificial. El número total de marcadores que se mapearon en cada locus de bases representó la cobertura (D) en ese locus. En esta muestra materna simulada, se espera que la contribución del alelo mayoritario de la madre (B) refleje la porción del 90% de los marcadores, y la contribución del alelo minoritario del hijo (A) refleje la porción del 10% de los marcadores.

Las Figuras 11A y B muestran los histogramas del número de observaciones de bases variantes (Frecuencia) en los cromosomas 1 y 7, respectivamente para los porcentajes de alelos minoritarios (A/D) para los cromosomas 1 y 7. El porcentaje del alelo minoritario es el porcentaje del número total de alelos en un determinado locus. Por ejemplo, para un determinado locus en el que hay 8 apariciones del alelo minoritario A y 56 apariciones del alelo mayoritario B, el porcentaje del alelo minoritario es 8%. Los datos muestran que el mayor número de apariciones (Frecuencia) para el alelo minoritario se observa cuando el alelo minoritario está presente en un 5%, lo que representa la mitad de la fracción fetal. Por consiguiente, los datos predijeron que la muestra contenía una fracción fetal del 10%, que se corresponde con la utilizada para crear la muestra materna artificial.

Las Figuras 12A y B muestran la distribución de la frecuencia alélica en los cromosomas 1 y 7, respectivamente. Ambos gráficos muestran que el número máximo de alelos variantes en los cromosomas se producen a una frecuencia del alelo minoritario del 5% y una frecuencia del alelo mayoritario del 95%. Algunos de los puntos de datos restantes representan loci bialélicos presentes en el genoma de la madre, mientras que otros representan el ruido de la metodología de secuenciación. La porción central de cada gráfico donde los alelos variantes no están representados coincide con los centrómeros del cromosoma, que se sabe son regiones ricas en repeticiones de los cromosomas, contra las que se mapean los marcadores en más de un locus y que, por lo tanto, se excluyen del análisis. En otras regiones, por ejemplo las regiones que flanquean el centrómero y las regiones correspondientes a los telómeros, los alelos variantes están sobrerrepresentados. La sobrerrepresentación de estas regiones puede atribuirse a la metodología de secuenciación por la cual algunas regiones se secuencian a mayores niveles que otras.

Por lo tanto, el método de la presente invención puede utilizarse para predecir la fracción fetal. El método es particularmente útil, ya que no requiere la identificación de secuencias diana, por ejemplo, SNPs, y cualquier variación en cualquier posición de cualquier cromosoma pueden servir para predecir el porcentaje de fracción fetal.

Otras formas de realización

Aunque lo anteriormente indicado ha descrito en general la presente invención según procesos y aparatos específicos, la presente descripción tiene un abanico de aplicabilidad mucho más amplio. En particular, la presente descripción se ha descrito en términos de detección de la fracción de ADN fetal en una muestra de ADN obtenida de una embarazada, pero no se limita a ello, ya que los conceptos y métodos presentados en el presente documento

también pueden aplicarse en otros contextos, tales como la detección de las cantidades relativas de los tipos de ADN en una muestra que tiene ADN procedente de dos o más genomas diferentes. Por supuesto, los expertos en la materia reconocerán otras variaciones, modificaciones y alternativas.

Por ejemplo, aunque la mayoría de los ejemplos y aplicaciones descritos en el presente documento tienen que ver con la estimación de la fracción fetal de ADN en una muestra de ADN obtenida de un individuo que lleva un feto, la descripción no se limita a ello. De manera más general, diversas formas de realización describen métodos para evaluar las cantidades relativas de ácidos nucleicos de dos genomas diferentes en una muestra de ensayo que contiene una mezcla de ácidos nucleicos de los dos genomas diferentes, y que se sabe o se sospecha difieren en la cantidad de una o más secuencias de interés. La mezcla de ácidos nucleicos se deriva de dos o más tipos de células.

Además, aunque la mayoría de los ejemplos presentados en el presente documento tienen que ver con muestras tomadas de una embarazada humana, la descripción no se limita a ello. Por ejemplo, el individuo que proporciona una muestra a ensayar puede ser un organismo que comprende secuencias polinucleotídicas, por ejemplo, una planta, un insecto tal como una mosca, o un animal. En algunas formas de realización, el sujeto es un mamífero, por ejemplo, un ratón, rata, perro, mono o ser humano. Como se indica, el sujeto puede ser una paciente embarazada. El sujeto podría ser un paciente con una enfermedad tal como un cáncer, o podría estar infectado con un cuerpo extraño tal como un microorganismo, por ejemplo, un virus. La muestra puede comprender un fluido corporal del sujeto, por ejemplo, sangre, plasma, suero, esputo, saliva, orina, excrementos, pus, linfa, moco o similares. Por ejemplo, la muestra puede ser una muestra de plasma materno que contiene una mezcla de ADN libre materno y fetal. En general, los métodos descritos pueden implicar secuenciar el ADN de una muestra; mapear las lecturas de secuencias contra los polimorfismos; clasificar los polimorfismos en base a la cigosidad; y estimar la fracción de ADN de una fuente secundaria en la muestra.

Anexo 1. Listado de secuencias de la base de datos de búsqueda de alelos

>rs560681.1|Cr.1|longitud=111|alelo=A

CACATGCACA GCCAGCAACC CTGTCAGCAG GAGTTCCCAC CAGTTTCTTT
CTGAGAACAT CTGTTCAGGT TTCTCTCCAT CTCTATTTAC TCAGGTCACA
GGACCTTGGG G

>rs560681.2|Cr.1|longitud=111|alelo=G

CACATGCACA GCCAGCAACC CTGTCAGCAG GAGTTCCCAC CAGTTTCTTT
CTGAGAACAT CTGTTCAGGT TTCTCTCCAT CTCTGTTTAC TCAGGTCACA
GGACCTTGGG G

>rs1109037.1|Cr.2|longitud=126|alelo=A

TGAGGAAGTG AGGCTCAGAG GGTAAGAAAC TTTGTCACAG AGCTGGTGGT
GAGGGTGGAG ATTTTACACT CCCTGCCTCC CACACCAGTT TCTCCAGAGT
GGAAAGACTT TCATCTCGCA CTGGCA

>rs1109037.2|Cr.2|longitud=126|alelo=G

TGAGGAAGTG AGGCTCAGAG GGTAAGAAAC TTTGTCACAG AGCTGGTGGT
GAGGGTGGAG ATTTTACACT CCCTGCCTCC CACACCAGTT TCTCCGGAGT
GGAAAGACTT TCATCTCGCA CTGGCA

>rs9866013.1|Cr.3|longitud=121|alelo=C

GTGCCTTCAG AACCTTTGAG ATCTGATTCT ATTTTTAAAG CTTCTTAGAA
GAGAGATTGC AAAGTGGGTT GTTTCTCTAG CCAGACAGGG CAGGCAAATA
GGGGTGGCTG GTGGGATGGGA

>rs9866013.2|Cr.3|longitud=121|alelo=T

GTGCCTTCAG AACCTTTGAG ATCTGATTCT ATTTTTAAAG CTTCTTAGAA
GAGAGATTGC AAAGTGGGTT GTTTCTCTAG CCAGACAGGG CAGGTAAATA
GGGGTGGCTG GTGGGATGGGA

>rs13182883.1|Cr.5|longitud=111|alelo=A

AGGTGTGTCT CTCTTTTGTG AGGGGAGGGG TCCCTTCTGG CCTAGTAGAG
GGCCTGGCCT GCAGTGAGCA TTCAAATCCT CAAGGAACAG GGTGGGGAGG
TGGGACAAAG G

>rs13182883.2|Cr.5|longitud=111|alelo=G

AGGTGTGTCT CTCTTTTGTG AGGGGAGGGG TCCCTTCTGG CCTAGTAGAG
GGCCTGGCCT GCAGTGAGCA TTCAAATCCT CGAGGAACAG GGTGGGGAGG
TGGGACAAAG G

>rs13218440.1|Cr.6|longitud=139|alelo=A

CCTCGCCTAC TGTGCTGTTT CTAACCATCA TGCTTTTCCC TGAATCTCTT
GAGTCTTTTT CTGCTGTGGA CTGAAACTTG ATCCTGAGAT TCACCTCTAG
TCCCTCTGAG CAGCCTCCTG GAATACTCAG CTGGGATGG

>rs13218440.2|Cr.6|longitud=139|alelo=G

CCTCGCCTAC TGTGCTGTTT CTAACCATCA TGCTTTTCCC TGAATCTCTT
GAGTCTTTTT CTGCTGTGGA CTGAAACTTG ATCCTGAGAT TCACCTCTAG
TCCCTCTGGG CAGCCTCCTG GAATACTCAG CTGGGATGG

>rs4606077.1|Cr.8|longitud=114|alelo=C

GCAACTCCCT CAACTCCAAG GCAGACACCA AAGCCCTCCC TGCCTGTGGC
TTTGTAGTTC TAGTGTGGGA TCTGACTCCC CACAGCCCAC CCAAAGCCGG
GGAACTCCTC ACTG

>rs4606077.2|Cr.8|longitud=114|alelo=T

GCAACTCCCT CAACTCCAAG GCAGACACCA AAGCCCTCCC TGCCTGTGGC
TTTGTAGTTC TAGTGTGGGA TCTGACTCCC CACAGCCTAC CCAAAGCCGG
GGAACTCCTC ACTG

>rs7041158.1|Cr.9|longitud=117|alelo=C

AATTGCAATG GTGAGAGGTT GATGGTAAAA TCAAACGGAA CTTGTTATTT
TGTCATTCTG ATGGACTGGA ACTGAGGATT TTCAATTTCC TCTCCAACCC
AAGACACTTC TCACTGG

>rs7041158.2|Cr.9|longitud=117|alelo=T

AATTGCAATG GTGAGAGGTT GATGGTAAAA TCAAACGGAA CTTGTTATTT
TGTCATTCTG ATGGACTGGA ACTGAGGATT TTCAATTTCC TTTCCAACCC
AAGACACTTC TCACTGG

>rs740598.1|Cr.10|longitud=114|alelo=A

GAAATGCCTT CTCAGGTAAT GGAAGGTTAT CCAAATATTT TTCGTAAGTA
TTTCAAATAG CAATGGCTCG TCTATGGTTA GTCTCACAGC CACATTCTCA
GAACTGCTCA AACC

>rs740598.2|Cr.10|longitud=114|alelo=G

GAAATGCCTT CTCAGGTAAT GGAAGGTTAT CCAAATATTT TTCGTAAGTA
TTTCAAATAG CAATGGCTCG TCTATGGTTA GTCTCGCAGC CACATTCTCA
GAACTGCTCA AACC

>rs10773760.1|Cr.12|longitud=128|alelo=A

ACCCAAAACA CTGGAGGGGC CTCTTCTCAT TTTCGGTAGA CTGCAAGTGT
TAGCCGTCGG GACCAGCTTC TGTCTGGAAG TTCGTCAAAT TGCAGTTAAG
TCCAAGTATG CCACATAGCA GATAAGGG

