(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 907 069**

(51) Int. CI.:

C12Q 1/6809 (2008.01)
C12Q 1/6827 (2008.01)
C12Q 1/6876 (2008.01)
C12Q 1/6883 (2008.01)
G16B 20/00 (2009.01)
G16B 30/00 (2009.01)
G16B 45/00 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **12.04.2012 E 19178858 (7)**

(97) Fecha y número de publicación de la concesión europea: **15.12.2021 EP 3567124**

(54) Título: **Resolución de fracciones genómicas usando recuentos de polimorfismos**

(30) Prioridad:

12.04.2011 US 201161474362 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
21.04.2022

(73) Titular/es:

VERINATA HEALTH, INC. (100.0%)
5200 Illumina Way
San Diego, California 92122, US

(72) Inventor/es:

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

(74) Agente/Representante:

LEHMANN NOVO, María Isabel

DESCRIPCIÓN

Resolución de fracciones genómicas usando recuentos de polimorfismos

ANTECEDENTES

El descubrimiento de ADN fetal que fluye libremente (a veces denominado "ADN libre de células" o "ADNcf") en sangre materna permite la posibilidad de detectar una anormalidad, aneuploidía, y una aberración cromosómicas a partir de muestras de sangre. La abundancia fraccionaria de ADN fetal en plasma sanguíneo materno no es constante y varía con una variedad de factores incluyendo el manejo de muestras y la edad gestacional.

Cuando se usa secuenciación de 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, el poder estadístico (la probabilidad de identificar casos anómalos, o la sensibilidad) se puede calcular mediante métodos de permutación o a través de la integración de combinaciones o convoluciones lineales de distribuciones F no centrales desde alfa al infinito, donde el punto crítico alfa para la significación (probabilidad máxima de considerar falsamente una anomalía) de la población de puntuaciones bajo la hipótesis nula de ausencia de aberración.

El documento US 7.332.277 muestra un método para detectar la presencia de ausencia de una anormalidad cromosómica fetal al cuantificar la relación de la cantidad relativa de alelos en un locus heterocigótico de interés.

Una desventaja de los métodos existentes para detectar la fracción fetal es que responden a medidas de la abundancia de cromosomas sexuales (que solo se pueden usar para medir fiablemente la abundancia relativa de ADN embrionario masculino) o la secuencia de ARNm de genes que se sabe que se expresan diferencialmente entre tejido grávido y embrionario (que está sometido a variabilidad de expresión debido a la edad gestacional u otros factores).

La estimación de la fracción fetal puede ser difícil debido a varios factores perjudiciales incluyendo: parámetros genéticos de población diferenciales étnicos parentales y errores de secuenciación. Por lo tanto, es deseable tener métodos robustos en presencia de estos y otros factores de confusión presentes comúnmente.

SUMARIO

La invención proporciona un método para estimar la fracción de ADN fetal en ADN obtenido de un líquido corporal de una embarazada, comprendiendo el método:

(a) alinear o asignar de otro modo secuencias de segmentos de ADN derivadas de la secuenciación del ADN del líquido corporal a uno o más polimorfismos indicados en una secuencia de referencia, en donde el alineamiento o la asignación de otro modo se realiza usando un aparato informático programado para asignar secuencias de ácido nucleico a los uno o más polimorfismos indicados;

(b) determinar frecuencias alélicas de las secuencias de segmentos de ADN asignadas a al menos uno de los polimorfismos indicados;

(c) clasificar el al menos un polimorfismo indicado basándose en una combinación de la cigosidad de la embarazada y la cigosidad del feto; y

(d) estimar la fracción del ADN fetal en el ADN obtenido de la embarazada usando las frecuencias alélicas determinadas en (b) junto con la clasificación de las cigosidades procedente de (c),

en donde (b)-(d) se realizan en uno o más procesadores que funcionan bajo las instrucciones del programa para la determinación, la clasificación y la estimación; y

en donde la clasificación en (c) clasifica el al menos un polimorfismo indicado en una de las siguientes combinaciones: (i) la embarazada es homocigótica y el feto es homocigótico, (ii) la embarazada es homocigótica y el feto es heterocigótico, (iii) la embarazada es heterocigótica y el feto es homocigótico, y (iv) la embarazada es heterocigótica y el feto es heterocigótico.

La invención proporciona además un aparato para estimar la fracción de ADN fetal en ADN obtenido de un líquido corporal de una embarazada, comprendiendo el aparato:

(a) un secuenciador configurado para (i) recibir ADN extraído de una muestra del líquido corporal que comprende ADN tanto de un genoma materno como de un genoma fetal, y (ii) secuenciar el ADN extraído bajo condiciones que produzcan secuencias de segmentos de ADN que contienen uno o más polimorfismos indicados; y

(b) un aparato informático configurado para dar instrucciones a uno o más procesadores para

alinear o asignar de otro modo secuencias de ácido nucleico a los uno o más polimorfismos indicados en una secuencia de referencia,

determinar las frecuencias alélicas de las secuencias de segmentos de ADN asignadas a al menos uno de los polimorfismos indicados,

clasificar el al menos un polimorfismo indicado basándose en una combinación de la cigosidad de la embarazada y la cigosidad del feto, y

estimar la fracción de ADN fetal en el ADN obtenido de la embarazada usando las frecuencias alélicas junto con la clasificación de cigosidades; y

en donde el aparato informático está configurado además para dar instrucciones a uno o más procesadores para clasificar el al menos un polimorfismo indicado en una de las siguientes combinaciones: (i) la embarazada es homocigótica y el feto es homocigótico, (ii) la embarazada es homocigótica y el feto es heterocigótico, (iii) la embarazada es heterocigótica y el feto es homocigótico, y (iv) la embarazada es heterocigótica y el feto es heterocigótico.

Ciertas realizaciones divulgadas se refieren a métodos informáticos para medir fiablemente la abundancia relativa de ADN fetal que flota libremente al secuenciar una muestra de sangre materna.

En realizaciones específicas, la invención que se define mediante las reivindicaciones adjuntas proporciona métodos para estimar fiablemente la fracción fetal a partir de polimorfismos tales como pequeñas variaciones o inserciones-eliminaciones de bases que son robustas con respecto a la etnicidad parental, el sexo del embrión, la edad gestacional y otros factores ambientales. Muchos ejemplos divulgados en la presente emplean SNPs como el polimorfismo pertinente. La invención se puede aplicar como parte de un estudio de resecuenciación prediseñado intencionada dirigido contra polimorfismos conocidos o se puede usar en un análisis retrospectivo de variaciones encontradas mediante coincidencia en secuencias solapadas generadas a partir de plasma materno (o cualquier otro entorno en el que esté presente una mezcla de ADN procedente de varias personas).

Este documento presenta técnicas para la estimación de la abundancia fraccionaria de ADN fetal en muestras de sangre materna. Ciertas técnicas divulgadas usan las frecuencias alélicas observadas de SNPs encontrados por casualidad o encontrados en grupos de SNPs conocidos previamente diseñados con el propósito de estimar la fracción fetal.

Las técnicas y los aparatos descritos en la presente se pueden emplear en muchos casos para estimar la fracción de ácido nucleico procedente de un genoma en una mezcla de dos genomas, que pueden estar relacionados como genomas de la madre y el hijo.

Ciertos aspectos de la divulgación tratan de métodos para estimar la fracción de ADN fetal en ADN obtenido de un líquido corporal de una embarazada. Estos métodos se pueden caracterizar por las siguientes operaciones: (a) recibir una muestra del líquido corporal; (b) extraer ADN de la muestra bajo condiciones que extraigan ADN tanto de un genoma materno como de un genoma fetal presentes en el líquido corporal; (c) secuenciar el ADN extraído con un secuenciador de ácidos nucleicos bajo condiciones que produzcan secuencias de segmentos de ADN que contienen uno o más polimorfismos; (d) asignar las secuencias de segmentos de ADN derivadas de la secuenciación del ADN en el líquido corporal a uno o más polimorfismos indicados sobre una secuencia de referencia; (e) determinar frecuencias alélicas de las secuencias de segmentos de ADN asignadas a al menos uno de los polimorfismos indicados; (f) clasificar el al menos un polimorfismo indicado 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 usando las frecuencias alélicas determinadas en (e) y la combinación de cigosidades procedente de (f).

La asignación se puede realizar usando un aparato informático programado para asignar secuencias de ácido nucleico al uno o más polimorfismos indicados. En general, cualquiera de las operaciones (d)-(g) se puede realizar en uno o más procesadores que funcionan bajo las instrucciones del programa.

En ciertas realizaciones, el ADN obtenido de un líquido corporal de una embarazada es ADN libre de células obtenido del plasma de la embarazada. Típicamente, la secuenciación se efectúa sin amplificar selectivamente ninguno de los uno o más polimorfismos indicados.

En ciertas realizaciones, la asignación de los segmentos de ADN obtenidos de la sangre de la embarazada comprende asignar informáticamente los segmentos a una base de datos de polimorfismos. En ciertas realizaciones, la clasificación en (f) clasifica el al menos un polimorfismo indicado en una de las siguientes combinaciones: (i) la embarazada es homocigótica y el feto es homocigótico, (ii) la embarazada es homocigótica y el feto es heterocigótico,

(iii) la embarazada es heterocigótica y el feto es homocigótico, y (iv) la embarazada es heterocigótica y el feto es heterocigótico.

Se pueden emplear diversas operaciones de filtrado. Estas incluyen, por ejemplo, dejar de considerar cualquier polimorfismo clasificado en la combinación (i) o la combinación (iv). En otro ejemplo, los métodos incluyen además filtrar el al menos uno de los polimorfismos indicados para dejar de considerar cualquier polimorfismo que tenga una frecuencia del alelo secundario mayor que un umbral definido. En otro ejemplo más, los métodos incluyen una operación de filtrado del al menos uno de los polimorfismos indicados para dejar de considerar cualquier polimorfismo que tenga una frecuencia del alelo secundario menor que un umbral definido.

La operación de clasificación se puede ejecutar de diversos modos. Por ejemplo, puede implicar aplicar un umbral a la frecuencia alélica determinada en (e). En otro ejemplo, la operación de clasificación implica aplicar los datos de frecuencia alélica procedentes de (e), obtenidos para una pluralidad de polimorfismos, a un modelo mixto. En una ejecución, el modelo mixto emplea momentos factoriales.

La fracción fetal determinada según se describe en la presente se puede usar para diversas aplicaciones. En algunos ejemplos, los métodos descritos en la presente incluyen una operación de ejecución de las instrucciones del programa en el uno o más procesadores para registrar automáticamente la fracción de fetal de ADN que se determina en (g) en una historia clínica del paciente, almacenada en un medio legible informáticamente, para la embarazada. La historia clínica del paciente puede ser mantenida por un laboratorio, una consulta médica, un hospital, un seguro médico restringido, una compañía aseguradora o un cibersitio de historias clínicas personales. En otra aplicación, la estimación de la fracción de ADN fetal se usa para prescribir, iniciar y/o alterar el tratamiento de un sujeto humano del que se tomó la muestra de ensayo materna. En otra aplicación, la estimación de la fracción de ADN fetal se usa para solicitar o realizar una o más pruebas adicionales.

Otro aspecto de la divulgación trata de un aparato para estimar la fracción de ADN fetal en ADN obtenido de un líquido corporal de una embarazada. Este aparato se puede caracterizar por las siguientes particularidades: (a) un secuenciador configurado para (i) recibir ADN extraído de una muestra del líquido corporal que comprende ADN tanto de un genoma materno como de un genoma fetal, y (ii) secuenciar el ADN extraído bajo condiciones que produzcan secuencias de segmentos de ADN que contienen uno o más polimorfismos indicados; y (b) un aparato informático configurado para (p. ej., programado para) dar instrucciones a uno o más procesadores para realizar diversas operaciones tales como las descritas con dos o más de las operaciones del método descritas en la presente. En algunas realizaciones, el aparato informático está configurado para (i) asignar secuencias de ácido nucleico al uno o más polimorfismos indicados en una secuencia de referencia, (ii) determinar frecuencias alélicas de las secuencias de segmentos de ADN asignadas a al menos uno de los polimorfismos indicados, (iii) clasificar el al menos un polimorfismo indicado 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 usando las frecuencias alélicas y la combinación de cigosidades.

En ciertas realizaciones, el aparato también incluye una herramienta para extraer ADN de la muestra bajo condiciones que extraigan ADN tanto del genoma materno como del genoma fetal. En algunas ejecuciones, el aparato incluye un módulo configurado para extraer ADN libre de células obtenido 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 informático se puede configurar adicionalmente para dar instrucciones al uno o más procesadores para asignar los segmentos de ADN obtenidos de la sangre de la embarazada al asignar informáticamente los segmentos a la base de datos de polimorfismos. Las secuencias de la base de datos son un ejemplo de una secuencia de referencia. Otros ejemplos de secuencias de referencia se presentan posteriormente.