>rs10773760.2|Cr.12|longitud=128|alelo=G

ACCCAAAACA CTGGAGGGGC CTCTTCTCAT TTTCGGTAGA CTGCAAGTGT
TAGCCGTCGG GACCAGCTTC TGTCTGGAAG TTCGTCAAAT TGCAGTTAGG
TCCAAGTATG CCACATAGCA GATAAGGG

>rs4530059.1|Cr.14|longitud=110|alelo=A

GCACCAGAAT TTAAACAACG CTGACAATAA ATATGCAGTC GATGATGACT
TCCCAGAGCT CCAGAAGCAA CTCCAGCACA CAGAGAGGCG CTGATGTGCC
TGTCAGGTGC

>rs4530059.2|Cr.14|longitud=110|alelo=G

GCACCAGAAT TTAAACAACG CTGACAATAA ATATGCAGTC GATGATGACT
TCCCAGAGCT CCAGAAGCAA CTCCAGCACA CGGAGAGGCG CTGATGTGCC
TGTCAGGTGC

>rs1821380.1|Cr.15|longitud=139|alelo=C

GCCCAGATTA GATGGAACCT TTTCCTCTTT TCCAGTGCAA GACAAGCGAT
TGAAAGAAGT GGATGTGTTA TTGCGGGCAC AATGGAGCCA CTGAACTGCA
GTGCAAAAAT GCAGTAAGGC ATACAGATAG AAGAAGGAG

>rs1821380.2|Cr.15|longitud=139|alelo=G

GCCCAGATTA GATGGAACCT TTTCCTCTTT TCCAGTGCAA GACAAGCGAT
TGAAAGAAGT GGATGTGTTA TTGCGGGCAC AATGGAGCCA CTGAACTGCA
GTGCAAAAAT GCAGTAAGGG ATACAGATAG AAGAAGGAG

>rs7205345.1|Cr.16|longitud=116|alelo=C

TGACTGTATA CCCCAGGTGC ACCCTTGGGT CATCTCTATC ATAGAACTTA
TCTCACAGAG TATAAGAGCT GATTTCTGTG TCTGCCTCTC ACACTAGACT
TCCACATCCT TAGTGC

>rs7205345.2|Cr.16|longitud=116|alelo=G

TGACTGTATA CCCCAGGTGC ACCCTTGGGT CATCTCTATC ATAGAACTTA
TCTCACAGAG TATAAGAGCT GATTTCTGTG TCTGCCTGTC ACACTAGACT
TCCACATCCT TAGTGC

>rs8078417.1|Cr.17|longitud=110|alelo=C

TGTACGTGGT CACCAGGGGA CGCCTGGCGC TGCGAGGGAG GCCCCGAGCC
TCGTGCCCCC GTGAAGCTTC AGCTCCCCTC CCCGGCTGTC CTTGAGGCTC
TTCTCACACT

>rs8078417.2|Cr.17|longitud=110|alelo=T

TGTACGTGGT CACCAGGGGA CGCCTGGCGC TGCGAGGGAG GCCCCGAGCC
TCGTGCCCCC GTGAAGCTTC AGCTCCCCTC CCTGGCTGTC CTTGAGGCTC
TTCTCACACT

>rs576261.1|Cr.19|longitud=114|alelo=A

CAGTGGACCC TGCTGCACCT TTCCTCCCCT CCCATCAACC TCTTTTGTGC
CTCCCCCTCC GTGTACCACC TTCTCTGTCA CCAACCCTGG CCTCACAACT
CTCTCCTTTG CCAC

>rs576261.2|Cr.19|longitud=114|alelo=C

CAGTGGACCC TGCTGCACCT TTCCTCCCCT CCCATCAACC TCTTTTGTGC
CTCCCCCTCC GTGTACCACC TTCTCTGTCA CCACCCCTGG CCTCACAACT
CTCTCCTTTG CCAC

>rs2567608.1|Cr.20|longitud=110|alelo=A

CAGTGGCATA GTAGTCCAGG GGCTCCTCCT CAGCACCTCC AGCACCTTCC
AGGAGGCAGC AGCGCAGGCA GAGAACCCGC TGGAAGAATC GGCGGAAGTT
GTCGGAGAGG

>rs2567608.2|Cr.20|longitud=110|alelo=A

CAGTGGCATA GTAGTCCAGG GGCTCCTCCT CAGCACCTCC AGCACCTTCC
AGGAGGCAGC AGCGCAGGCA GAGAACCCGC TGGAAGGATC GGCGGAAGTT
GTCGGAGAGG

>rs2073383.1|Cr.22|longitud=140|alelo=C

GCTGCAGAAT CCACAGAGCC AGACGCCCCC TGGGCCCCCA GCGCCCCCCT
GCACAAGTGG GGAAACTAGG TCATGGGGCC CAGGCAGTGT GGAAGGCGTT
GCAGGAGTTG CCCAGGGCGT GGGGTCCTCC AGCCTCAGTG

>rs2073383.2|Cr.22|longitud=140|alelo=T

GCTGCAGAAT CCACAGAGCC AGACGCCCCC TGGGCCCCCA GCGCCCCCCT
GCACAAGTGG GGAAACTAGG TCATGGGGCC CAGGCAGTGT GGAAGGCGTT
GCAGGAGTTG CCCAGGGTGT GGGGTCCTCC AGCCTCAGTG

LISTADO DE SECUENCIAS

<110> RAVA, RICHARD P.
RHEES, BRIAN K.
BURKE, JOHN P.

<120> RESOLUCIÓN DE FRACCIONES DE GENOMA MEDIANTE RECUENTO DE POLIMORFISMOS

<130> ARTEP002WO

<140> PCT/US2012/033391
<141> 12-04-2012

<150> 61/474,362
<151> 12-04-2011

<160> 32

<170> PatentIn versión 3.5

<210> 1
<211> 111
<212> ADN
<213> *Homo sapiens*

<400> 1

cacatgcaca gccagcaacc ctgtcagcag gagttcccac cagtttcttt ctgagaacat

ctgttcaggt ttctctccat ctctatttac tcaggtcaca ggaccttggg g

<210> 2
<211> 111
<212> ADN
<213> *Homo sapiens*

<400> 2

cacatgcaca gccagcaacc ctgtcagcag gagttcccac cagtttcttt ctgagaacat

ctgttcaggt ttctctccat ctctgtttac tcaggtcaca ggaccttggg g

<210> 3
<211> 126
<212> ADN
<213> *Homo sapiens*

<400> 3

tgaggaagtg aggctcagag ggtaagaaac tttgtcacag agctggtggt gagggtggag

attttacact ccctgcctcc cacaccagtt tctccagagt ggaaagactt tcatctcgca

ctggca

<210> 4
<211> 126
<212> ADN
<213> *Homo sapiens*

<400> 4

tgaggaagtg aggctcagag ggtaagaaac tttgtcacag agctggtggt gagggtggag 60

attttacact ccctgcctcc cacaccagtt tctccggagt ggaaagactt tcatctcgca 120

ctggca 126

<210> 5
<211> 121
<212> ADN
<213> *Homo sapiens*

<400> 5

gtgccttcag aacctttgag atctgattct atttttaaag cttcttagaa gagagattgc 60

aaagtgggtt gtttctctag ccagacaggg caggcaaata ggggtggctg gtgggatggg 120

a 121

<210> 6
<211> 121
<212> ADN
<213> *Homo sapiens*

<400> 6

gtgccttcag aacctttgag atctgattct atttttaaag cttcttagaa gagagattgc 60

aaagtgggtt gtttctctag ccagacaggg caggtaaata ggggtggctg gtgggatggg 120

a 121

<210> 7
<211> 111
<212> ADN
<213> *Homo sapiens*

<400> 7

aggtgtgtct ctcttttgtg aggggagggg tcccttctgg cctagtagag ggcctggcct 60

gcagtgagca ttcaaatcct caaggaacag ggtggggagg tgggacaaag g 111

<210> 8
<211> 111
<212> ADN
<213> *Homo sapiens*

<400> 8

aggtgtgtct ctcttttgtg aggggagggg tcccttctgg cctagtagag ggcctggcct 60

gcagtgagca ttcaaatcct cgaggaacag ggtggggagg tgggacaaag g 111

<210> 9
<211> 139
<212> ADN
<213> *Homo sapiens*

<400> 9

cctcgcctac tgtgctgttt ctaaccatca tgcttttccc tgaatctctt gagtcttttt 60

ctgctgtgga ctgaaacttg atcctgagat tcacctctag tccctctgag cagcctcctg 120

gaatactcag ctgggatgg 139

<210> 10
<211> 139
<212> ADN
<213> *Homo sapiens*

<400> 10

cctcgcctac tgtgctgttt ctaaccatca tgcttttccc tgaatctctt gagtcttttt 60

ctgctgtgga ctgaaacttg atcctgagat tcacctctag tccctctggg cagcctcctg 120

gaatactcag ctgggatgg 139

<210> 11
<211> 114
<212> ADN
<213> *Homo sapiens*

<400> 11

gcaactccct caactccaag gcagacacca aagccctccc tgcctgtggc tttgtagttc 60

tagtgtggga tctgactccc cacagcccac ccaaagccgg ggaactcctc actg 114

<210> 12
<211> 114
<212> ADN
<213> *Homo sapiens*

<400> 12

gcaactccct caactccaag gcagacacca aagccctccc tgcctgtggc tttgtagttc 60

tagtgtggga tctgactccc cacagcctac ccaaagccgg ggaactcctc actg 114

<210> 13
<211> 117
<212> ADN
<213> *Homo sapiens*

<400> 13

aattgcaatg gtgagaggtt gatggtaaaa tcaaacggaa cttgttattt tgtcattctg 60

atggactgga actgaggatt ttcaatttcc tctccaaccc aagacacttc tcactgg 117

<210> 14
<211> 117
<212> ADN
<213> *Homo sapiens*

<400> 14

aattgcaatg gtgagaggtt gatggtaaaa tcaaacggaa cttgttattt tgtcattctg 60

atggactgga actgaggatt ttcaatttcc tttccaaccc aagacacttc tcactgg 117

<210> 15
<211> 114
<212> ADN
<213> *Homo sapiens*

<400> 15

gaaatgcctt ctcaggtaat ggaaggttat ccaaatattt ttcgtaagta tttcaaatag 60

caatggctcg tctatggtta gtctcacagc cacattctca gaactgctca aacc 114

<210> 16
<211> 114
<212> ADN
<213> *Homo sapiens*

<400> 16

gaaatgcctt ctcaggtaat ggaaggttat ccaaatattt ttcgtaagta tttcaaatag 60

caatggctcg tctatggtta gtctcgcagc cacattctca gaactgctca aacc 114

<210> 17
<211> 128
<212> ADN
<213> *Homo sapiens*

<400> 17

acccaaaaca ctggaggggc ctcttctcat tttcggtaga ctgcaagtgt tagccgtcgg 60

gaccagcttc tgtctggaag ttcgtcaaat tgcagttaag tccaagtatg ccacatagca 120

gataaggg 128

<210> 18
<211> 128
<212> ADN
<213> *Homo sapiens*

<400> 18

acccaaaaca ctggaggggc ctcttctcat tttcggtaga ctgcaagtgt tagccgtcgg 60

gaccagcttc tgtctggaag ttcgtcaaat tgcagttagg tccaagtatg ccacatagca 120

gataaggg 128

<210> 19
<211> 110
<212> ADN
<213> *Homo sapiens*

<400> 19

gcaccagaat ttaaacaacg ctgacaataa atatgcagtc gatgatgact tcccagagct 60

ccagaagcaa ctccagcaca cagagaggcg ctgatgtgcc tgtcaggtgc 110

<210> 20
<211> 110
<212> ADN
<213> *Homo sapiens*

<400> 20

gcaccagaat ttaaacaacg ctgacaataa atatgcagtc gatgatgact tcccagagct 60

ccagaagcaa ctccagcaca cggagaggcg ctgatgtgcc tgtcaggtgc 110

<210> 21
<211> 139
<212> ADN
<213> *Homo sapiens*

<400> 21

gcccagatta gatggaacct tttcctcttt tccagtgcaa gacaagcgat tgaaagaagt 60

ggatgtgtta ttgcgggcac aatggagcca ctgaactgca gtgcaaaaat gcagtaaggc 120

atacagatag aagaaggag 139

<210> 22
<211> 139
<212> ADN
<213> *Homo sapiens*

<400> 22

gcccagatta gatggaacct tttcctcttt tccagtgcaa gacaagcgat tgaaagaagt 60

ggatgtgtta ttgcgggcac aatggagcca ctgaactgca gtgcaaaaat gcagtaaggg 120

atacagatag aagaaggag 139

<210> 23
<211> 116
<212> ADN
<213> *Homo sapiens*

<400> 23

tgactgtata ccccaggtgc acccttgggt catctctatc atagaactta tctcacagag 60

tataagagct gatttctgtg tctgcctctc acactagact tccacatcct tagtgc 116

<210> 24
<211> 116
<212> ADN
<213> *Homo sapiens*

<400> 24

tgactgtata ccccaggtgc acccttgggt catctctatc atagaactta tctcacagag 60

tataagagct gatttctgtg tctgcctgtc acactagact tccacatcct tagtgc 116

<210> 25
<211> 110
<212> ADN
<213> *Homo sapiens*

<400> 25

tgtacgtggt caccagggga cgcctggcgc tgcgagggag gccccgagcc tcgtgccccc 60

gtgaagcttc agctcccctc cccggctgtc cttgaggctc ttctcacact 110

<210> 26
<211> 110
<212> ADN
<213> *Homo sapiens*

<400> 26

tgtacgtggt caccagggga cgcctggcgc tgcgagggag gccccgagcc tcgtgccccc 60

gtgaagcttc agctcccctc cctggctgtc cttgaggctc ttctcacact 110

<210> 27
<211> 114
<212> ADN
<213> *Homo sapiens*

<400> 27

cagtggaccc tgctgcacct ttcctcccct cccatcaacc tcttttgtgc ctccccctcc 60

gtgtaccacc ttctctgtca ccaaccctgg cctcacaact ctctcctttg ccac 114

<210> 28
<211> 114
<212> ADN
<213> *Homo sapiens*

<400> 28

cagtggaccc tgctgcacct ttcctcccct cccatcaacc tcttttgtgc ctccccctcc 60

gtgtaccacc ttctctgtca ccacccctgg cctcacaact ctctcctttg ccac 114

<210> 29
<211> 110
<212> ADN
<213> *Homo sapiens*

<400> 29

cagtggcata gtagtccagg ggctcctcct cagcacctcc agcaccttcc aggaggcagc 60

agcgcaggca gagaacccgc tggaagaatc ggcggaagtt gtcggagagg 110

<210> 30
<211> 110
<212> ADN
<213> *Homo sapiens*

<400> 30

cagtggcata gtagtccagg ggctcctcct cagcacctcc agcaccttcc aggaggcagc 60

agcgcaggca gagaacccgc tggaaggatc ggcggaagtt gtcggagagg 110

<210> 31
<211> 140
<212> ADN
<213> *Homo sapiens*

<400> 31

gctgcagaat ccacagagcc agacgccccc tgggccccca gcgcccccct gcacaagtgg 60

ggaaactagg tcatggggcc caggcagtgt ggaaggcgtt gcaggagttg cccagggcgt 120

ggggtcctcc agcctcagtg 140

<210> 32
<211> 140
<212> ADN
<213> *Homo sapiens*

<400> 32

gctgcagaat ccacagagcc agacgccccc tgggccccca gcgcccccct gcacaagtgg 60

ggaaactagg tcatggggcc caggcagtgt ggaaggcgtt gcaggagttg cccagggtgt 120

ggggtcctcc agcctcagtg 140

Reivindicaciones

1. Método de estimación de la fracción de ADN fetal en el ADN obtenido a partir de un fluido corporal de una embarazada, comprendiendo el método:

(a) recibir una muestra del fluido corporal;
(b) extraer ADN de la muestra en condiciones que extraen el ADN tanto de un genoma materno como de un genoma fetal presente en el fluido corporal;
(c) secuenciar el ADN extraído con un secuenciador de ácidos nucleicos en condiciones que producen secuencias de segmentos de ADN que contienen uno o más polimorfismos;
(d) mapear las secuencias de segmentos de ADN derivadas de la secuenciación del ADN en el fluido corporal contra uno o más polimorfismos designados en una secuencia de referencia, en el que el mapeo se realiza utilizando un aparato computacional programado para mapear secuencias de ácidos nucleicos contra el uno o más polimorfismos designados;
(e) determinar las frecuencias alélicas de las secuencias de segmentos de ADN mapeadas para al menos uno de los polimorfismos designados;
(f) clasificar el al menos un polimorfismo designado basándose en una combinación de la cigosidad de la embarazada y la cigosidad del feto; y
(g) estimar la fracción de ADN fetal en el ADN obtenido de la embarazada utilizando las frecuencias alélicas determinadas en (e), junto con la clasificación de cigosidades de (f),

en el que (e)-(g) se realizan en uno o más procesadores que ejecutan las instrucciones de un programa para la determinación, la clasificación y la estimación,
en el que la secuenciación de la etapa (c) produce secuencias de segmentos de ADN que contienen una pluralidad de polimorfismos y las etapas (d)-(f) se realizan basándose en la pluralidad de polimorfismos; y
en el que la clasificación en (f) clasifica cada una de la pluralidad de polimorfismos en una de las siguientes combinaciones: (i) la embarazada es homocigoto y el feto es homocigoto, (ii) la embarazada es homocigoto y el feto es heterocigoto, (iii) la embarazada es heterocigoto y el feto es homocigoto, y (iv) la embarazada es heterocigoto y el feto es heterocigoto.

2. Método según la reivindicación 1, en el que la clasificación de la pluralidad de polimorfismos comprende aplicar los datos de frecuencia alélica de (e), obtenidos para la pluralidad de polimorfismos, a un modelo de mezcla.

3. Método según la reivindicación 2, que comprende adicionalmente resolver una serie de ecuaciones para los momentos factoriales de los datos de frecuencia alélica para cada una de la pluralidad de secuencias de polimorfismos, siendo la serie de ecuaciones:

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i}{d_i}$$

$$F_2 = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i(a_i-1)}{d_i(d_i-1)}$$

$$\cdots$$

$$F_j = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i(a_i-1)\cdots(a_i-j+1)}{d_i(d_i-1)(d_i-j+1)}$$

en las que F_j es el j-ésimo momento factorial, a_i es la frecuencia del alelo minoritario del i-ésimo sitio de polimorfismo, d_i es la cobertura del i-ésimo sitio de polimorfismo, y n es el número de sitios de polimorfismo.

4. Método según la reivindicación 2 ó 3, que comprende adicionalmente relacionar los momentos factoriales con $\{\alpha_i, p_i\}$ mediante

$$F_1 \approx \sum_{i=1}^{m} \alpha_i p_i^1$$

$$F_2 \approx \sum_{i=1}^{m} \alpha_i p_i^2$$

...

$$F_j \approx \sum_{i=1}^{m} \alpha_i p_i^j$$

...

$$F_g \approx \sum_{i=1}^{m} \alpha_i p_i^g$$

en las que α_i es la porción de la pluralidad de secuencias de polimorfismos en el caso de cigosidad *i*, p_i es la probabilidad binomial de alelo minoritario en el caso de cigosidad *i*, *m* es el número de casos de cigosidad, y g es el número de momentos factoriales calculados.

5. Método según la reivindicación 3 ó 4, que comprende adicionalmente, antes de resolver una serie de ecuaciones para los momentos factoriales, eliminar computacionalmente las frecuencias alélicas para los polimorfismos identificados como:

(A) heterocigoto tanto en el feto como en la embarazada;
(B) homocigoto tanto en el feto como en la embarazada; o
(C) heterocigoto en la embarazada.

6. Método según cualquiera de las reivindicaciones 2-5, en el que el modelo de mezcla tiene en cuenta el error de secuenciación.

7. Método según la reivindicación 1, que comprende adicionalmente no tener en cuenta ningún polimorfismo clasificado en la combinación (i) o en la combinación (iv).

8. Método según cualquiera de las reivindicaciones anteriores, en el que la clasificación de la pluralidad de polimorfismos comprende aplicar un umbral a la frecuencia alélica determinada en (e).

9. Método según cualquiera de las reivindicaciones anteriores, que comprende adicionalmente filtrar la pluralidad de polimorfismos para no tener en cuenta ningún polimorfismo entre la pluralidad de polimorfismos que tenga una frecuencia del alelo minoritario superior o inferior a un umbral definido.

10. Método según cualquiera de las reivindicaciones anteriores, en el que el mapeo en (d) comprende identificar una pluralidad de secuencias de polimorfismos bialélicos.

11. Método según cualquiera de las reivindicaciones anteriores, en el que la estimación de la fracción de ADN fetal en el ADN en (g) comprende transformar los datos del caso (iii) en datos del caso (ii).

12. Método según la reivindicación 11, en el que:

(a) la transformación de datos del caso (iii) en datos del caso (ii) comprende transformar los datos (D, A) en (D1, A1) como:

$$A1 = 0,5D - A$$

$$D1 = D$$

en las que D es la cobertura del polimorfismo y A es el recuento del alelo minoritario del polimorfismo;
(b) la transformación de datos del caso (iii) en datos del caso (ii) comprende la transformación trigonométrica o el uso de matrices de rotación; o

(c) la fracción de ADN fetal se estima como 2 A/D para el caso (ii) y como 1 - 2 A/D para el caso (iii), donde A es la frecuencia del alelo minoritario y D es la cobertura para el polimorfismo en cuestión.