En ciertas realizaciones, el aparato informático está configurado además para dar instrucciones al uno o más procesadores para clasificar el al menos un polimorfismo indicado en una de las siguientes combinaciones: (i) la embarazada es homocigótica y el feto es homocigótico, (ii) la embarazada es homocigótica y el feto es heterocigótico, (iii) la embarazada es heterocigótica y el feto es homocigótico, y (iv) la embarazada es heterocigótica y el feto es heterocigótico. En algunas realizaciones, el aparato informático se configura además para dar instrucciones al uno o más procesadores para dejar de considerar cualquier polimorfismo clasificado en la combinación (i) o la combinación (iv).

En ciertas realizaciones, el aparato informático se configura adicionalmente para dar instrucciones al uno o más procesadores para dejar de considerar cualquier polimorfismo que tenga una frecuencia del alelo secundario mayor que un umbral definido. En algunas realizaciones, el aparato informático está configurado además para dar instrucciones al uno o más procesadores para filtrar el uno o más polimorfismos indicados para dejar de considerar cualquier polimorfismo que tenga una frecuencia del alelo secundario menor que un umbral definido. En ciertas realizaciones, el aparato informático está configurado además para dar instrucciones al uno o más procesadores para clasificar el al menos un polimorfismo indicado al aplicar un umbral a la frecuencia alélica.

En ciertas realizaciones, el aparato informático está configurado además para dar instrucciones al uno o más procesadores para clasificar el al menos un polimorfismo indicado al aplicar los datos de frecuencia alélica obtenidos para una pluralidad de polimorfismos, a un modelo mixto. El modelo mixto puede emplear momentos factoriales.

En ciertas realizaciones, el aparato informático está configurado además para dar instrucciones al uno o más procesadores para registrar automáticamente la fracción de fetal de ADN en una historia clínica del paciente, almacenada en un medio legible informáticamente, para la embarazada. La historia clínica del paciente puede ser mantenida por un laboratorio, una consulta médica, un hospital, un seguro médico restringido, una compañía aseguradora o un cibersitio de historias clínicas personales.

Otro aspecto de la divulgación trata de métodos para estimar una fracción de ADN fetal en ADN obtenido de un líquido corporal de una embarazada según las siguientes operaciones: (a) asignar segmentos de ADN obtenidos del líquido corporal de la embarazada a una pluralidad de secuencias polimórficas, en donde el ADN se secuenció bajo condiciones que identificaban la pluralidad de secuencias polimórficas; (b) determinar una frecuencia alélica de los ácidos nucleicos asignada a cada una de la pluralidad de secuencias polimórficas; y (c) aplicar las frecuencias alélicas a un modelo mixto para obtener una estimación de la fracción de ADN fetal en el ADN obtenido de la sangre de la embarazada. Una cualquiera o más de las operaciones (a)-(c) se pueden realizar en uno o más procesadores que funcionan bajo las instrucciones del programa. En ciertas realizaciones, la operación (c) implica ejecutar las instrucciones en el uno o más procesadores para resolver una serie de ecuaciones para momentos factoriales de datos de frecuencia alélica para cada una de la pluralidad de secuencias polimórficas. En algunas realizaciones, el modelo mixto explica el error de secuenciación.

En ciertas realizaciones, los métodos incluyen adicionalmente retirar informáticamente frecuencias alélicas para polimorfismos identificados como heterocigóticos tanto en el feto como en la embarazada. En algunas ejecuciones, antes de (c), los métodos incluyen una operación de retirar informáticamente frecuencias alélicas para polimorfismos identificados como homocigóticos tanto en el feto como en la embarazada. En algunas ejecuciones, antes de (c), los métodos incluyen una operación de retirar informáticamente frecuencias alélicas para polimorfismos identificados como heterocigóticos en la embarazada.

El ADN obtenido de un líquido corporal de una embarazada puede ser ADN libre de células obtenido del plasma de la embarazada. La asignación de los ácidos nucleicos obtenidos a partir de líquido corporal se puede ejecutar al asignar los segmentos a una base de datos de polimorfismos.

Los métodos de este aspecto de la divulgación pueden incluir además secuenciar el ADN procedente del líquido corporal de la embarazada con un secuenciador de ácidos nucleicos bajo condiciones que produzcan secuencias de segmentos de ADN que contienen las secuencias polimórficas.

En algunas ejecuciones, la asignación en (a) comprende identificar una pluralidad de secuencias polimórficas bialélicas. En otras realizaciones, la asignación (a) comprende asignar los segmentos de ADN a una pluralidad de secuencias polimórficas predefinidas.

En algunas realizaciones, los métodos de este aspecto incluyen adicionalmente ejecutar las instrucciones del programa en uno o más procesadores para registrar automáticamente la fracción de fetal del ADN según se determina en (c) en una historia clínica del paciente, almacenada en un medio legible informáticamente, para la embarazada. La historia clínica del paciente puede ser mantenida por un laboratorio, una consulta médica, un hospital, un seguro médico restringido, una compañía aseguradora o un cibersitio de historias clínicas personales.

Basándose en la estimación de la fracción de ADN fetal, los métodos de este aspecto pueden incluir además prescribir, iniciar y/o alterar el tratamiento de un ser humano del que se tomó una muestra de ensayo materna. Basándose en la estimación de la fracción de ADN fetal, los métodos de este aspecto pueden incluir además encargar y/o realizar una o más pruebas adicionales.

Según otro aspecto más de la divulgación, se proporcionan métodos para estimar la fracción de ADN fetal en ADN obtenido de un líquido corporal de una embarazada usando las siguientes operaciones: (a) recibir una muestra del líquido corporal; (b) extraer ADN de la muestra bajo condiciones que extraigan ADN tanto de un genoma materno como de un genoma fetal presente en el líquido corporal; (c) secuenciar el ADN extraído con un secuenciador de ácido nucleicos bajo condiciones que produzcan secuencias de segmentos de ADN; (d) comparar las secuencias de segmentos de ADN derivadas del líquido corporal y de la comparación que identifica uno o más polimorfismos bialélicos; (e) determinar 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 usando las frecuencias alélicas determinadas en (e) y la combinación de cigosidades procedente de (f).

La asignación se puede realizar usando un aparato informático programado para asignar secuencias de ácido nucleico a los uno o más polimorfismos indicados. En general, cualquiera de las operaciones (d)-(g) se puede realizar en uno o más procesadores que funcionan bajo las instrucciones del programa.

En ciertas ejecuciones 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 ciertas realizaciones de este aspecto, la clasificación en (f) clasifica el al menos un polimorfismo identificado en una de las siguientes combinaciones: (i) la embarazada es homocigótica y el feto es homocigótico, (ii) la embarazada es homocigótica y el feto es heterocigótico, (iii) la embarazada es heterocigótica y el feto es homocigótico, y (iv) la embarazada es heterocigótica y el feto es heterocigótico. Los métodos pueden incluir además dejar de considerar cualquier polimorfismo clasificado en la combinación (i) o la combinación (iv).

Según diversas realizaciones, los métodos de este aspecto puede incluir operaciones de filtrado y/o clasificación como las descritas en la presente en relación con otros aspectos. Por ejemplo, los métodos de este aspecto pueden incluir filtrar el uno o más polimorfismos identificados para dejar de considerar cualquier polimorfismo que tenga una frecuencia del alelo secundario mayor que 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). El uso de modelos mixtos según se describe en la presente se puede emplear para clasificar los polimorfismos identificados.

Otro aspecto de la divulgación trata de un aparato para estimar una fracción de ADN fetal y que incluye los siguientes elementos: (a) un secuenciador configurado para (i) recibir ADN extraído de una muestra del líquido 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 informático configurado para dar instrucciones a uno o más procesadores para (i) asignar los segmentos de secuencias de ADN obtenido del líquido corporal de la embarazada a una pluralidad de secuencias polimórficas, (ii) determinar una frecuencia alélica para cada una de la pluralidad se secuencias polimórficas a partir de los segmentos de secuencias de ADN asignados, y (iii) aplicar las frecuencias alélicas a un modelo mixto para obtener una estimación de la fracción de ADN fetal en el ADN obtenido de la sangre de la embarazada.

Otro aparato más para estimar la fracción de ADN fetal incluye los siguientes elementos: (a) un secuenciador configurado para (i) recibir ADN extraído de una muestra del líquido corporal que comprende ADN tanto de un genoma materno como de un genoma fetal, y (ii) secuenciar el ADN extraído bajo condiciones que produzcan secuencias de segmentos de ADN; y (b) un aparato informático configurado para dar instrucciones a uno o más procesadores para (i) comparar las secuencias de segmentos de ADN derivado del líquido corporal y a partir de la comparación identificar uno o más polimorfismos bialélicos, (ii) determinar frecuencias alélicas de las secuencias de segmentos de ADN para al menos uno de los polimorfismos identificados, (iii) clasificar el al menos un polimorfismo identificado basándose en una combinación de la cigosidad de la embarazada y la cigosidad del feto, y (iii) estimar la fracción de ADN fetal en el ADN obtenido de la embarazada usando las frecuencias alélicas y la combinación de cigosidades.

Las instrucciones y/o el equipo informático empleados en los aspectos del aparato descritos en la presente pueden proporcionar la ejecución de una cualquiera o más de las operaciones informáticas o algorítmicas de los aspectos del método divulgados en la presente, independientemente de si estas operaciones se citan explícitamente anteriormente.

Estas y otras características y ventajas de las realizaciones divulgadas se describirán con más detalle posteriormente con referencia a los dibujos adjuntos.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

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

La Figura 2 es un esquema de procesamiento ejemplar para ejecutar algunas de las realizaciones divulgadas.

La Figura 3 presenta estimaciones de error mediante la posición de bases secuenciada a lo largo de 30 carriles de datos de Illumina GA2 alineados al genoma humano HG18 usando Eland con parámetros por defecto.

La Figura 4 es una gráfica de recuento del alelo secundario A frente a la cobertura D (suponiendo ausencia de error) para los casos de heterocigosidad 1 a 4.

La Figura 5 representa la transformación de los datos del Caso 3 en el Caso 2.

La Figura 6 presenta datos posteriores a la rotación, en los que D1 se seleccionaba de modo que el caso 1 y los casos 2, 3 no se solaparan. 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 usando un modelo mixto y la fracción fetal conocida y la fracción fetal estimada.

La Figura 8 muestra que usar el grado de error de la máquina como un parámetro conocido reduce la desviación ascendente en un punto.

En la Figura 9 muestra que los datos simulados que usan el grado de error de la máquina como un parámetro conocido, potenciar los modelos de error del caso 1 y 2 reduce mucho la desviación ascendente hasta menos de un punto para una fracción fetal por debajo de 0,2.

La Figura 10 es una representación esquemática de un sistema informático que, cuando se configura (p. ej., se programa) o diseña apropiadamente, puede servir como un aparato de análisis para realizaciones divulgadas.

Las Figuras 11A y B muestran un histograma del número de observaciones de variantes (Frecuencia) al porcentaje del alelo secundario (A/D) para el cromosoma cromosomas 1(A) y el cromosoma 7 que se producen en un ejemplo.

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

DESCRIPCIÓN DETALLADA

Introducción y Visión General

Ciertas realizaciones divulgadas implican analizar ADN tomado de sangre de una embarazada y usar el análisis para estimar la fracción de ese ADN que proviene del feto. La fracción fetal de ADN se puede usar a continuación para atribuir algún nivel de confianza a otra medida o caracterización del feto basándose en un análisis independiente del ADN recogido de la sangre de la madre. Por ejemplo, una muestra de ADN fetal tomada de sangre materna se puede analizar separadamente para detectar aneuploidía en el feto. La determinación de la aneuploidía realizada mediante este análisis separado se puede dar mediante un nivel de confianza fundamentado estadísticamente basándose en la cantidad fraccionaria de ADN fetal presente en el ADN tomado de la sangre de la madre. Fracciones de ADN fetal relativamente bajas en el complemento total de ADN sugiere una baja confianza en cualquier caracterización basada en ADN fetal.

Típicamente, aunque no necesariamente, el ADN analizado en la sangre de la madre es ADN libre de células, aunque, en algunas realizaciones, puede ser ADN unido a células. El ADN libre de células se toma del plasma de la madre. La cantidad de ADN fetal en el contenido de ADN libre de células tomado de embarazadas varía ampliamente dependiendo de una variedad de factores incluyendo la edad gestacional del feto. Para embarazadas típicas, actualmente se cree que aproximadamente 5-20% del ADN libre de células es ADN fetal. Sin embargo, no es infrecuente que la fracción fetal sea significativamente menor (p. ej., aproximadamente 1% o menor). En estos casos, cualquier caracterización separada del ADN fetal puede ser inherentemente sospechosa. Por otra parte, algunos investigadores han presentado muestras de ADN libre de células materno que tienen fracciones de ADN fetal tan altas como 40% o 50%.

En ciertas ejecuciones descritas en la presente, la determinación de la fracción fetal de ADN materno se basa en múltiples lecturas de secuencias de ADN en sitios de secuencia que se sabe que albergan uno o más polimorfismos. Típicamente, aunque no necesariamente, estos polimorfismos son polimorfismos de un solo nucleótido (SNP). Otros tipos de polimorfismos adecuados incluyen eliminaciones, STRs (repeticiones cortas en tándem), inserciones, índeles (incluyendo microíndeles), etc. Ejemplos adicionales se presentan posteriormente. En ciertas realizaciones, los sitios polimórficos se encuentran en una "secuencia de referencia" según se describe posteriormente. En algunas realizaciones, los sitios polimórficos se descubren mientras se alinean marcadores de secuencia entre sí y/o a una secuencia de referencia.

Ciertos métodos divulgados hacen uso del hecho de que las secuencias de ADN de un feto en los sitios polimórficos considerados pueden no corresponder a las de su madre. Por ejemplo, el ADN de la madre en el sitio de un SNP particular puede ser homocigótico, mientras que la versión fetal del SNP será heterocigótica. De ahí que una colección de muestras de secuencias tomadas para el SNP en cuestión será heterogénea con la mayoría de las secuencias que contienen el alelo principal y la fracción restante que contiene el alelo secundario. Las cantidades relativas de los alelos principal y secundario se determinan mediante la fracción de ADN fetal de la muestra.

Se debe mencionar que en una muestra homocigótica ambas copias de un SNP u otro polimorfismo dado contienen el mismo alelo, mientras que un SNP u otro polimorfismo heterocigótico contiene una copia del alelo principal y una copia del alelo secundario. Por lo tanto, se sabe que el ADN tomado exclusivamente de un individuo heterocigótico debe contener 50% del alelo principal y 50% del alelo secundario. Este conocimiento se puede usar para elucidar la fracción de ADN fetal según se esboza posteriormente. Según se explica más a fondo posteriormente, diversos métodos divulgados en la presente consideran solamente polimorfismos en los que solo hay dos alelos en el ADN materno y fetal, colectivamente.

En algunas ejecuciones, el ADN tomado de la sangre de la madre se lee muchas veces, asignándose el número total de lecturas a un sitio particular de un polimorfismo que se considera la "cobertura" del polimorfismo, y considerándose el número de lecturas que se asigna al alelo secundario para ese polimorfismo el recuento del alelo secundario. La relación del recuento del alelo secundario a la cobertura es importante en diversas ejecuciones.

Ciertos métodos divulgados en la presente identifican y caracterizan cuatro casos de polimorfismos en muestras de ADN que comprenden ADN procedente tanto de la madre como del feto. La Figura 1 posterior representa estos cuatro casos. Específicamente, en un primer caso, que es bastante poco interesante, tanto la madre como el feto son homocigóticos en el polimorfismo particular que se considera. En este caso, cada secuencia en la muestra de ADN que contiene el polimorfismo en cuestión contendrá el mismo alelo y no se puede recoger información acerca de las cantidades relativas de ADN procedentes de la madre y el feto. Sin embargo, se debe apuntar que este caso podría ser interesante en el sentido de que permitiría al investigador o técnico tener alguna idea del grado de error relativo del aparato de secuenciación de ADN usado para generar los datos de secuencia considerados.

El segundo caso que encontrará el análisis es un polimorfismo para el que la embarazada es homocigótica y el feto es heterocigótico. En este caso, una fracción relativamente pequeña, pero no obstante significativa, de las secuencias detectadas contendrá el alelo secundario. Específicamente, en este segundo caso, la frecuencia del alelo secundario está dada nominalmente por la fracción de ADN fetal en la corriente sanguínea de la madre dividida por dos.

En un tercer caso, el polimorfismo considerado es heterocigótico en el ADN de la madre y homocigótico en el ADN del feto. En esta situación, la frecuencia del alelo secundario está dada nominalmente por 0,5 menos la mitad de la fracción de ADN fetal en la muestra de ADN.

Finalmente, en el cuarto caso, el polimorfismo considerado es heterocigótico tanto en la madre como en el feto. En este caso, se espera que la frecuencia de los alelos principal y secundario sea 0,5 en ambos. Como con el primer caso, el cuarto caso es relativamente poco interesante para determinar la fracción fetal de ADN.

Si el investigador, el técnico o el programa encargados de determinar la fracción de ADN fetal en una muestra conocida para un polimorfismo dado supiera a cuál de los cuatro casos pertenecía ese polimorfismo, entonces la fracción de ADN fetal se podría estimar directamente, suponiendo que el polimorfismo considerado entrara bien en el caso dos o bien en el caso tres. Sin embargo, en la práctica, nunca se tiene este conocimiento *a priori.* Por lo tanto, se requiere un aparato informático para realizar las operaciones descritas en la presente.

En ciertas realizaciones, descritas en cualquier parte en la presente, se emplea una técnica de umbralización para clasificar un solo polimorfismo en uno de los cuatro casos. Una vez que el polimorfismo se ha clasificado así y se ha encontrado que se encuentra en el caso bien 2 o bien 3, se puede estimar la fracción fetal. En otras realizaciones, la técnica considera múltiples polimorfismos distribuidos a través de la totalidad o una porción del genoma. Según se ilustra en los ejemplos específicos, se pueden usar con este propósito múltiples SNPs diferentes a través del genoma.

En realizaciones particulares, la frecuencia alélica se determina para un número de diferentes polimorfismos en una muestra de ADN tomada de una muestra de sangre de la madre. Para esta pluralidad de polimorfismos, alguna 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. Estas fracciones sumarán un valor de 1. Se puede emplear un modelo mixto o técnica relacionada para extraer una o más propiedades estadísticas de los polimorfismos en cada una de estas cuatro categorías. Específicamente, se puede emplear un modelo mixto para determinar una media y opcionalmente la varianza para cada uno de los cuatro casos encontrados en una muestra de ADN tomada de la sangre de una embarazada. En realizaciones específicas, esta es la media y la varianza asociada con la frecuencia del alelo secundario en relación con el número total de recuentos para un polimorfismo considerado (cobertura). Según se elabora en cualquier parte en la presente, 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 tomado de la sangre de la madre.

En una realización específica que emplea modelos mixtos, se calculan uno o más momentos factoriales para cada posición en la que se considere un polimorfismo. Por ejemplo, un momento factorial (o una colección de momentos factoriales) se calcula usando múltiples posiciones de SNP consideradas en la secuencia de ADN. Según se muestra en la ecuación 4 posteriormente, cada uno de los diversos momentos factoriales es una suma de todas las diversas posiciones de SNP consideradas para la relación de frecuencia de alelo secundario a cobertura para una posición dada. Según se muestra en la ecuación 5 posteriormente, estos momentos factoriales también están relacionados con los parámetros asociados con cada uno de los casos de cigosidad descritos anteriormente. Específicamente, se refieren a la probabilidad para cada uno de los casos así como las cantidades relativas de cada uno de los cuatro casos en la colección de polimorfismos considerada. Según se explica, la probabilidad es una función de la fracción de ADN fetal en el ADN libre de células en la sangre de la madre. Según se explica más a fondo posteriormente, al calcular un número suficiente de estos momentos factoriales (que se muestran en la ecuación 4), el método proporciona un número suficiente de expresiones para resolver todas las incógnitas. Las incógnitas en este caso serían las cantidades relativas de cada uno de los cuatro casos en la población de polimorfismos considerada así como las probabilidades (y de ahí las fracciones de ADN fetal) asociadas con cada uno de estos cuatro casos. Véase la ecuación

5. Se pueden obtener resultados similares usando otras versiones de modelos mixtos según se representa en las ecuaciones 7-12 posteriormente. Estas versiones particulares hacen uso solamente de polimorfismos que entran dentro de los casos 1 y 2, filtrándose los polimorfismos para los casos 3 y 4 mediante una técnica de umbralización.

Así, los momentos factoriales se pueden usar como parte de un modelo mixto para identificar las probabilidades de cualquier combinación de los cuatro casos de cigosidad. Y, según se menciona, estas probabilidades, o al menos las de los casos segundo y tercero, están directamente relacionadas con la fracción de ADN fetal en el ADN libre de células total en la sangre de la madre.

También se debe mencionar que se puede emplear el error de secuenciación para reducir la complejidad del sistema de ecuaciones de momentos factoriales que se deben resolver. A este respecto, se debe conocer que el error de secuenciación puede tener realmente uno cualquiera de cuatro resultados (correspondientes a cada una de las cuatro bases posible en cualquier posición de polimorfismo dada).

En ciertas realizaciones, los marcadores se alinean con un cromosoma o genoma de referencia, y se identifican polimorfismos bialélicos. Estos polimorfismos no están predefinidos o identificados de otro modo antes del alineamiento. Simplemente se identifican durante el alineamiento y a continuación se caracterizan basándose en sus cigosidades y recuentos del alelo secundario según se describe en la presente. Esta información se usa para estimar fracciones genómicas según se describe anteriormente.

Las longitudes de los marcadores usados en realizaciones descritas en la presente se determinarán generalmente mediante el método de secuenciación empleado para generar los marcadores. Los métodos son robustos a través de un amplio intervalo de longitudes de los marcadores. En ciertas ejecuciones, los marcadores tienen una longitud entre aproximadamente 20 y 300 pares de bases (o una longitud de aproximadamente 30 a 100 pares de bases).

Un esquema de procesamiento ejemplar para ejecutar algunas de las realizaciones divulgadas se muestra en la Figura 2. Según se representa allí, el procedimiento comienza en 201 con la recogida de ADN (libre de células o unido a células) procedente de sangre u otro líquido corporal materno. A partir de este ADN, múltiples secuencias se asignan a uno o más polimorfismos en una secuencia de referencia. Esta asignación proporciona una frecuencia alélica para cada uno de los polimorfismos. Véase el bloque 203.

Más específicamente, el procedimiento en el bloque 203 puede implicar leer secuencias del ADN recogido en las posiciones de múltiples polimorfismos. En algunos casos, se pueden generar como parte del procedimiento para determinaciones de la ploidía u otra determinación realizada con respecto al ADN fetal. Así, en algunas realizaciones, no se necesita generar secuencias separadas. Las secuencias leídas se alinean con una secuencia de referencia para maximizar el alineamiento usando BLAST o una herramienta similar.

La secuencia de referencia se puede proporcionar como una base de datos de polimorfismos. En algunos casos, esta es un grupo de referencia de búsqueda de alelos producido a partir de una expansión combinatoria de todas las definiciones de polimorfismo (p. ej., en el caso en el que los polimorfismos sean SNPs, todas las secuencias de SNP). Véase el apéndice, 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. Un modelo mixto se puede emplear con este propósito en ciertas realizaciones. Según se menciona, las combinaciones son como sigue: M y F homocigóticos, M homocigótica y F heterocigótico, M heterocigótica y F homocigótico, y M y F heterocigóticos.

Finalmente, según se ilustra en el bloque 207, el método usa la combinación de frecuencia alélica de casos de cigosidad en uno o más de los polimorfismos para estimar la cantidad fraccionaria de componente fetal en el ADN procedente de la muestra materna.

Definiciones

El siguiente análisis se proporciona como una ayuda para comprender ciertos aspectos y ventajas de las realizaciones divulgadas.

El término "lectura" se refiere a una secuencia leída a partir de una porción de una muestra de ácido nucleico. Típicamente, aunque no necesariamente, una lectura representa una secuencia corta de pares de bases contiguos en la muestra. La lectura se puede representar simbólicamente mediante la secuencia de pares de bases (en ATCG) de una porción de la muestra. Se puede almacenar en un dispositivo de memoria y procesar según sea apropiado para determinar si coincide con una secuencia de referencia o cumple otros criterios. Una lectura se puede obtener directamente de un aparato de secuenciación o indirectamente a partir de información de secuencias relativa a la muestra.

El término "marcador" también se refiere a secuencias cortas procedentes de una muestra de ácido nucleico. Típicamente, un marcador contiene información asociada tal como la posición de la secuencia en el genoma. Para algunos propósitos, los términos lectura y marcador son intercambiables en la presente. Sin embargo, típicamente, las lecturas de la secuencia se alinean con una secuencia de referencia, y las lecturas que se asignan a un solo sitio en el genoma de referencia se denominan marcadores. La "secuencia segmentada" se usa a veces en la presente intercambiablemente con "marcador".