13. Método según la reivindicación 11 ó 12, que comprende adicionalmente aplicar técnicas de regresión a los datos del caso (ii) y los datos transformados del caso (iii), en el que la fracción de ADN fetal se estima como dos veces la pendiente de la línea de regresión de un modelo de regresión lineal.

14. Método según cualquiera de las reivindicaciones anteriores, en el que:

(a) el ADN obtenido a partir de un fluido corporal de una embarazada es ADN libre obtenido a partir de plasma de la embarazada; o
(b) la secuenciación se lleva a cabo sin amplificar selectivamente ninguno del uno o más polimorfismos designados.

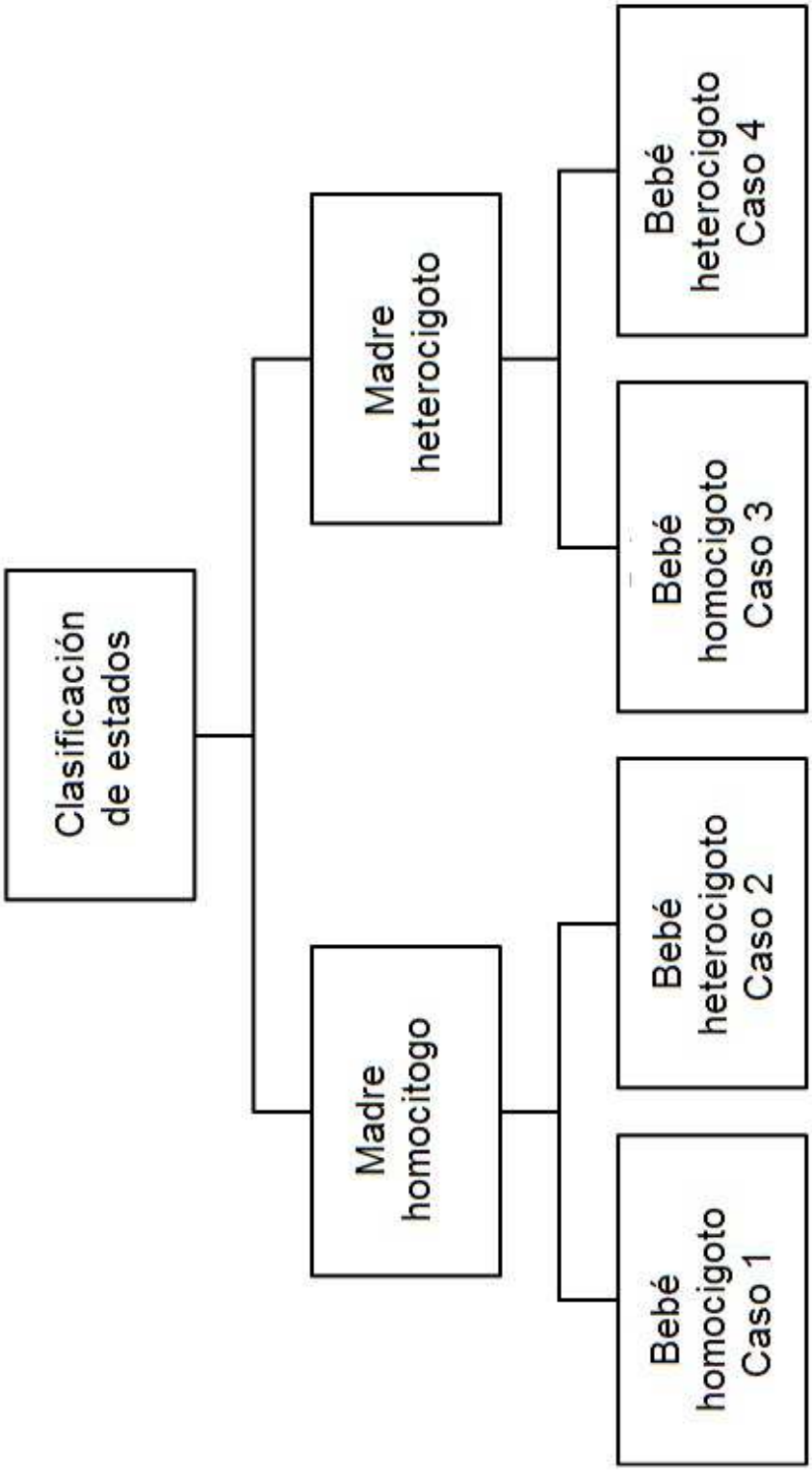
Clasificación de estados
Madre homocitogo
Madre heterocigoto
Bebé homocigoto Caso 1
Bebé heterocigoto Caso 2
Bebé homocigoto Caso 3
Bebé heterocigoto Caso 4

FIG. 1

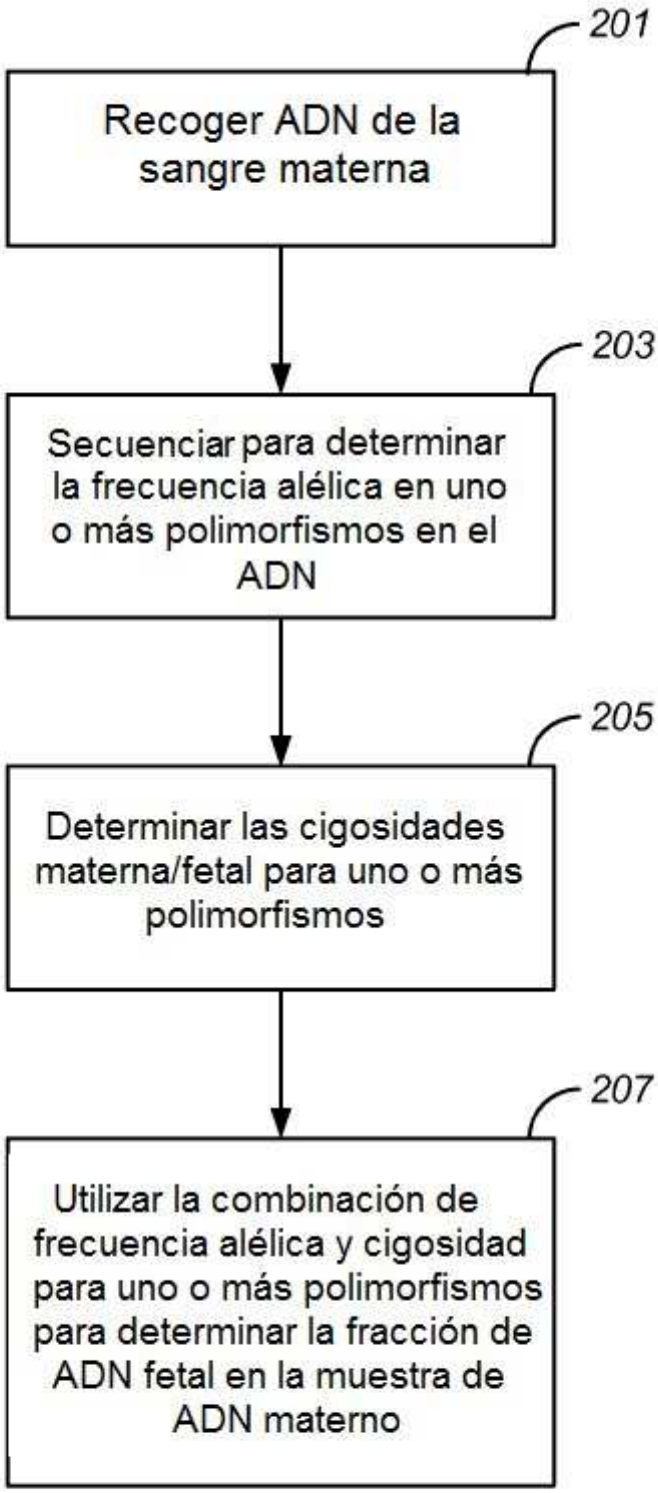
201
Recoger ADN de la sangre materna
203
Secuenciar para determinar la frecuencia alélica en uno o más polimorfismos en el ADN
205
Determinar las cigosidades materna/fetal para uno o más polimorfismos
207
Utilizar la combinación de frecuencia alélica y cigosidad para uno o más polimorfismos para determinar la fracción de ADN fetal en la muestra de ADN materno

FIG. 2

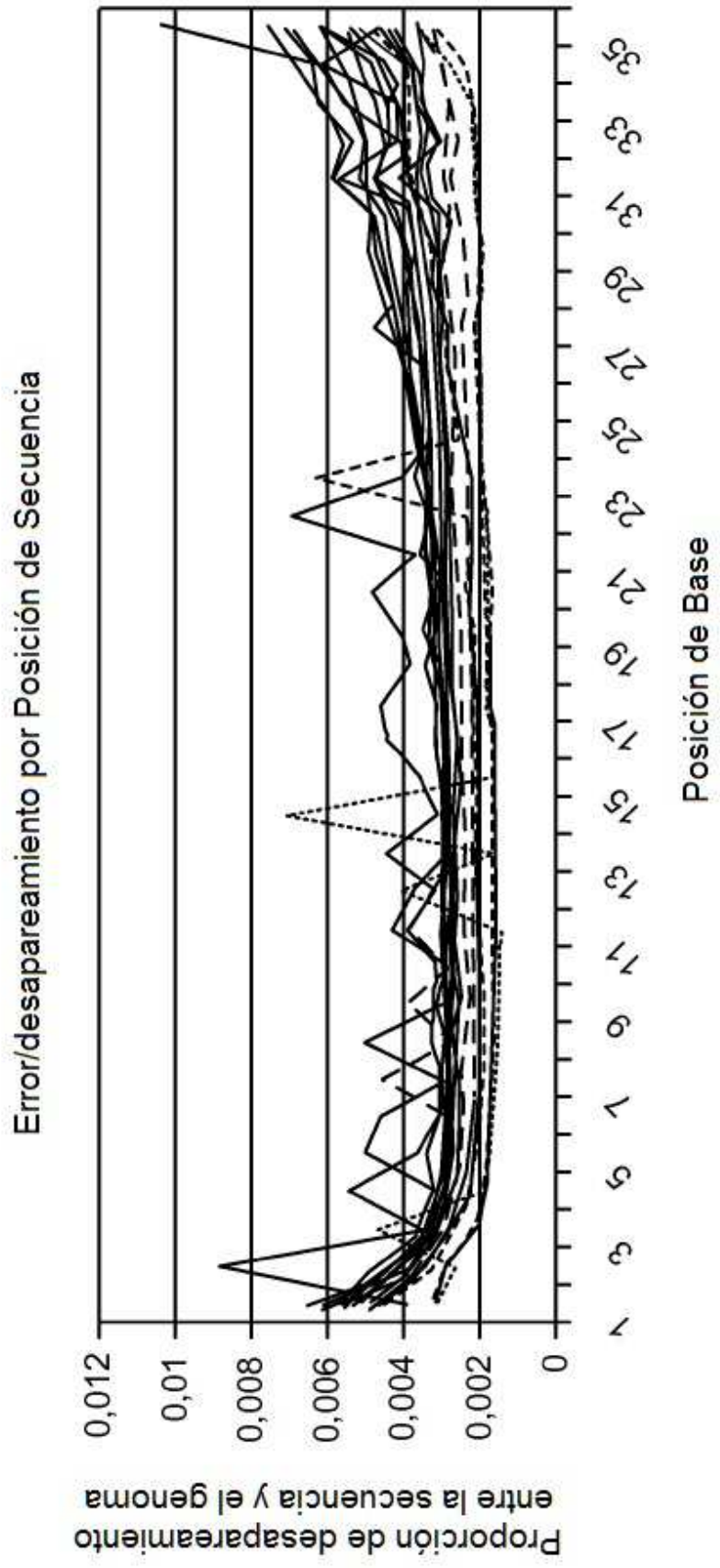
Error/desapareamiento por Posición de Secuencia
Proporción de desapareamiento entre la secuencia y el genoma
0,012
0,01
0,008
0,006
0,004
0,002
0
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
Posición de Base