Frecuentemente en la presente, las "lecturas" se describen como secuencias de ácidos nucleicos que tienen una longitud de 36 pares de bases (36meros). Por supuesto, las realizaciones divulgadas no se limitan a este tamaño. Lecturas menores y mayores son adecuadas en muchas aplicaciones. Para aplicaciones que alinean lecturas con el genoma humano, una lectura de un tamaño de 30 pares de bases o mayor se considera generalmente suficiente para asignar una muestra a un solo cromosoma. Marcadores/lecturas mucho mayores son adecuadas para algunas aplicaciones. Con la secuenciación del genoma completo, se pueden usar lecturas del orden de 1000 pares de bases o mayores. En ciertas realizaciones, una lectura puede tener una longitud de entre aproximadamente 20 y 10.000 pares de bases, o entre aproximadamente 30 y 1000 pares de bases, o entre aproximadamente 30 y 50 pares de bases.

Una "secuencia de referencia" es una secuencia de una molécula biológica, que es frecuentemente un ácido nucleico tal como un cromosoma o genoma. Típicamente, múltiples lecturas son miembros de una secuencia de referencia dada. En ciertas realizaciones, una lectura o un marcador se compara con una secuencia de referencia para determinar si la secuencia de referencia contiene la secuencia de lectura. Este procedimiento se denomina a veces alineamiento.

En diversas realizaciones, la secuencia de referencia es significativamente mayor que las lecturas que se alinean a ella. Por ejemplo, puede ser al menos aproximadamente 100 veces mayor, o al menos aproximadamente 1000 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. Estas secuencias se pueden denominar secuencias de referencia genómicas. En otro ejemplo, la secuencia de referencia se limita a un cromosoma humano específico tal como el cromosoma 13. Estas secuencias se pueden denominar secuencias de referencia cromosómicas. Otros ejemplos de secuencias de referencia incluyen genomas de otras especies, así como cromosomas, regiones subcromosómicas (tales como cadenas), etc. de cualquier especie.

En diversas realizaciones, la secuencia de referencia es una secuencia de consenso u otra combinación derivada de múltiples individuos. Sin embargo, en ciertas aplicaciones, la secuencia de referencia se puede tomar de un individuo particular.

El término "alineamiento" se refiere al procedimiento de comparación de una lectura o un marcador con una secuencia de referencia y de ese modo determinar si la secuencia de referencia contiene la secuencia de lectura. Si la secuencia de referencia contiene la lectura, la lectura se puede asignar a la secuencia de referencia o, en ciertas realizaciones, a una posición particular en la secuencia de referencia. En algunos casos, el alineamiento simplemente indica si una lectura es o no un miembro 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 humano 13 indicará si la lectura está presente en la secuencia de referencia para el cromosoma 13. Una herramienta que proporciona esta información se puede denominar un comprobador de miembros de un grupo. En algunos casos, un alineamiento indica adicionalmente una posición en la secuencia de referencia a la que se asigna la lectura o el marcador. Por ejemplo, si la secuencia de referencia es la secuencia de todo el genoma humano, un alineamiento puede indicar que una lectura está presente en el cromosoma 13, y puede indicar además que la lectura está en una cadena particular del cromosoma 13.

Un "sitio" es una posición única en una secuencia de referencia correspondiente a una lectura o un marcador. En ciertas realizaciones, especifica la identidad de un cromosoma (p. ej., el cromosoma 13), una hebra del cromosoma y una posición exacta en el cromosoma.

"Sitio polimórfico" es un locus en el que se produce una divergencia en la secuencia nucleotídica. El locus puede ser tan pequeño como un par de bases. Los marcadores ilustrativos tienen al menos dos alelos, cada uno presente en una frecuencia de más de 1%, y más típicamente mayor de 10% o 20% de una población seleccionada. Un sitio polimórfico puede ser tan pequeño como un par de bases. Los términos "locus polimórfico" y "sitio polimórfico" se usan en la presente intercambiablemente.

"Secuencia polimórfica" se refiere en la presente a una secuencia de ácido nucleico, p. ej. una secuencia de ADN, que comprende uno o más sitios polimórficos, p. ej. un SNP o un SNP en tándem. Secuencias polimórficas según la presente tecnología se pueden usar para diferenciar específicamente entre alelos maternos y no maternos en la muestra materna que comprende una mezcla de ácidos nucleicos fetales y maternos.

Realizaciones Detalladas

Típicamente, los procedimientos descritos en la presente emplean una secuencia de referencia que abarca uno o más polimorfismos y está asociada con el ADN que se muestrea. Una secuencia de referencia puede ser, por ejemplo, el genoma humano, un cromosoma o una región de un cromosoma. Uno o más de los polimorfismos puede estar indicado con el propósito de estimar la fracción de ADN fetal. Los polimorfismos que están indicados para el uso en la determinación de la fracción fetal son polimorfismos que son previamente conocidos. Por ejemplo, una lista exhaustiva de referencias, datos e información de secuencia sobre STRs conocidos previamente y datos de población relacionados se compilan en STRBase, a la que se puede acceder a través de Internet en ibm4.carb.nist.gov:8800/dna/home.htm. Información de secuencias de GenBank® (http://www2.ncbi.nlm.nih.gov/cgi-bin/genbank) para locus de STR usados comúnmente también es accesible a través de STRBase. Se puede acceder a información de SNPs conocidos previamente a través de están disponibles de bases de datos disponibles públicamente incluyendo, pero no limitadas a, bases de datos de SNP humanos en la dirección de Internet wi.mit.edu, la página de inicio de dbSNP del NCIB en la dirección de Internet ncbi.nlm.nih.gov, la dirección de Internet lifesciences.perkinelmer.com, Applied Biosystems by Life Technologies ™ (Carlsbad, CA) en la dirección de Internet appliedbiosystems.com, la base de datos de SNP humanos de Celera en la dirección de Internet celera.com, la base de datos de SNP del Genome Analysis Group (GAN) en la dirección de Internet gan.iarc.fr. En una realización, los SNPs indicados para determinar la fracción fetal se seleccionan del grupo de 92 SNPs de identificación individuales (IISNPs) descrito por Pakstis y cols. (Pakstis et el. Hum Genet 127:315-324 [2010]), que se ha mostrado que tienen una variación muy pequeña en la frecuencia a través de poblaciones (Fst<0,06), y que son altamente informativos en todo el mundo, teniendo una heterocigosidad promedio ≥0,4. SNPs que se usan en el método de la invención incluyen SNPs conectados y no conectados. Para denominar secuencias de SNP en tándem adecuadas, se puede buscar la base de datos del International HapMap Consortium (The International HapMap Project, Nature 426:789-796 [2003]). La base de datos está disponible en Internet en hapmap.org.

Los polimorfismos así empleados pueden ser conjuntos de polimorfismos conocidos previamente indicados para determinar la fracción de ADN fetal o se pueden encontrar por casualidad en un análisis de ADN materno para otros propósitos tales como asignar ADN de una muestra a cromosomas.

En ciertas realizaciones, el método comprende secuenciar ADN en una muestra usando una mezcla de genomas, p. ej. una muestra materna que comprende ADN libre de células fetal y materno, para proporcionar una pluralidad de marcadores de secuencia que se asignan a secuencias que comprenden sitios polimórficos previamente conocidos en un genoma de referencia, y usar los marcadores asignados en los sitios conocidos previamente para determinar la fracción fetal según se describe con detalle posteriormente. Alternativamente, después de la secuenciación del ADN, los marcadores de secuencia que se obtienen mediante la tecnología de secuenciación, p. ej. NGS, se asignan a un genoma de referencia, p. ej. hg19, y los marcadores de secuencia que se asignan a sitios en los que se producen polimorfismos por casualidad, es decir no conocidos previamente, se usan para determinar la fracción fetal.

La secuencia de referencia cuyos marcadores de secuencia se asignan a sitios polimórficos conocidos previamente puede ser un genoma de referencia publicado o puede ser una base de datos artificial u otra colección predefinida de secuencias para los polimorfismos considerados. Cada una de las secuencias de la base de datos abarcará el uno o más nucleótidos asociados con el polimorfismo. Como un ejemplo, véase la lista de secuencias polimórficas presentadas posteriormente en el "Apéndice 1."

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

En un ejemplo, la cobertura y la frecuencia alélica del SNP se determinan al alinear 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 las variaciones bialélicas rodeada por, p. ej., al menos aproximadamente 50 bases de secuencia de flanqueo. Por ejemplo, un amplicón con una cadena de información de variaciones "[g/c]" (que representa alelos alternos "g" y "c" puede parecer:
atcg.....accg[g/c]ccgt....

En algunos casos, el procedimiento para introducir la base de datos de amplicones y las secuencias generadas y obtener recuentos de SNP/alelos es como sigue.

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

a. Por ejemplo, considerando la secuencia del amplicón ejemplar anterior, se crearían dos secuencias: 1) atcg.....accgGccgt.... y 2) atcg.....accgCccgt....

b. Un ejemplo de un conjunto de referencia de búsqueda de alelos completo se puede encontrar en la lista de secuencias de la base de datos de búsqueda de alelos.

2. Asígnense las secuencias al conjunto de referencia de búsqueda de alelos manteniendo solo asignaciones que se ajusten solo a una secuencia del conjunto de búsqueda.

3. El recuento de alelos se determina al contar el número de secuencias que se ajusta a su secuencia alélica.

Los métodos divulgados en la presente suponen un embarazado "normal", es decir, un embarazo en el que la madre está embarazada de un solo feto, y no gemelos, trillizos, etc. Los expertos apreciarán modificaciones que tienen en cuenta embarazos anormales, particularmente aquellos en los que se conoce el número de fetos.

Según se indica, cuando se determina la fracción fetal, el método secuencia el ADN de la muestra procedente de sangre materna y recuenta los marcadores de secuencia que se asignan a cada secuencia del polimorfismo o los polimorfismos considerados. Para cada polimorfismo, el método cuenta el número de lecturas que se asignan a él (la cobertura) y los números de marcadores de secuencia asociados con cada alelo (los recuentos alélicos). En un ejemplo simple, un polimorfismo que tiene una cobertura de 5 puede tener 3 lecturas del alelo B y 2 lecturas del alelo A. En este ejemplo, el alelo A se considera el alelo secundario y el alelo B se considera el alelo principal.

En algunas realizaciones, esta operación hace uso de herramientas de secuenciación muy rápidas tales como herramientas de secuenciación de ADN masivamente paralelas. Ejemplos de estas herramientas se describen con más detalle posteriormente. En algunos casos, se leen muchos miles o millones de secuencias marcadoras para una sola muestra. Preferiblemente, la secuenciación se realiza de un modo que permita una asignación rápida y directa de ADN secuenciado a secuencias predefinidas particulares que albergan polimorfismos considerados. Generalmente, existe una información suficiente para este propósito en marcadores de un tamaño de 30 pares de bases o mayores. Los marcadores de este tamaño se pueden asignar inequívocamente a secuencias de interés. En una realización específica, las secuencias marcadoras empleadas en el procedimiento tienen una longitud de 36 pares de bases.

Los marcadores se asignan a un genoma de referencia o a secuencias de una base de datos de secuencias alélicas (p. ej., véase el Apéndice 1 que se menciona previamente) y se determina el número de marcadores así asignados. Esto proporcionará tanto la cobertura como el recuento del alelo secundario para cada polimorfismo considerado. En algunos casos, esto se puede realizar simultáneamente con la asignación de cada marcador a uno de los 23 cromosomas humanos y la determinación del número de marcadores asignados por cromosoma.

Según se menciona, la cobertura es el número total de secuencias leídas que se asignan a un polimorfismo dado en una secuencia de referencia. El recuento alélico es el número total de secuencias leídas que se asignan a este polimorfismo que tienen un alelo. La suma de todos los recuentos alélicos debe ser igual a la cobertura. El alelo con el recuento más alto es el alelo principal y el alelo con el recuento más bajo es el alelo secundario. En ciertas realizaciones, la única información necesaria para estimar la fracción de ADN fetal es la cobertura y el recuento del alelo secundario para cada uno de una pluralidad de polimorfismos. En algunas realizaciones, también se usa un grado de error de la designación de bases del aparato de secuenciación de ADN.

Es útil considerar los fundamentos matemáticos o simbólicos de ciertos métodos divulgados en la presente. Según se menciona, en diversos ejemplos, las secuencias generadas a partir de sangre materna se alinean (superpuestas de modo que las bases idénticas se maximicen) con un genoma u otra secuencia de ácido nucleico de referencia. Dada una posición genómica, j, y un conjunto de secuencias alineadas a la referencia, déjese que el número de presencias de cada una de las cuatro bases de ADN ("a","t","g" y "c", también llamadas "alelos"), entre las secuencias alineadas, sea w(j,1), w(j,2), w(j,3) y w(j,4), respectivamente. Para los propósitos de este análisis, se puede suponer sin pérdida de generalidad que todas las variaciones son bialélicas. De ahí que se puedan usar las siguientes notaciones:

Recuento del alelo principal en la posición genómica j como $B \equiv B_j \equiv \{b_j\} \equiv w_{j,i}^{(1)} = \max_{i \in \{1,2,3,4\}} \{w_{j,i}\}$ como la estadística de primer orden de los recuentos en la posición j (El alelo principal, b, es el correspondiente argmax. Se usan subíndices cuando se está considerando más de un SNP.),

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

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

El grado de error de la máquina secuenciadora se indica e.

Cuando el contexto esté claro, por comodidad, las notaciones se usan intercambiablemente; por ejemplo, A, Ai o {ai} se pueden usar intercambiablemente para el alelo secundario o el recuento del alelo secundario. Se pueden usar

subíndices o no dependiendo de si se está considerando más de un SNP. (Los SNPs se usan solamente con propósitos de ejemplo. Se pueden usar otros tipos de polimorfismos según se analiza en cualquier parte en la presente.).

En la Figura 1, se representa la base para los cuatro estados de cigosidad polimórfica. Según se ilustra, la madre puede ser homo o heterocigótica en un polimorfismo dado. De forma similar, el hijo puede ser bien heterocigótico o bien homocigótico en la misma posición. Según se ilustra, los casos 1 y 2 son los casos de polimorfismo en los que la madre es homocigótica. Si el hijo y la madre son ambos homocigóticos, el polimorfismo es un polimorfismo de caso 1. Según se indica anteriormente, esta situación típicamente no es particularmente interesante. Si la madre es homocigótica y el hijo es heterocigótico, la fracción fetal, *f*, está dada nominalmente por dos veces la relación del alelo secundario a la cobertura. En el caso del polimorfismo en el que la madre es heterocigótica y el hijo es homocigótico (caso 3 en la Figura 1), la fracción fetal es nominalmente uno menos dos veces la relación del alelo secundario a la cobertura. Finalmente, en el caso en el que tanto la madre como el feto sean heterocigóticos, la fracción de alelo secundario debe ser siempre 0,5, salvo error. La fracción fetal no se puede derivar para polimorfismos que se encuentren dentro del caso 4.

Los cuatro casos se elaborarán adicionalmente ahora.

Caso 1: Madre e Hijo Homocigóticos

- En este caso, salvo error de secuenciación o contaminación, no se deben observar diferencias.
- E(frecuencia del alelo secundario) = E(A) = 0.
- En la práctica A ~ (se distribuye como) una distribución binomial que se aproxima bien mediante la distribución de Poisson para np bajo. El parámetro del grado de distribución para binomial o Poisson está relacionado con el grado de error de secuenciación, e, y la cobertura D. La Figura 3 muestra frecuencias discrepantes de secuencias 36meras generadas alineadas con un genoma de referencia humano.
- Este caso no contiene información acerca de la fracción fetal.

La Figura 3 presenta estimaciones de error mediante la posición de bases secuenciadas sobre 30 carriles de datos de Illumina GA2 alineados al genoma humano HG18 usando Eland con parámetros por defecto.

Caso 2: Madre Homocigótica e Hijo Heterocigótico

- En este caso, para una fracción fetal (*f*) pequeña, las frecuencias alélicas observadas serán notablemente diferentes. Con el alelo principal presentándose habitualmente a una frecuencia varias veces mayor que el alelo secundario.
- Salvo error, dada una posición de SNP individual (D,A), E(A) = Df/2 y una estimación no desviada 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 Heterocigótica e Hijo Homocigótico

- En este caso, las frecuencias observadas para los alelos principal y secundario están cercanas y A/D está justo por debajo de 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 Heterocigótica e Hijo Heterocigótico

Nótese que, salvo error, existen dos subcasos para este.

Caso 4.1: El alelo procedente del padre es diferente a los alelos de la madre

Esto introduciría un tercer alelo que sería el alelo secundario con E(A) = Df/2. Estos casos no deben tener un efecto sobre las estimaciones para f debido a que el procedimiento para asignar secuencias a amplicones filtrará estos casos cuando los SNPs de referencia sean bialélicos.

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

- En este case, salvo error, los dos alelos aparecerían en proporción 1:1 de modo que este caso no es 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 una gráfica de recuento de alelo secundario A frente a cobertura D (suponiendo ausencia de error) para los casos de heterocigosidad 1 a 4.

En diversas realizaciones, el método se refiere ampliamente a analizar la frecuencia alélica en uno o más SNPs (u otros polimorfismos) para clasificar los polimorfismos por estar bien en el caso 2 y/o bien en el caso 3. Usando la frecuencia alélica junto con la clasificación, el método puede estimar la fracción fetal.

En algunos casos, un recuento de alelo secundario A y una cobertura D dados, en otra palabras un solo punto (D,A), para una posición del SNP individual permite que los métodos hagan una estimación de un solo punto. Por ejemplo, ciertos métodos clasifican un SNP con un recuento alélico (D,A) en un solo caso y derivan una estimación de la fracción fetal como sigue:

ES1.1 Umbrales Simples para Decidir el Caso

Dada una posición individual (SNP),

1. Decídase por el caso 1 con una función de decisión como 2A/D < e o un valor crítico definido de Binomial(e,D) o Poisson(De). También se puede usar una distribución alternativa según se divulga en la presente). Sin estimación de la fracción fetal (f).

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

3. De otro modo, decídase por el caso 2 si 2A/D < 0,25 (o algún otro umbral fijado manualmente o estimado automáticamente). Fracción fetal f estimada como 2A/D

4. De otro modo, caso 3. úsese la estimación de fracción fetal f = (1-2A/D).

Se pueden ganar precisión al combinar la información del recuento alélico procedente de varios SNPs para estimar la fracción fetal.

Método EM1: Combínense Múltiples SNPs mediante Promediado.

Tómese la media, la mediana, otra medida central (por ejemplo: Tukey bicuadrática, estimadores de M, etc...). También se pueden usar promedios ponderados. Para un ejemplo de cómo se pueden definir las ponderaciones véase EM2.4 posteriormente. Se pueden usar medidas del centro adicionalmente robustas.

Método EM2 Estimación simultánea del caso 2 y el caso 3 mediante transformación

Para ocasiones en las que f es menor de X%, los puntos del caso 3 (D,A) se pueden transformar para que sean coincidentes con los puntos del caso 2. A partir de esta línea, se puede calcular una pendiente común mediante regresión a través del origen (véase la Figura 5).

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

La Figura 5 representa la transformación de los datos del Caso 3 en el Caso 2. Ahora, una sola regresión puede estimar f a partir de ambos casos simultáneamente.

Método para calcular EM2.3:

Etapa 1: Deséchense los datos del Caso 4

Para cada punto de datos (D,A) si A > (0,5D-T1) entonces exclúyase (D,A) de un análisis adicional. T1(D,A) una función de valor real.

Etapa 2: Transfórmense los datos del Caso 3

Véase la Figura 6. Para cada punto de datos (D,A) que se determina que no es 4, si A > T2*D entonces transfórmense en puntos hasta nuevas coordenadas (D1,A1). T2(D,A) una función de valor real.

$$\alpha = 2A/D$$

$$A1 = -1(0.5D - A)$$

$$D1 = D$$

Etapa 3: Establézcase DT liminar para reducir la contaminación procedente de los datos del caso 1

Ignórense todos los puntos de datos T2(D,A) una función de valor real.

Etapa 4: Estimación de la regresión para los restantes datos de los casos 2 y 3 transformados.

Aplíquese regresión a través del origen hasta los restantes puntos. La estimación de la fracción fetal es dos veces la pendiente de la semejanza de regresión.

Nótese que existen muchas clases de transformaciones que se pueden construir para efectuar la misma coincidencia de los datos de los casos 2 y 3. Ejemplos incluyen transformación trigonométrica o uso de matrices de rotación. Estas desviaciones están destinadas a ser incluidas en el alcance de esta divulgación. Por otra parte, se pueden usar muchas clases de regresión (L2, L1, ...) u optimización. Cambiar el algoritmo de optimización es un cambio trivial y cubierto bajo el alcance de esta divulgación.

La Figura 6 presenta datos posteriores a la rotación. Selección de D1 de modo que el caso 1 y los casos 2 y 3 no se solapen. E1 representa un límite superior para el 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 una varianza igual. Es más apropiado tener en cuenta la heterocedasticidad de las diferentes fuentes de datos e incluso de puntos procedentes del 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 procedentes de los datos del caso 2 tendrán una varianza v2(f,D) = [0,5*Df – 0,25*Df^2] y los puntos procedentes de los datos del caso 3 tendrán una varianza v3(f,D)=[0,25D(1 - f^2)]. Suponiendo que se dé a cada punto una ponderación, w, diferente que en EM2.3, se busca minimizar

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

Ecuación 1

Fijando las primeras derivadas a cero y resolviendo para 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 secundario (transformado para el caso 3) de SPN i

Ecuación 2

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

En ciertas realizaciones, se puede emplear un modelo mixto para clasificar una colección de polimorfismos en dos o más de los casos de cigosidad y simultáneamente estimar la fracción de ADN fetal a partir de frecuencias alélicas medias para cada uno de estos casos. Generalmente, un modelo mixto supone que una colección de datos particular está constituida por una mezcla de diferentes tipos de datos, cada uno de los cuales tiene su propia distribución esperada (p. ej., una distribución normal). El procedimiento intenta encontrar la media y posiblemente otras características para cada tipo de datos. En realizaciones divulgadas en la presente, existen hasta cuatro tipos de datos diferentes (los casos de cigosidad) que constituyen los datos de frecuencia del alelo secundario para los polimorfismos considerados.

Una ejecución de un modelo mixto se presenta en la siguiente sección. En esta realización, la frecuencia del alelo secundario A es una suma de cuatro términos según 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 polimórfica α y una distribución binomial de la frecuencia del alelo secundario. Las αs son las fracciones de los polimorfismos que se encuentran dentro de cada uno de los cuatro casos. Cada distribución binominal tiene una probabilidad asociada, *p*, y una cobertura, *d*. La probabilidad del alelo secundario para el caso 2, por ejemplo, está dada por *f*/2.

Las realizaciones divulgadas hacen uso de "momentos factoriales" para los datos de frecuencia alélica considerados. Como se sabe bien, una media de la distribución es el primer momento. Es el valor esperado de la frecuencia del alelo secundario. La varianza es el segundo momento. Se calcula a partir del valor esperado de la frecuencia alélica al cuadrado.

Los datos de frecuencia alélica a través de todos los polimorfismos se pueden usar para calcular momentos factoriales (un primer momento factorial, un segundo momento factorial, etc.) según se muestra en la ecuación 4. Según se indica mediante estas ecuaciones, los momentos factoriales son sumas de términos por encima de i, los polimorfismos individuales en el conjunto de datos, donde existen *n* de estos polimorfismos en el conjunto de datos. Los términos que se resumen son funciones de los recuentos del alelo secundario, a_i, y las coberturas d_i.

Útilmente, los momentos factoriales tienen relaciones con los valores de α_i y p_i como las ilustradas en la ecuación 5. A partir de las probabilidades, p_i, se puede determinar la fracción fetal, *f*. Por ejemplo, $p_2 = f/2$ y p_3 es 1 - *f*/2. Así, la lógica solvente puede resolver un sistema de ecuaciones que relacionan αs y *p*s desconocidas con las expresiones de momentos factoriales para fracciones del alelo secundario a través de los múltiples polimorfismos considerados. Por supuesto, existen otras técnicas para resolver los modelos mixtos dentro del alcance de esta invención.

Es útil considerar además los fundamentos matemáticos o simbólicos de realizaciones de modelos mixtos divulgadas en la presente. Los cuatro casos de heterocigosidad descritos anteriormente sugieren el siguiente modelo mixto binomial para la distribución de ai en 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

Se describen posteriormente diversos modelos para relacionar la p_i con la fracción fetal y los grados de error de secuenciación. Los parámetros α_i se refieren a 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 etnicidad y la descendencia de los progenitores.

Para diversos casos de heterocigosidad, la ecuación anterior se puede resolver para la fracción fetal. Quizá el método más fácil de resolver para la fracción fetal sea a través de momentos factoriales en los que los parámetros de la mezcla se puedan expresar en cuanto a momentos que se puedan estimar fácilmente a partir de los datos observados.