FIG. 3

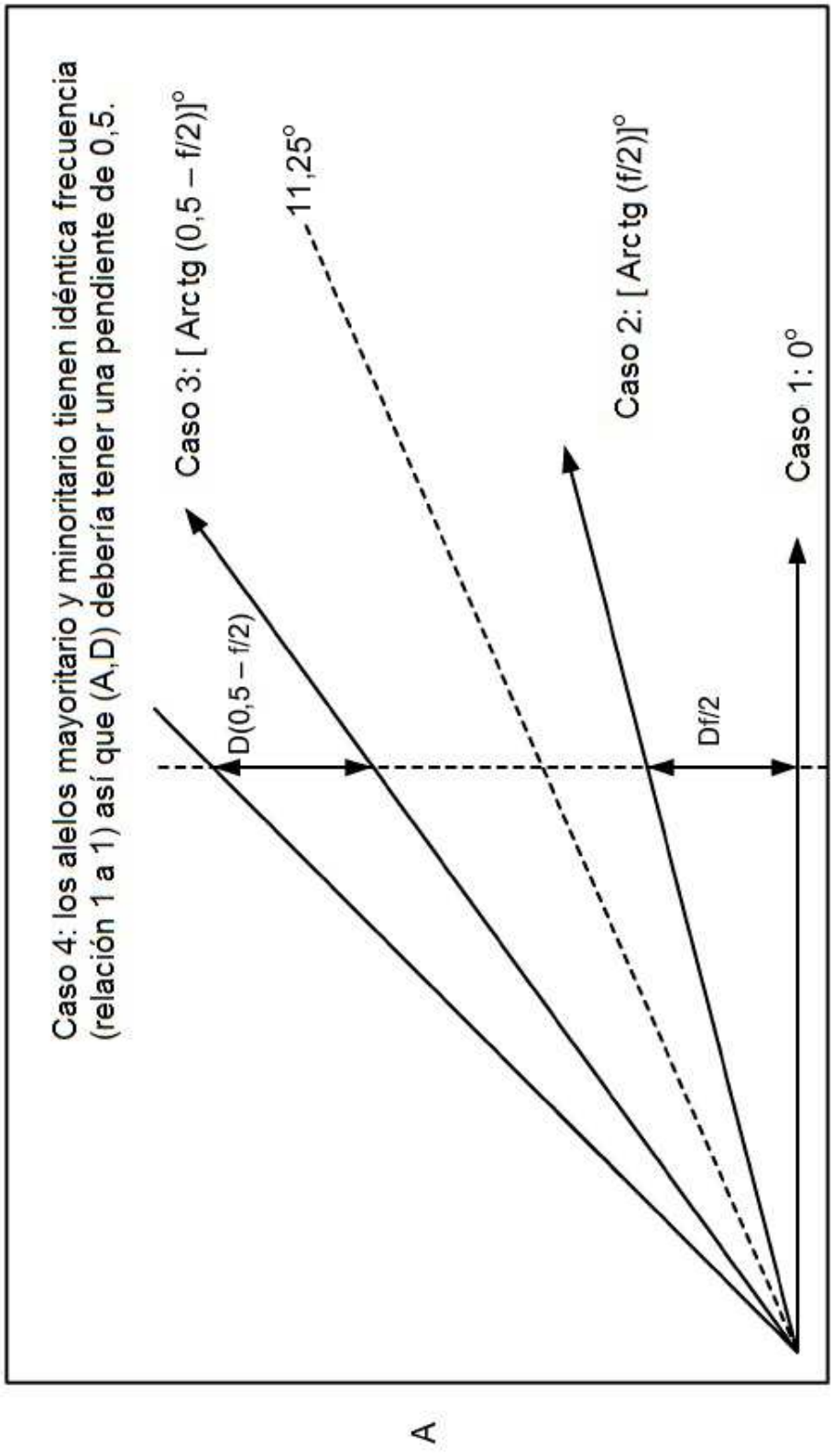
Caso 4: los alelos mayoritario y minoritario tienen idéntica frecuencia (relación 1 a 1) así que (A,D) debería tener una pendiente de 0,5.
Caso 3: [Arc tg (0,5 – f/2)]°
11,25°
Caso 2: [Arc tg (f/2)]°
Caso 1: 0°
D(0,5 – f/2)
Df/2
D=A+B
A

FIG. 4

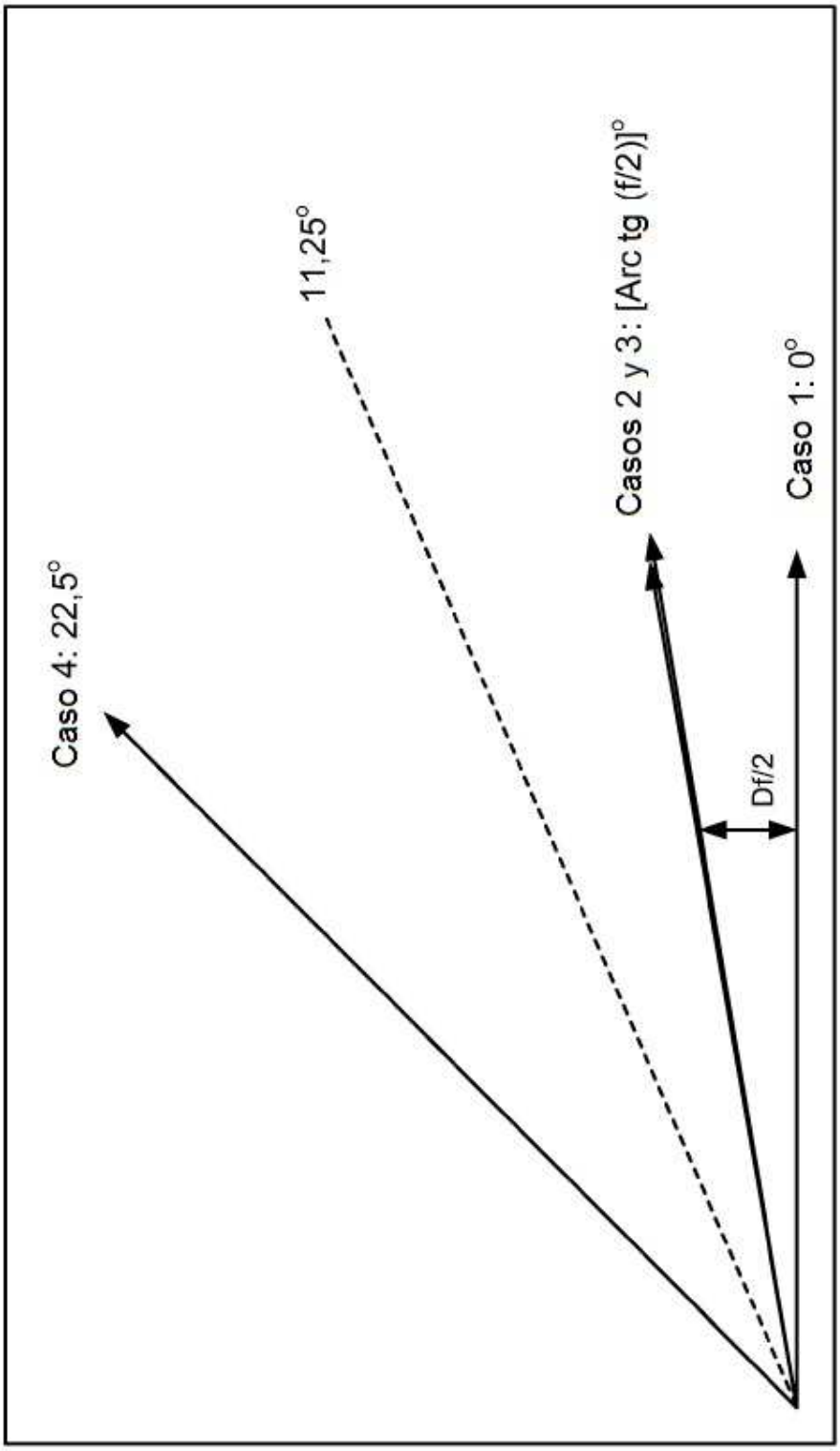

FIG. 5

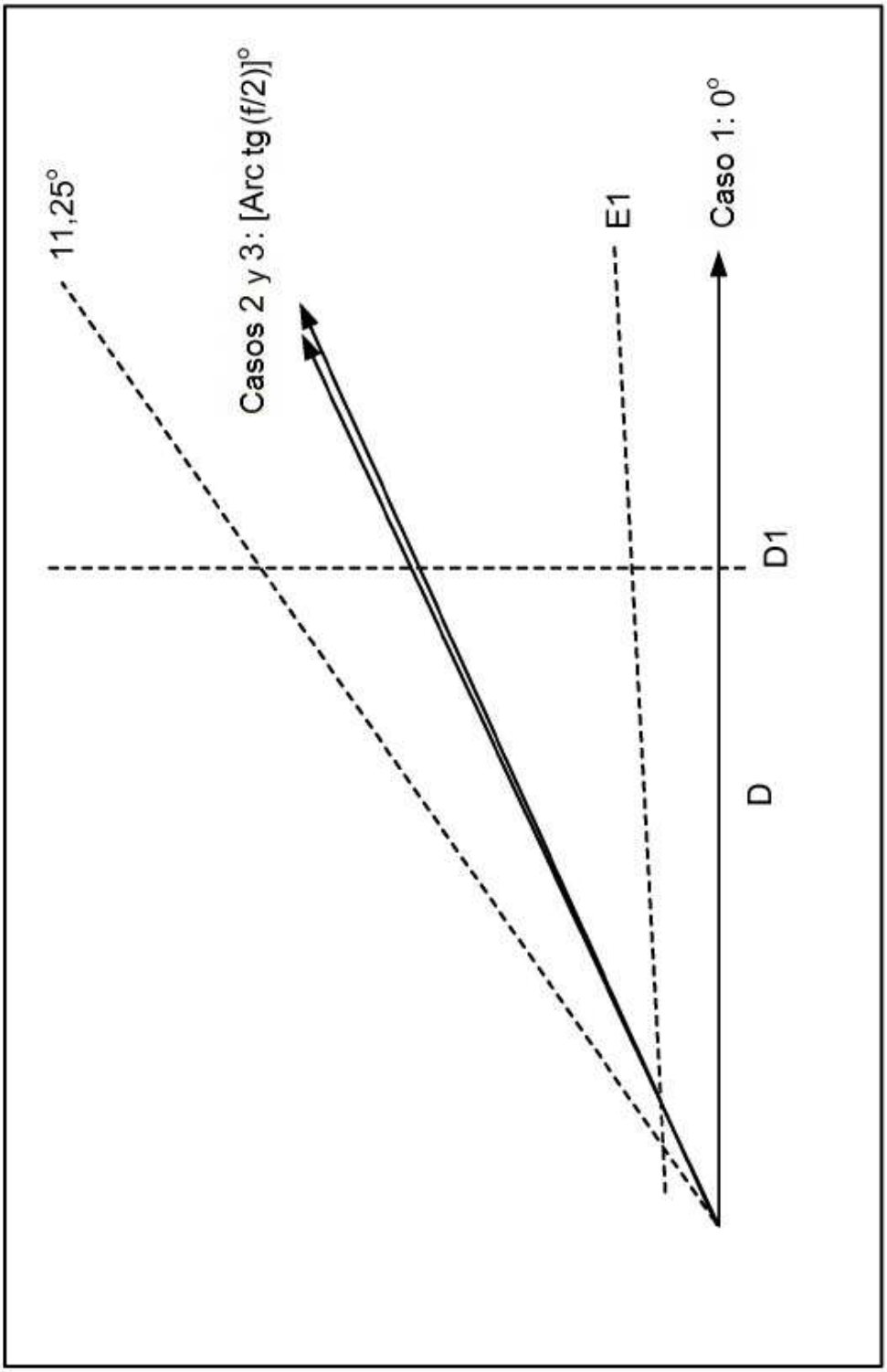

FIG. 6

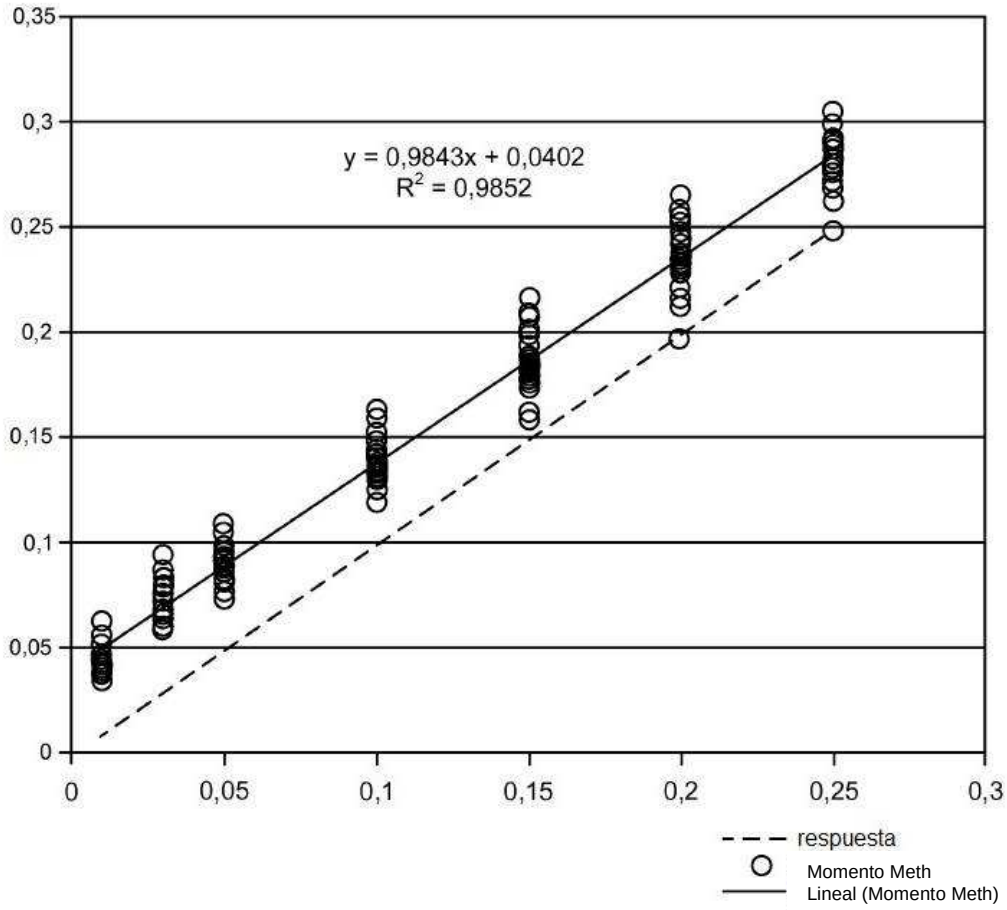
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
y = 0,9843x + 0,0402
R2 = 0,9852
0
0,05
0,1
0,15
0,2
0,25
0,3
respuesta
Momento Meth
Lineal (Momento Meth)

FIG. 7

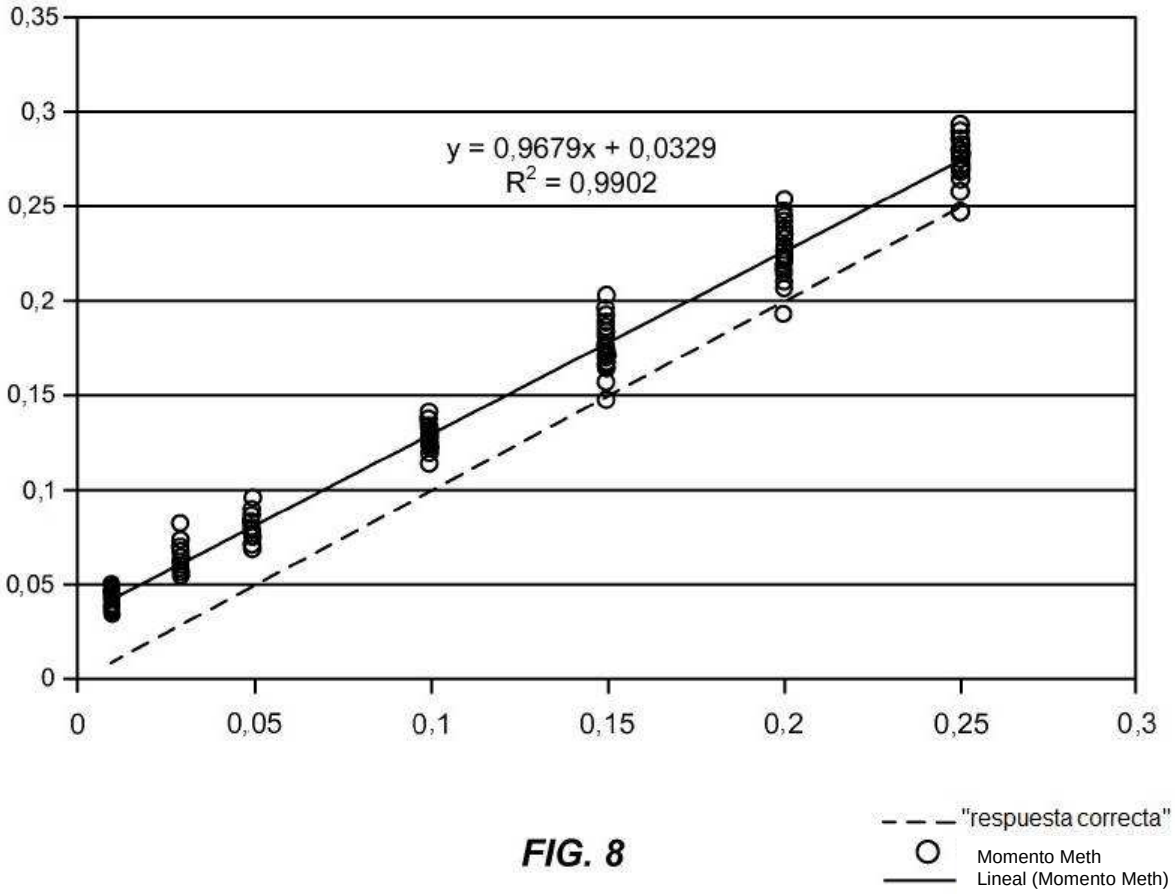
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
y = 0,9679x + 0,0329
R2 = 0,9902
0
0,05
0,1
0,15
0,2
0,25
0,3
"respuesta correcta"
Momento Meth
Lineal (Momento Meth)

FIG. 8

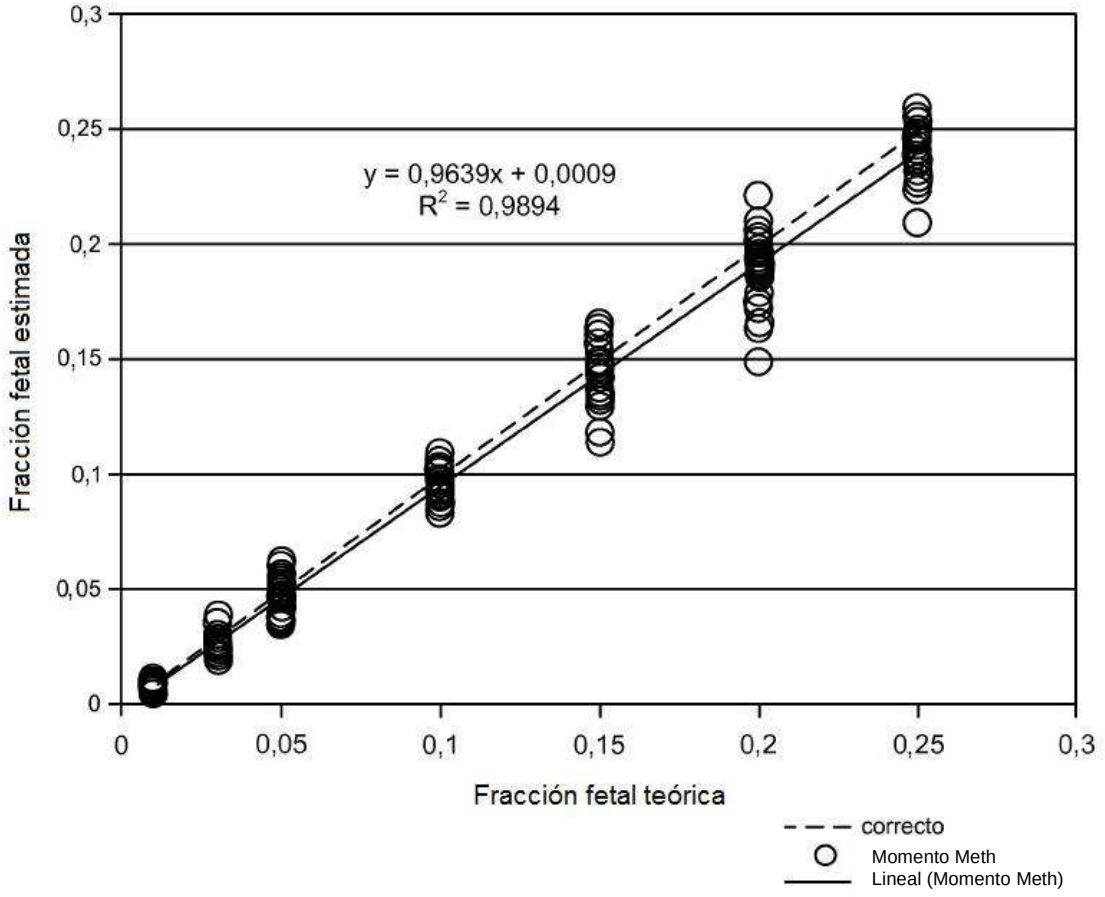
0,3
0,25
0,2
0,15
0,1
0,05
0
y = 0,9639x + 0,0009
R2 = 0,9894
Fracción fetal estimada
0
0,05
0,1
0,15
0,2
0,25
0,3
Fracción fetal teórica
correcto
Momento Meth
Lineal (Momento Meth)