Dadas n posiciones de SNP, los momentos factoriales se definen como sigue:

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

Ecuación 4

Los momentos factoriales se pueden relacionar con el $\{\alpha_i,p_i\}$ con

$$F_1 \approx \sum_{i=1}^{m}\alpha_i p_i^1$$
$$F_2 \approx \sum_{i=1}^{m}\alpha_i p_i^2$$
$$\cdots$$
$$F_j \approx \sum_{i=1}^{m}\alpha_i p_i^j$$
$$\cdots$$
$$F_g \approx \sum_{i=1}^{m}\alpha_i p_i^g$$

Ecuación 5

Una solución se puede identificar mediante resolución para el $\{\alpha_i,,p_i\}$ en un sistema de ecuaciones derivadas de la relación anterior Ecuación 5 cuando n > 2* (número de parámetros que se ha de estimar). Obviamente, el problema se hace mucho más difícil matemáticamente para una g superior ya que se necesitan estimar más $\{\alpha_i,p_i,\}$.

Típicamente, no es posible discriminar precisamente entre los datos del caso 1 y 2 (o el caso 3 y 4) mediante umbrales simples a fracciones fetales inferiores. Afortunadamente para el uso de modelos de casos reducidos, los datos del caso 1/2 se separan fácilmente de los datos del caso 3/4 al discriminar en el punto (2A/D)=T. Se ha encontrado que el uso de T=0,5 funciona satisfactoriamente.

Nótese que el método del modelo mixto que emplea las ecuaciones 4 y 5 hace uso de datos para todos los polimorfismos pero no explica el error de secuenciación. Métodos apropiados que separan los datos para los casos primero y segundo de los datos para los casos tercero y cuarto pueden explicar el error de secuenciación.

En ejemplos adicionales, el conjunto de datos proporcionado para un modelo mixto contiene datos solo para polimorfismos del caso 1 y el caso 2. Estos son polimorfismos para los que la madre es homocigótica. Se puede emplear una técnica liminar para retirar los polimorfismos del caso 3 y 4. Por ejemplo, los polimorfismos con frecuencias del alelo secundario mayores que un umbral particular se eliminan antes de emplear el modelo mixto. Usando datos apropiadamente filtrados y momentos factoriales como los reducidos a las ecuaciones 7 y 8, se puede calcular la fracción fetal, *f*, según se muestra en la ecuación 9. Nótese que la ecuación 7 es una reformulación de la ecuación 3 para esta ejecución de un modelo mixto. Nótese también que, en este ejemplo particular, no se conoce el error de secuenciación asociado con la lectura de la máquina. Como consecuencia, el sistema de ecuaciones debe resolver separadamente el error, *e.*

La Figura 7 muestra una comparación de los resultados usando este modelo mixto y la fracción fetal conocida (eje x) y la fracción fetal estimada. Si el modelo mixto predijera perfectamente la fracción fetal, los resultados representados seguirían la línea de puntos. No obstante, las fracciones estimadas son notablemente buenas, particularmente considerando que muchos de los datos se eliminaban antes de aplicar el modelo mixto.

Para más elaboración, están disponibles varios otros métodos para la estimación paramétrica del modelo de la Ecuación 3. En algunos casos, se puede encontrar una solución tratable al fijar derivadas a cero de la estadística chi cuadrado. En casos en los que no se pueda encontrar una solución fácil mediante diferenciación directa, puede ser eficaz la expansión por la serie de Taylor del PDF binomial u otros polinomios aproximativos. Se sabe bien que los estimadores de 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, pp. 129-156] usa la iteración de Ralph-Newton de la función de probabilidad. El método de resoluciones de momentos a partir de la Ecuación 5 se puede usar como un punto de partida para la iteración.

Bajo otra aplicación, se analiza un método para resolver modelos mixtos que implica métodos de esperanza-maximización que funcionan sobre mezclas de distribuciones beta aproximativas.

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

Considérese un modelo reducido que solo explica los casos de heterocigosidad 1 y 2. En este caso, la distribución de la mezcla se puede escribir como

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

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 la e (grado de error de secuenciación), la alfa (proporción de puntos del caso 1) y la f (fracción fetal). Donde las Fi se definen como en la Ecuación 4 anteriormente. Se elige una solución de forma cerrada para la fracción fetal para que sea la solución real 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 está entre 0 y 1.

Para calibrar el funcionamiento de los estimadores, se construyó un conjunto de datos simulado de puntos de equilibrio de Hardy-Weinberg (ai,di) con una fracción fetal diseñada para ser {1%, 3%, 5%, 10%, 15%, 20% y 25%} y un grado de error de secuenciación constante de 1%. El grado de error de 1% es el grado actualmente aceptado para las máquinas y los protocolos de secuenciación que se están usando y está de acuerdo con la gráfica de los datos del analizador genómico II de Illumina mostrados en la Figura 3 anteriormente. La ecuación 9 se aplicó a los datos y se encontró, con la excepción de un desvío ascendente de cuatro puntos, de acuerdo generalmente con la fracción fetal "conocida". De forma interesante, se estima que el grado de error de secuenciación, e, está justo por encima de 1%.

En el siguiente ejemplo de modelo mixto, se emplea de nuevo umbralización u otra técnica de filtrado para retirar datos para polimorfismos que se encuentran dentro de 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*, según se muestra en las ecuaciones 10. La Figura 8 muestra que esta versión de un modelo mixto proporcionaba resultados mejorados en comparación con el enfoque empleado en la ecuación 9.

Un enfoque similar se muestra en las ecuaciones 11 y 12. Este enfoque identifica que solo algunos errores de secuenciación se suman al recuento del alelo secundario. En cambio, solo uno de cada cuatro errores de secuenciación debe incrementar el recuento del alelo secundario. La Figura 9 muestra una concordancia notablemente buena entre las fracciones fetales real y estimada usando esta técnica.

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

Puesto que el grado de error de secuenciación de las máquinas usadas se conoce en gran medida, la desviación y la complejidad de los cálculos se pueden reducir al eliminar e como una variable que se va a resolver. Así, 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 usar el grado de error de la máquina como un parámetro conocido reduce la desviación ascendente en un punto.

Casos Modélicos (1+2), error de secuenciación conocido, Modelos de Error Mejorados

Para mejorar la desviación en el modelo, se expandió el modelo de error de las ecuaciones anteriores para tener en cuenta el hecho de que no todo episodio de error de secuenciación se sumará al recuento del alelo secundario A=ai

en el caso de heterocigosidad 1. Por otra parte, se admite el hecho de que los episodios de errores de secuenciación puedan contribuir a recuentos del caso de heterocigosidad 2. De ahí que se determine la fracción fetal F al resolver 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

lo 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 que usan el grado de error de la máquina como un parámetro conocido, potenciando los modelos de error del caso 1 y 2, reducen mucho la desviación ascendente hasta menos de un punto para una fracción fetal por debajo de 0,2.

Opciones de Ejecución

MUESTRAS

Las muestras que se usan en las realizaciones divulgadas en la presente comprenden ADN genómico que es celular o está libre de células. El ADN celular se deriva de células enteras al extraer manualmente o mecánicamente el ADN genómico de células enteras de composiciones genéticas iguales o diferentes. El ADN celular se puede derivar, por ejemplo, de células enteras de la misma composición genética derivadas 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 se derivan de un sujeto. Se conocen en la técnica métodos para extraer ADN genómico de células enteras, y difieren dependiendo de la naturaleza de la fuente.

En algunas ocasiones, puede ser ventajoso fragmentar el ADN genómico celular. La fragmentación puede ser aleatoria o puede ser específica, según se obtenga, por ejemplo, usando digestión con endonucleasas de restricción. Métodos para la fragmentación aleatoria son muy conocidos en la técnica e incluyen, por ejemplo, digestión con ADNasa limitada, tratamiento alcalino y corte físico. En ciertas realizaciones, los ácidos nucleicos de muestra se someten a fragmentación en fragmentos de aproximadamente 500 o más pares de bases, y a los que se pueden aplicar fácilmente métodos de secuenciación de última generación (NGS). En una realización, se obtienen ácidos nucleicos de muestra a partir de ADNcf, que no está sometido a fragmentación.

El ADN libre de células es ADN genómico que está presente naturalmente como una mezcla de fragmentos genómicos encontrados típicamente en líquidos biológicos, p. ej., sangre, de un sujeto. La mezcla genómica se puede derivar de células que se rompen naturalmente para liberar su contenido genómico mediante procesos biológicos, p. ej., apoptosis. Una muestra de ADNcf puede comprender ADNcf derivado de una mezcla de células de diferentes sujetos de la misma especie, de una mezcla de células procedentes de un sujeto que difieren en composición genética o de una mezcla de células procedentes de diferentes especies, p. ej., un sujeto.

Los ácidos nucleicos libres de células, incluyendo ADN libre de células, se pueden obtener mediante diversos métodos conocidos en la técnica a partir de muestras biológicas incluyendo, pero no limitadas a, plasma, suero y orina (Fan y cols., ProcNatlAcadSci 105:16266-16271 [2008]; Koide y cols., Prenatal Diagnosis 25:604-607 [2005]; Chen y cols., Nature Med. 2: 1033-1035 [1996]; Lo y cols., Lancet 350: 485-487 [1997]; Botezatu y cols., Clin Chem. 46: 1078-1084, 2000; y Su y cols., J Mol. Diagn. 6: 101-107 [2004]). Para separar ADNcf de las células, se pueden usar métodos de fraccionación, centrifugación (p. ej., centrifugación por gradiente de densidad), precipitación específica de ADN o clasificación y/o separación celular de alto rendimiento. Están disponible comercialmente estuches para la separación manual y automatizada de ADNcf (Roche Diagnostics, Indianapolis, IN, Qiagen,Valencia, CA, Macherey-Nagel, Duren, DE).

La muestra que comprende la mezcla de ácidos nucleicos a la que se aplican los métodos descritos en la presente puede ser una muestra biológica tal como una muestra de tejido, una muestra de líquido biológico o una muestra de células. En algunas realizaciones, la mezcla de ácidos nucleicos se purifica o aísla de la muestra biológica mediante

uno cualquiera de los métodos conocidos. Una muestra puede ser un polinucleótido purificado o aislado. Un líquido biológico incluye, como ejemplos no limitativos, sangre, plasma, suero, sudor, lágrimas, esputos, orina, esputos, flujo ótico, linfa, saliva, líquido cefalorraquídeo, ravages, suspensión de médula ósea, flujo vaginal, lavado transcervical, líquido cerebral, ascitis, leche, secreciones de los tractos respiratorio, intestinal y genitourinario, líquido amniótico y muestras de leucoforesis. En algunas realizaciones, la muestra es una muestra que es fácilmente obtenible mediante procedimientos no invasivos, p. ej., sangre, plasma, suero, sudor, lágrimas, esputos, orina, esputos, flujo ótico, saliva o heces. Preferiblemente, la muestra biológica es una muestra de sangre periférica, o las fracciones de plasma y suero. En otras realizaciones, la muestra biológica es un hisopo o frotis, un espécimen de biopsia o un cultivo celular. En otra realización, la muestra es una mezcla de dos o más muestras biológicas, p. ej., una muestra biológica puede comprender dos o más de una muestra de fluido biológico, una muestra de tejido y una muestra de cultivo celular. Según se usa en la presente, los términos "sangre," "plasma" y "suero" abarcan expresamente fracciones o porciones procesadas de las mismas. De forma similar, cuando una muestra se recoge de una biopsia, un hisopo, un frotis, etc., la "muestra" abarca expresamente una fracción o porción procesada derivada de la biopsia, el hisopo, el frotis, etc.

En algunas realizaciones, se pueden obtener muestras de fuentes, incluyendo, pero no limitadas a, muestras de diferentes individuos, diferentes estados de desarrollo del mismo o diferentes individuos, diferentes individuos enfermos (p. ej., individuos con cáncer o que se sospecha que tienen un trastorno genético), individuos normales, muestras obtenidas en diferentes estadios 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 patológico infeccioso (p. ej., VIH).