FIG. 9

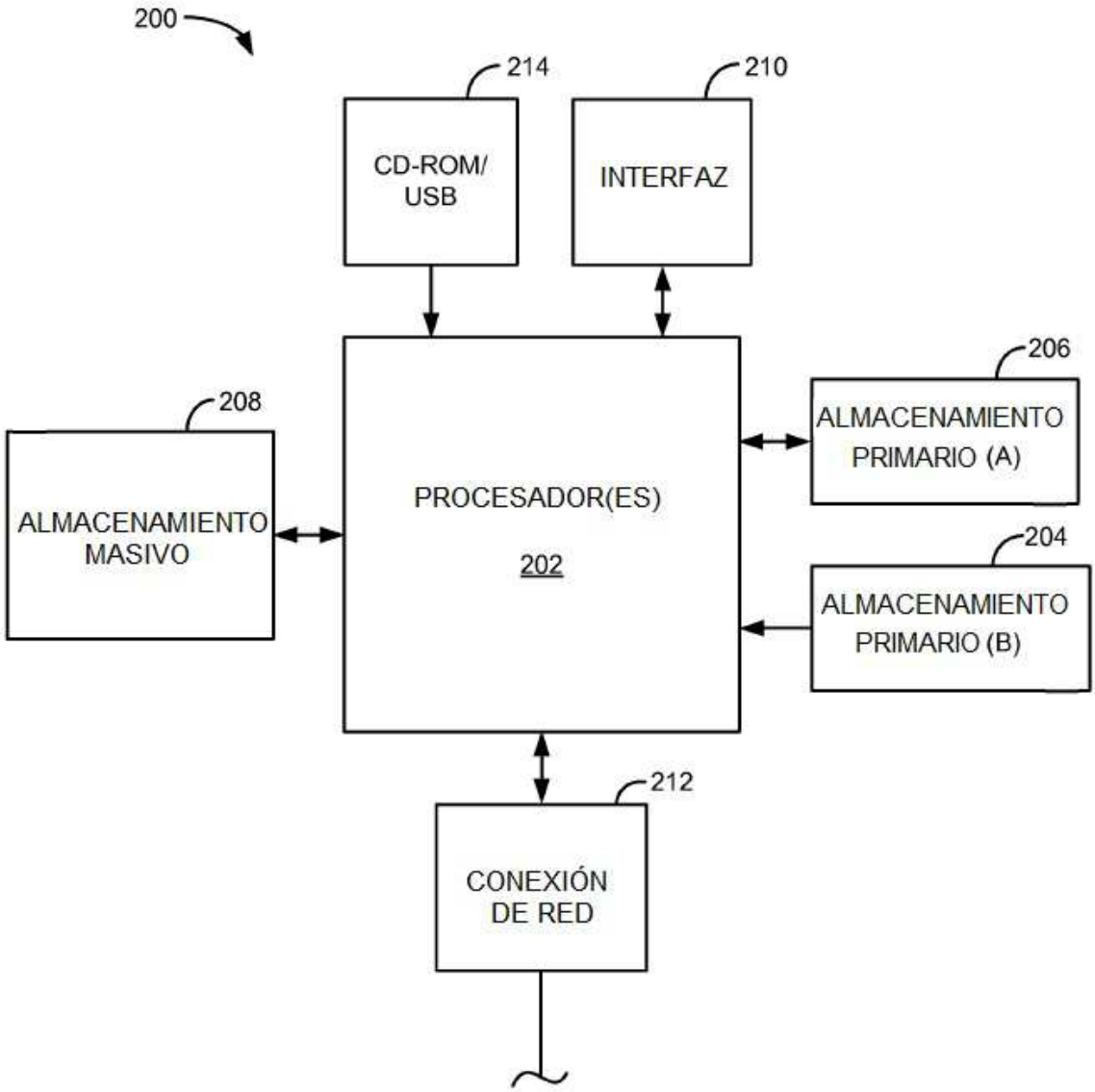
200
214
CD-ROM/
USB
210
INTERFAZ
206
ALMACENAMIENTO
PRIMARIO (A)
208
ALMACENAMIENTO
MASIVO
PROCESADOR(ES)
202
204
ALMACENAMIENTO
PRIMARIO (B)
212
CONEXIÓN
DE RED

FIG. 10

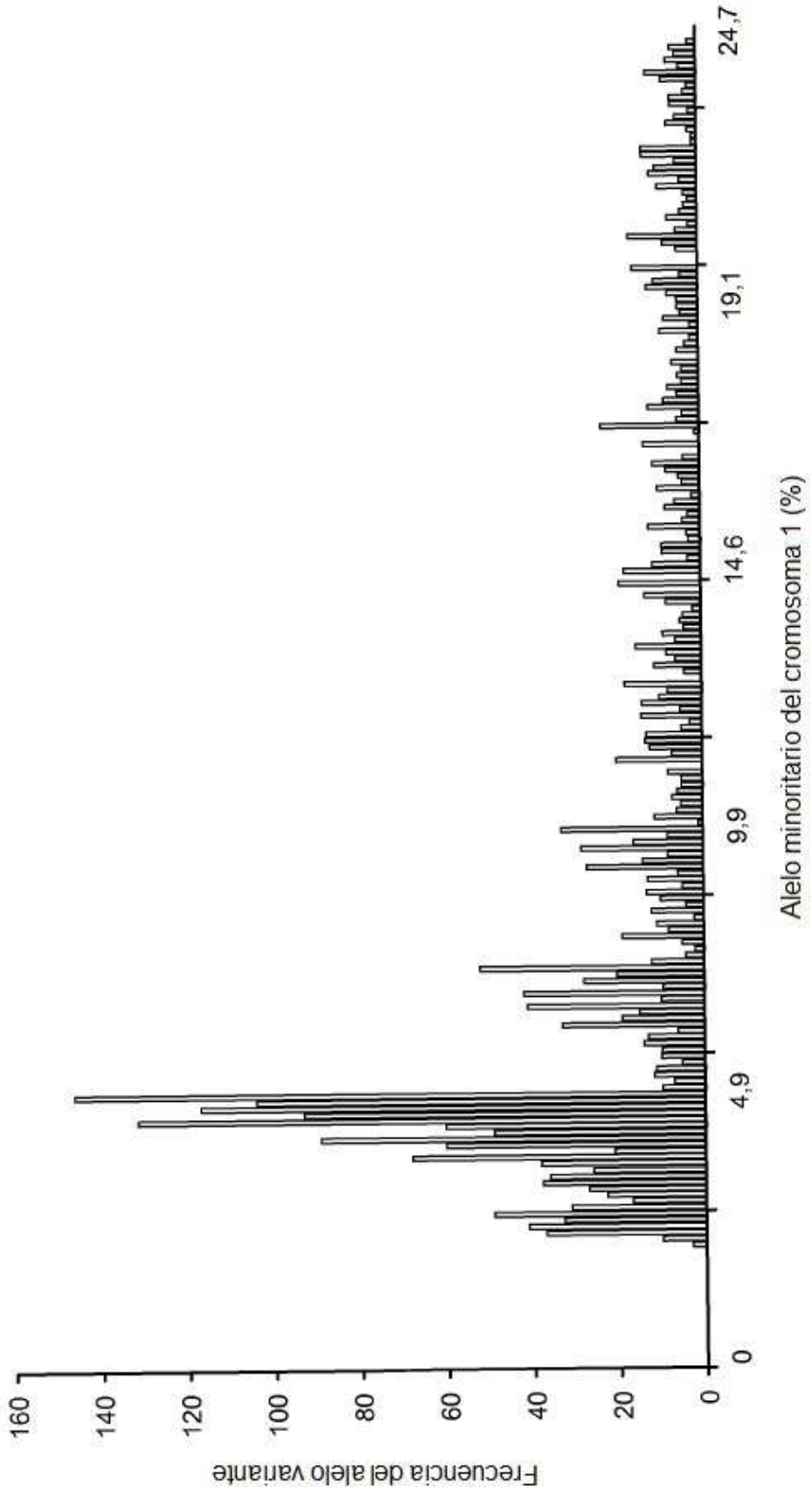
Frecuencia del alelo variante
160
140
120
100
80
60
40
20
0
0
4,9
9,9
14,6
19,1
24,7
Alelo minoritario del cromosoma 1 (%)

FIG. 11A

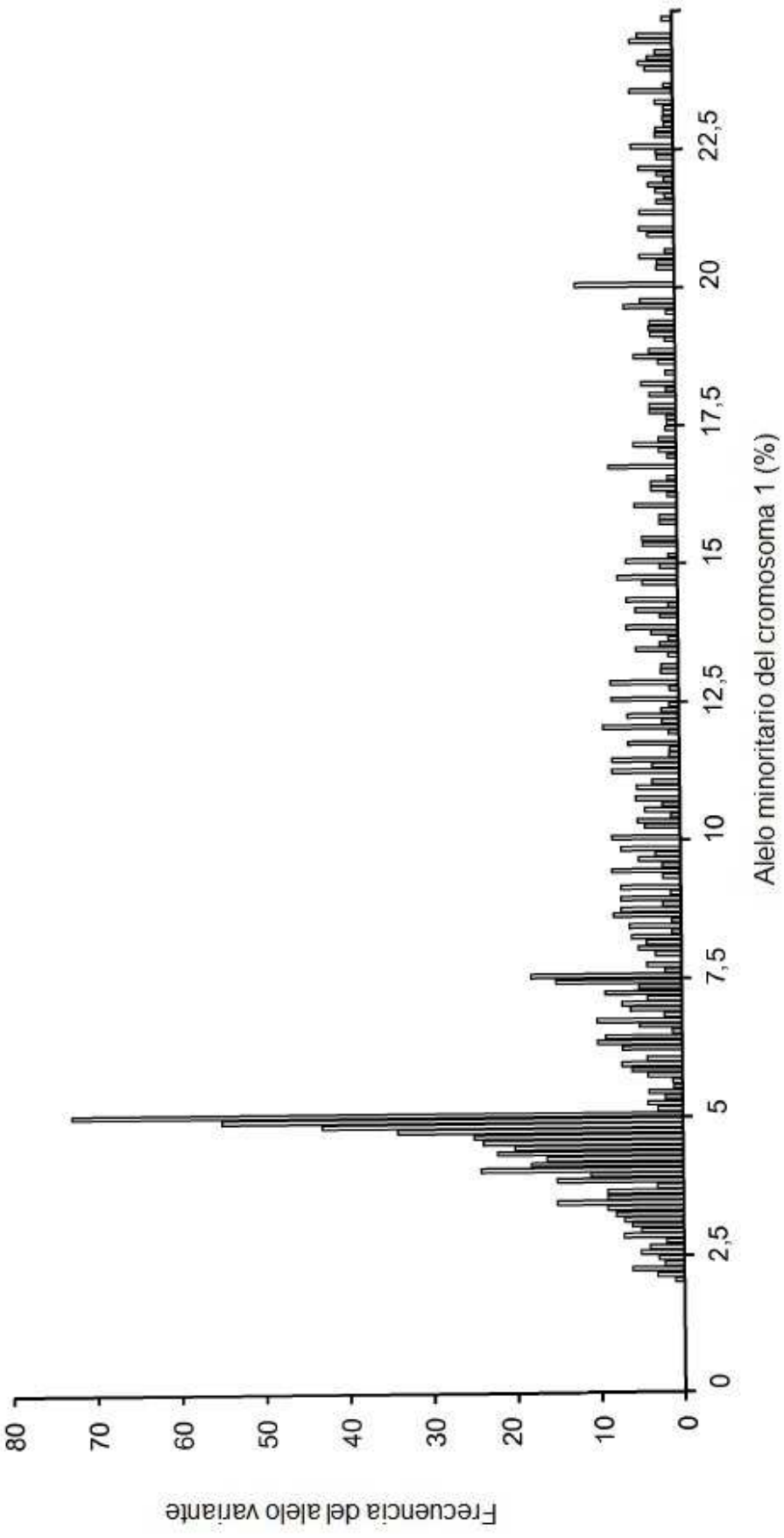
Frecuencia del alelo variante
80
70
60
50
40
30
20
10
0
0
2,5
5
7,5
10
12,5
15
17,5
20
22,5
Alelo minoritario del cromosoma 1 (%)

FIG. 11B

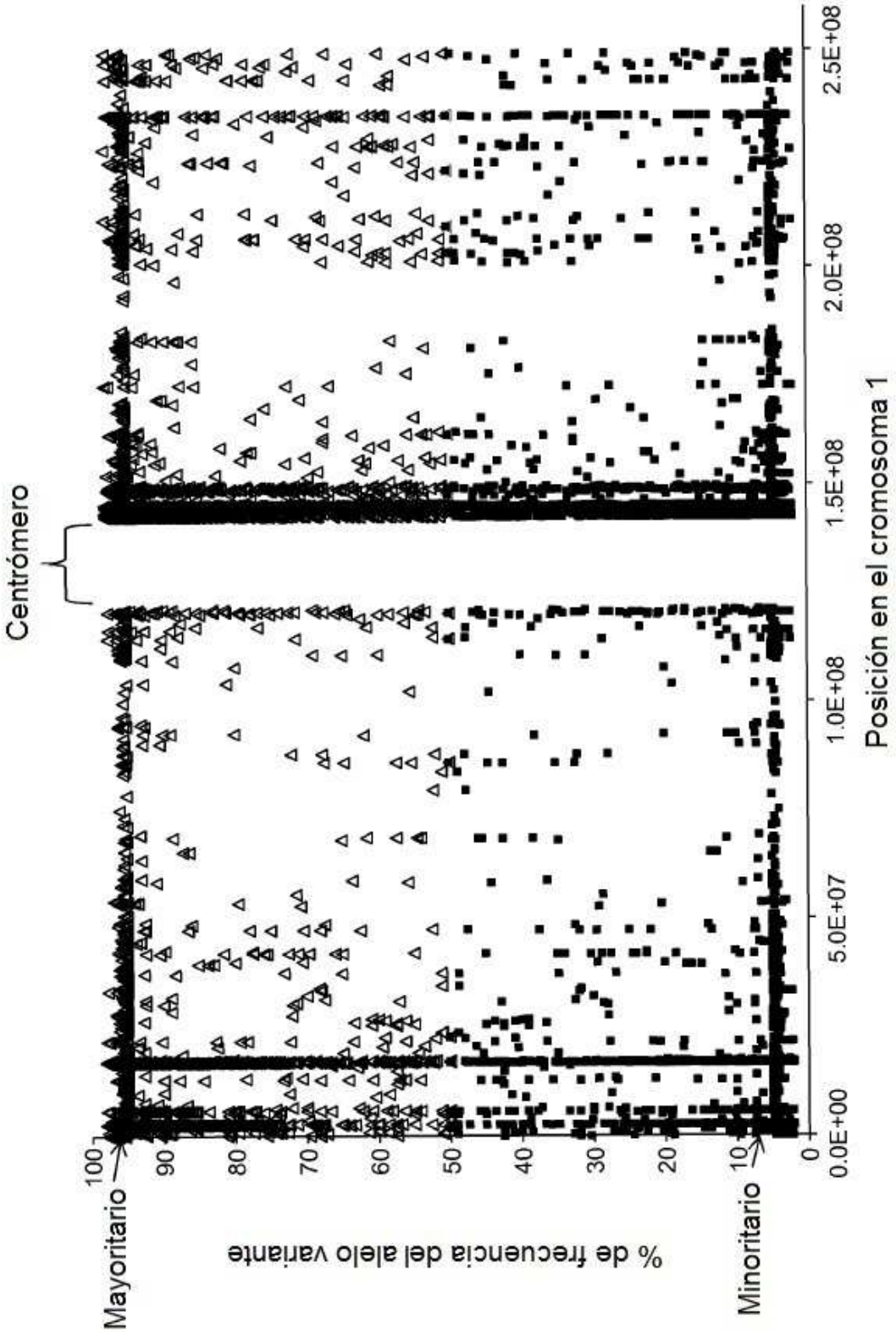
Centrómero
Mayoritario
Minoritario
% de frecuencia del alelo variante
100
90
80
70
60
50
40
30
20
10
0
0.0E+00
5.0E+07
1.0E+08
1.5E+08
2.0E+08
2.5E+08
Posición en el cromosoma 1

FIG. 12A

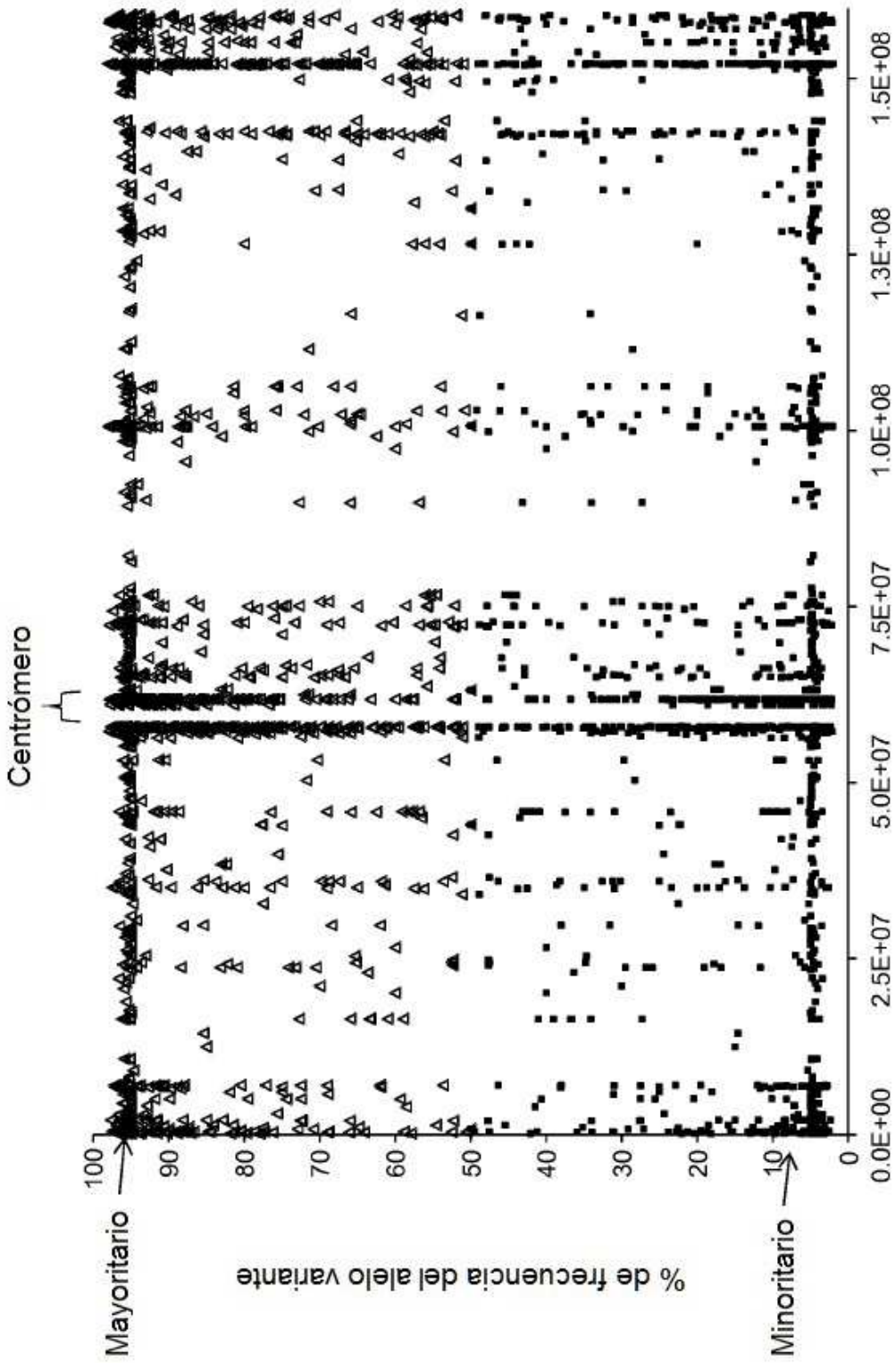
Centrómero
Mayoritario
Minoritario
% de frecuencia del alelo variante
100
90
80
70
60
50
40
30
20
10
0
0.0E+00
2.5E+07
5.0E+07
7.5E+07
1.0E+08
1.3E+08
1.5E+08
Posición en el cromosoma 7

FIG. 12B