En una realización, la muestra es una muestra materna que se obtiene de una hembra preñada, por ejemplo una mujer embarazada. En esta ocasión, la muestra se puede analizar usando los métodos descritos en la presente para proporcionar un diagnóstico prenatal de anormalidades cromosómicas potenciales en el feto. La muestra materna puede ser una muestra de tejido, una muestra de líquido biológico o una muestra celular. Un líquido biológico incluye, como ejemplos no limitativos, sangre, plasma, suero, sudor, lágrimas, esputos, orina, esputos, flujo ótico, linfa, saliva, líquido cefalorraquídeo, ravages, suspensión de médula ósea, flujo vaginal, lavado transcervical, líquido cerebral, ascitis, leche, secreciones de los tractos respiratorio, intestinal y genitourinario y muestras de leucoforesis. En otra realización, la muestra materna es una mezcla de dos o más muestras biológicas, p. ej., una muestra biológica puede comprender dos o más de una muestra de líquido biológico, una muestra de tejido y una muestra de cultivo celular. En algunas realizaciones, la muestra es una muestra que es fácilmente obtenible mediante procedimientos no invasivos, p. ej. sangre, plasma, suero, sudor, lágrimas, esputos, orina, esputos, flujo ótico, saliva y heces. En algunas realizaciones, la muestra biológica es una muestra de sangre periférica, o las fracciones de plasma y suero. En otras realizaciones, la muestra biológica es un hisopo o un frotis, un espécimen de biopsia o un cultivo celular.

Las muestras también se pueden obtener a partir de tejidos, células u otras fuentes que contienen polinucleótidos, cultivados in vitro. Las muestras cultivadas se pueden recoger de fuentes incluyendo, pero no limitadas a, cultivos (p. ej., tejido o células) mantenidos en diferentes medios y condiciones (p. ej., pH, presión o temperatura), cultivos (p. ej., tejido o células) mantenidos durante diferentes períodos, cultivos (p. ej., tejido o células) tratados con diferentes factores o reactivos (p. ej., un candidato a fármaco o un modulador) o cultivos de diferentes tipos de tejido o células. Métodos para aislar ácidos nucleicos de fuentes biológicas son muy conocidos y diferirán dependiendo de la naturaleza de la fuente según se explica anteriormente.

POLIMORFISMOS PARA EL USO EN LA IDENTIFICACIÓN DE UNA FRACCIÓN GENÓMICA

Según se explica, se pueden usar polimorfismos para determinar la fracción fetal. Se usa en la determinación la fracción alélica y la cigosidad de uno o más polimorfismos. Ejemplos de polimorfismos útiles incluyen, sin limitación, polimorfismos de un solo nucleótido (SNPs), SNPs en tándem, eliminaciones o inserciones de múltiples bases a pequeña escala, llamadas IN-DELS (también llamados polimorfismos de eliminación-inserción o DIPs), polimorfismos de múltiples nucleótidos (MNPs), repeticiones cortas en tándem (STRs), polimorfismos de longitud de fragmentos de restricción (RFLPs), eliminaciones, incluyendo microeliminaciones, inserciones, incluyendo microinserciones, duplicaciones, inversiones, translocaciones, multiplicaciones, variantes complejas de múltiples sitios, variantes del número de copias (CNVs) y polimorfismos que comprenden cualquier otro cambio de secuencia en un cromosoma.

En algunas realizaciones, polimorfismos que se usan en el método divulgado incluyen SNPs y/o STRs. Los polimorfismos SNP pueden ser un SNP simple, SNPs en tándem. SNPs simples incluyen SNPs individuales y SNPs marcadores, es decir SNPs presentes en un haplotipo, y/o un bloque de haplotipos. En algunas realizaciones, se usan combinaciones de polimorfismos. Por ejemplo, se pueden detectar diferencias en el número de copias mediante comparación de una combinación de secuencias polimórficas que comprende uno o más SNPs y uno o más STRs.

En general, cualquier sitio polimórfico que pueda ser abarcado por las lecturas generadas por los métodos de secuenciación descritos en la presente se puede usar para identificar la fracción genómica en muestras que comprenden ADN de diferentes genomas. Secuencias polimórficas útiles para poner en práctica los métodos de la

invención están disponibles de una variedad de bases de datos públicamente accesibles, que se están expandiendo continuamente. Por ejemplo, bases de datos útiles incluyen sin limitación la base de datos de SNP humanos en la dirección de Internet wi.mit.edu, la página de inicio de dbSNP del NCBI en la dirección de Internet ncbi.nlm.nih.gov, la dirección de Internet lifesciences.perkinelmer.com, la base de datos de SNP humanos de Celera en la dirección de Internet celera.com, la base de datos de SNP del Genome Analysis Group (GAN) en la dirección de Internet gan.iarc.fr, la base de datos de repeticiones en tándem cortas (STR) del ATCC en la dirección de Internet atcc.org y la base de datos HapMap en la dirección de Internet hapmap.org.

El número de polimorfismos que se puede usar 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, 1000 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 se puede genotipar en una muestra procedente de un sujeto humano puede ser al menos aproximadamente 10 millones de SNPs, así como muchos otros tipos de polimorfismos que están presentes en un genoma humano cualquiera. En algunas realizaciones, la identificación de uno o más polimorfismos en un primer genoma de una muestra que comprende una mezcla de ADN, p. ej. ADNcf, de un primer y un segundo genoma se realiza mediante la secuenciación del genoma completo usando un método de NGS según se describe en la presente. En algunas realizaciones, el método de secuenciación del genoma completo es un método de NGS que identifica las secuencias polimórficas mediante secuenciación masivamente paralela de moléculas de ácido nucleico amplificadas clonalmente o mediante secuenciación masivamente paralela de moléculas de ácido nucleico individuales, es decir, secuenciación de moléculas individuales.

APLICACIONES

La fracción de ácido nucleico que se origina de cada una de dos fuentes genómicas distintas en una muestra se puede usar con diversos propósitos. En diversas realizaciones descritas en la presente, la fracción de ADN fetal en ADN libre de células de una muestra de material se usa para facilitar diagnósticos prenatales y para ayudar a tomar decisiones relativas al tratamiento de los embarazados. En otras realizaciones, los genomas considerados no son materno y fetal. Se presentan posteriormente diversos ejemplos de fuentes genómicas para determinar la presencia de genoma fraccionado.

Se pueden usar ADN y ARN fetales libres de células que circulan en sangre materna para el diagnóstico prenatal no invasivo (NIPD) inicial de un número creciente de afecciones genéticas, tanto para el manejo del embarazo como para ayudar a tomar una decisión reproductiva. Pequeñas cantidades de ADN fetal circulatorio están presentes en la corriente sanguínea materna durante el embarazo (Lo y cols., Lancet 350:485-487 [1997]). Creyendo que se origina a partir de células placentarias muertas, se ha mostrado que el ADN fetal libre de células consiste en fragmentos cortos con una longitud típicamente menor de 200 pb Chan y cols., ClinChem 50:88-92 [2004]), que se pueden discernir tan temprano como a las 4 semanas de gestación (Illanes y cols., Early Human Dev 83:563-566 [2007]), y se sabe que se depuran de la circulación materna horas después del parto (Lo y cols., Am J Hum Genet 64:218-224 [1999]). Además de fragmentos de ADNcf, también se pueden discernir fragmentos de ARN fetal libre de células (ARNcf) en la corriente sanguínea materna, que se originan a partir de genes que se transcriben en el feto o la placenta. La extracción y el análisis posterior de estos elementos genéticos fetales procedentes de una muestra de sangre materna ofrece nuevas oportunidades para el NIPD.

Según se explica, los métodos divulgados determinan la fracción de un segundo genoma en una muestra biológica. Opcionalmente, los métodos determinan la presencia o ausencia de un número de trastornos en una muestra de sangre que comprende una mezcla de ADN (tal como ADNcf) de un primer y un segundo genoma. En algunas realizaciones, la determinación de la fracción fetal puede comprender (a) secuenciar genómicamente al menos una porción de la mezcla de ADNcf para obtener una pluralidad de marcadores de secuencia; (b) determinar en la pluralidad de marcadores de secuencia la presencia o ausencia de múltiples polimorfismos, y (c) asociar los múltiples polimorfismos con el primer y/o el segundo genoma en la mezcla. En realizaciones preferidas, la mezcla no está enriquecida para los múltiples polimorfismos. La identificación de los múltiples polimorfismos en la mezcla de ADN se realiza al comparar la secuencia de los marcadores asignados obtenidos mediante el método de secuenciación del genoma completo con múltiples polimorfismos de referencia, según se describe en la presente.

En una realización divulgada anteriormente, el primer genoma es un genoma fetal y el segundo genoma es un genoma materno. En otra realización, el primer genoma es un genoma de una célula no afectada y el segundo genoma es un genoma procedente de una célula afectada, p. ej. una célula cancerosa. En algunas realizaciones, las células afectadas y no afectadas se derivan del mismo sujeto. Por ejemplo, la célula afectada puede ser una célula cuyo genoma ha sido alterado por un trastorno. En algunas realizaciones, el trastorno es un trastorno monogénico. En otras realizaciones, el trastorno es un trastorno poligénico. Los trastornos se pueden identificar por un solo polimorfismo, p. ej. un SNP marcador, o por múltiples polimorfismos presentes en un haplotipo. En algunas realizaciones, los múltiples polimorfismos identificados según el presente método están presentes en un bloque de haplotipos.

Los trastornos que se pueden identificar con la ayuda del presente método son trastornos genéticos, que son dolencias provocadas al menos en parte por anormalidades en genes o cromosomas. El conocimiento de una fracción fetal en una muestra puede ayudar a identificar estos trastornos en un contexto prenatal. Trastornos identificados mediante el presente método incluyen trastornos monogénicos, es decir de un solo gen, y trastornos poligénicos, es decir, complejos. Trastornos de un solo gen incluyen dominante autosómico, recesivo autosómico, dominante ligado al cromosoma X, recesivo ligado al cromosoma X, y ligado al cromosoma Y.

En trastornos dominantes autosómicos, solo una copia mutada del gen será necesaria para una persona que vaya a estar afectada por el trastorno. Típicamente, un sujeto afectado tiene un progenitor afectado, y existe una posibilidad de 50% de que la descendencia herede el gen mutado. Las afecciones que son dominantes autosómicas tienen a veces una penetración reducida, lo que significa que aunque solo se necesite una copia mutada, no todos los individuos que hereden esa mutación van a desarrollar la enfermedad. Ejemplos de trastornos dominantes autosómicos que se pueden identificar mediante el presente método incluyen sin limitación hipercolesterolemia familiar, esferocitosis hereditaria, síndrome de Marfan, neurofibromatosis tipo 1, cáncer colorrectal no polipósico hereditario y exostosis múltiple hereditaria y enfermedad de Huntington.

Trastornos recesivos autosómicos detectados usando el presente método incluyen anemia drepanocítica, fibrosis quística, enfermedad de Tay-Sachs, enfermedad de Tay-Sachs, mucopolisacaridosis, enfermedades del almacenamiento del glucógeno y galactosemia. Trastornos ligados al cromosoma X detectados por el presente método incluyen distrofia muscular de Duchenne y hemofilia. En los trastornos recesivos autosómicos, se deben mutar dos copias del gen para que un sujeto esté afectado por un trastorno autosómico recesivo. Habitualmente, un sujeto afectado tiene progenitores no afectados que tienen una sola copia del gen mutado (y se denominan portadores). Dos personas no afectadas que tengan cada una una copia del gen mutado tienen un 25% de posibilidades con cada embarazo de tener un hijo afectado por el trastorno. Ejemplos de este tipo de trastorno que se pueden identificar mediante el presente método incluyen fibrosis quística, drepanocitosis, enfermedad de Tay-Sachs, enfermedad de Niemann-Pick, atrofia del músculo espinoso y síndrome de Roberts. Ciertos otros fenotipos, tales como cerumen húmedo frente a seco, también se determinan de un modo recesivo autosómico. Los trastornos dominantes ligados al cromosoma X están provocados por mutaciones en genes sobre el cromosoma X. Solo unos pocos trastornos tienen este patrón hereditario, siendo un ejemplo principal el raquitismo hipofosfatémico ligado al cromosoma X. Tanto los hombres como las mujeres están afectados en estos trastornos, siendo los hombres más gravemente afectados que las mujeres. Algunas afecciones dominantes ligadas al cromosoma X tales como síndrome de Rett, incontinencia pigmentaria tipo 2 y síndrome de Aicardi habitualmente son letales en hombres, y por lo tanto se observan predominantemente en mujeres. Las excepciones a este hallazgo son casos extremadamente raros en los que niños con síndrome de Klinefelter (47,XXY) también heredan una afección dominante ligada al cromosoma X y exhiben síntomas más similares a los de una mujer en cuanto a la gravedad de la enfermedad. La posibilidad de transmitir un trastorno dominante ligado al cromosoma X difieren entre hombres y mujeres. Ninguno de los hijos de un hombre con un trastorno dominante ligado al cromosoma X está afectado (puesto que reciben su cromosoma Y del padre) y todas sus hijas heredarán la afección. Una mujer con un trastorno dominante ligado al cromosoma X tiene un 50% de posibilidades de tener un feto afectado con cada embarazo, aunque se debe apuntar que en casos tales como la incontinencia pigmentaria solo las descendientes femeninas son generalmente viable. Además, aunque estas afecciones no alteran la fertilidad de por sí, los individuos con síndrome de Rett o síndrome de Aicardi raramente se reproducen.

El presente método también puede facilitar la identificación de polimorfismos asociados con trastornos ligados al cromosoma X. Las afecciones recesivas ligadas al cromosoma X también están provocadas por mutaciones en genes sobre el cromosoma X. Los hombres están afectados más frecuentemente que las mujeres, y la posibilidad de transmitir el trastorno difiere entre hombres y mujeres. Los hijos de un hombre con un trastorno recesivo ligado al cromosoma X no estarán afectados, y sus hijas portarán una copia del gen mutado. Una mujer que sea una portadora de un trastorno recesivo ligado al cromosoma X (X^RX^r) tiene un 50% de posibilidades de tener hijos que estén afectados y una posibilidad de 50% de tener hijas que porten una copia del gen mutado y por lo tanto sean portadoras. Afecciones recesivas ligadas al cromosoma X incluyen sin limitación las enfermedades graves hemofilia A, distrofia muscular de Duchenne y síndrome de Lesch-Nyhan así como afecciones comunes y menos graves tales como alopecia androgénica y daltonismo al rojo-verde. Las afecciones recesivas ligadas al cromosoma X se pueden manifestar a veces en mujeres debido a inactivación sesgada del cromosoma X o monosomía X (síndrome de Turner).

Los trastornos ligados al cromosoma Y están provocados por mutaciones en el cromosoma Y. Debido a que los varones heredan un cromosoma Y de sus padres, todos los hijos varones de un padre afectado estarán afectados. Debido a que las mujeres heredan un cromosoma X de sus padres, la descendencia femenina de padres afectados nunca estará afectada. Puesto que el cromosoma Y es relativamente pequeño y contiene muy pocos genes, existen relativamente pocos trastornos ligados al cromosoma Y. A menudo, los síntomas incluyen esterilidad, que se puede evitar con la ayuda de algunos tratamientos de fertilidad. Ejemplos son esterilidad masculina e hipertricosis lanuginosa.

Según se explica, los métodos divulgados para detectar fracciones genómicas en una muestra se pueden usar para facilitar la detección de aneuploidía a partir de muestras de material. En algunas realizaciones, la aneuploidía es una trisomía o monosomía cromosómica completa, o una trisomía o monosomía parcial. Las aneuploidías parciales están provocadas por pérdida o ganancia de parte de un cromosoma, y abarcan desequilibrios cromosómicos resultantes

de translocaciones desequilibradas, inversiones, eliminaciones e inserciones desequilibradas. Con mucho, la aneuploidía conocida más común compatible con la vida es la trisomía 21, es decir, el síndrome de Down (DS), que está provocado por la presencia de parte o la totalidad del cromosoma 21. Raramente, el DS puede estar provocado por un defecto heredado o esporádico por el que una copia adicional de la totalidad o parte del cromosoma 21 se une a otro cromosoma (habitualmente el cromosoma 14) para formar un solo cromosoma aberrante. El DS está asociado con deterioro intelectual, graves dificultades de aprendizaje y exceso de mortalidad provocado por problemas de salud a largo plazo tales como cardiopatía. Otras aneuploidías con significación clínica conocida incluyen el síndrome de Edward (trisomía 18) y el síndrome de Patau (trisomía 13), que frecuentemente son letales en los primeros meses de vida. Anormalidades asociadas con el número de cromosomas sexuales también son conocidas e incluyen monosomía X, p. ej. síndrome de Turner (XO) y síndrome de triple X (XXX) en nacimientos femeninos y síndrome de Kleinefelter (XXY) y síndrome XYY en nacimientos masculinos, que están todos asociados con diversos fenotipos incluyendo esterilidad y reducción en las capacidades intelectuales. La monosomía X [45,X] es una causa común de abortos involuntarios tempranos que explica aproximadamente 7% de los abortos espontáneos. Basándose en la frecuencia de nacidos vivos de 45,X (también llamado síndrome de Turner) de 1-2/10.000, se estima que menos de 1% de concebidos 45,X sobrevivirán a término. Aproximadamente 30% de los pacientes con síndrome de Turner son mosaicos tanto con una línea celular 45,X como bien con una línea celular 46,XX o bien una que contiene un cromosoma X reorganizado (Hook y Warburton 1983). El fenotipo en un recién nacido vivo es relativamente leve considerando la alta letalidad embrionaria y se ha establecido como hipótesis que posiblemente todas las niñas nacidas vivas con síndrome de Turner portan una línea celular que contiene dos cromosomas sexuales. La monosomía X se puede producir en mujeres como 45,X o como 45,X/46XX, y en hombres como 45,X/46XY. Se sugiere generalmente que las monosomías autosómicas son incompatibles con la vida; sin embargo, existe un cierto número de informes citogenéticos que describen monosomía completa de un cromosoma 21 en niños nacidos vivos (Vosranoval y cols., Molecular Cytogen. 1:13 [2008]; Joosten y cols., Prenatal Diagn. 17:271-5 [1997]. El método de la invención se puede usar para ayudar en el diagnóstico de estas y otras anormalidades cromosómicas prenatalmente.

Según algunas realizaciones, la fracción fetal puede ser útil para determinar la presencia o ausencia de trisomías cromosómicas de uno cualquiera de los cromosomas 1-22, X e Y. Ejemplos de trisomías cromosómicas que se pueden detectar según el presente método incluyen sin limitación trisomía 21 (T21; síndrome de Down), trisomía 18 (T18; síndrome de Edward), trisomía 16 (T16), trisomía 20 (T20), trisomía 22 (T22; síndrome de ojo de gato), trisomía 15 (T15; síndrome de PraderWilli), trisomía 13 (T13; síndrome de Patau), trisomía 8 (T8; síndrome de Warkany), trisomía 9 y las trisomías XXY (síndrome de Kleinefelter), XYY o XXX. Las trisomías completas de otros autosomas que existen en un estado que no es de mosaico son letales, pero pueden ser compatibles con la vida cuando están presentes en estado de mosaico. Se apreciará que diversas trisomías completas, ya existan en un estado de mosaico o no, y las trisomías parciales se pueden determinar en ADNcf fetal según el presente método. El método de la invención se puede usar para ayudar en la determinación de estas y otras anormalidades cromosómicas prenatalmente.

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

Los métodos divulgados en la presente también se pueden usar para ayudar a determinar monosomía cromosómica X, monosomía cromosómica 21 y monosomías parciales tales como monosomía 13, monosomía 15, monosomía 16, monosomía 21 y monosomía 22, que se sabe que están implicadas en abortos espontáneos. La monosomía parcial de cromosomas típicamente implicados en aneuploidía completa también se puede determinar mediante el presente método. El método de la invención se puede usar para ayudar en la determinación de estas y otras anormalidades cromosómicas prenatalmente. Ejemplos no limitativos de síndromes de eliminación que se pueden determinar según el presente método incluyen síndromes provocados por eliminaciones parciales de cromosomas. Ejemplos de eliminaciones parciales que se pueden determinar según el método incluyen sin limitación eliminaciones parciales de los cromosomas 1, 4, 5, 7, 11, 18, 15, 13, 17, 22 y 10, que se describen en lo siguiente. El método de la invención se puede usar para ayudar en la determinación de estas y otras anormalidades cromosómicas prenatalmente.

El síndrome de eliminación 1q21.1 o microeliminación (recurrente) 1q21.1 es una aberración rara del cromosoma 1. Junto al síndrome de eliminación, también hay un síndrome de duplicación 1q21.1. Aunque hay una parte del ADN que se pierde con el síndrome de eliminación en un lugar particular, existen dos o tres copias de una parte similar del ADN en el mismo lugar con el síndrome de duplicación. La bibliografía se refiere tanto a la eliminación como a la duplicación como las variaciones del número de copias (CNV) de 1q21.1. La eliminación 1q21.1 se puede asociar con el síndrome de TAR (trombocitopenia con ausencia de radios).

El síndrome de Wolf-Hirschhorn (WHS) (OMIN #194190) es un síndrome de eliminación de un gen contiguo asociado con una eliminación hemicigótica del cromosoma 4p16.3. El síndrome de Wolf-Hirschhorn es un síndrome de malformación congénita caracterizado por deficiencia en el crecimiento pre- y posnatal, discapacidad de desarrollo de grado variable, rasgos craneofaciales característicos (apariencia de 'casco de guerrero griego' de la nariz, frente alta, entrecejo prominente, hipertelorismo, cejas arqueadas, ojos saltones, epicantos, surco subnasal corto, boca marcada con comisuras hacia abajo, y micrognatia) y un trastorno convulsivo.

La eliminación parcial del cromosoma 5, también conocida como 5p- o 5p menos, y llamada síndrome de Cris du Chat (OMIN#123450), está provocada por una eliminación del brazo corto (brazo p) del cromosoma 5 (5p15.3-p15.2). Los niños con esta afección tienen a menudo un llanto agudo que suena como un gato. El trastorno se caracteriza por discapacidad intelectual y retraso en el desarrollo, tamaño de cabeza pequeño (microcefalia), bajo peso al nacer y tono muscular débil (hipotonía) en la lactancia, rasgos faciales marcados y posiblemente defectos cardíacos.

El síndrome de Williams-Beuren también conocido como síndrome de eliminación en el cromosoma 7q11.23 (OMIN 194050) es un síndrome de eliminación del gen contiguo que da como resultado un trastorno multisistémico provocado por eliminación hemicigótica de 1,5 a 1,8 Mb en el cromosoma 7q11.23, que contiene aproximadamente 28 genes.

El síndrome de Jacobsen, también conocido como trastorno de eliminación 11q, es un trastorno congénito raro resultante de la eliminación de una región terminal del cromosoma 11 que incluye la banda 11q24.1. Puede provocar incapacidades intelectuales, una apariencia facial distintiva y una variedad de problemas físicos incluyendo defectos cardíacos y un trastorno hemorrágico.

La monosomía parcial del cromosoma 18, conocida como monosomía 18p, es un trastorno cromosómico raro en el que se elimina la totalidad o parte del brazo corto (p) del cromosoma 18 (monosómico). El trastorno se caracteriza típicamente por estatura corta, grados variables de retardo mental, retrasos en el habla, malformaciones de la región craneal y facial (craneofacial) y/o anormalidades físicas adicionales. Los defectos craneofaciales asociados pueden variar mucho en alcance y gravedad de caso a caso.

Afecciones provocadas en la estructura o el número de copias del cromosoma 15 incluyen síndrome de Angelman y síndrome de Prader-Willi, que implican una pérdida de actividad génica en la misma parte del cromosoma 15, la región 15q11-q13. Se apreciará que varias translocaciones y microeliminaciones pueden ser asintomáticas en el progenitor portador, y sin embargo pueden provocar una enfermedad genética importante en la descendencia. Por ejemplo, una madre sana que tenga la microeliminación 15q11-q13 puede parir un niño con síndrome de Angelman, a un trastorno neurodegenerativo grave. Así, el presente método se puede usar para identificar esta eliminación parcial y otras eliminaciones en el feto. El método de la invención se puede usar para ayudar en la identificación de estas y otras anormalidades cromosómicas prenatalmente.

La monosomía parcial 13q es un trastorno cromosómico raro que resulta cuando se pierde un brazo largo (q) del cromosoma 13 (monosómico). Los niños nacidos con monosomía parcial 13q pueden exhibir bajo peso al nacer, malformaciones de la cabeza y la cara (región craneofacial), anormalidades esqueléticas (especialmente de las manos y los pies) y otras anormalidades físicas. El retardo mental es característico de esta afección. La tasa de mortalidad durante la lactancia es alta entre individuos nacidos con este trastorno. Casi todos los casos de monosomía parcial 13q se producen aleatoriamente por razones no evidentes (esporádicos).

El síndrome de Smith-Magenis (SMS - OMIM #182290) está provocado por una eliminación, o pérdida de material genético, en una copia del cromosoma 17. Este síndrome bien conocido está asociado con retraso del desarrollo, retardo mental, anomalías congénitas tales como defectos cardíacos y renales, y anormalidades neuroconductuales tales como perturbaciones graves del sueño y comportamiento autolesivo. El síndrome de Smith-Magenis (SMS) está provocado en la mayoría de los casos (90%) por una eliminación intersticial de 3,7 Mb en el cromosoma 17p11.2.

El síndrome de eliminación 22q11.2, también conocido como síndrome de DiGeorge, es un síndrome provocado por la eliminación de un pequeño fragmento del cromosoma 22. La eliminación (22 q11.2) se produce cerca del medio del cromosoma en el brazo largo de uno del par del cromosoma. Las características de este síndrome pueden variar ampliamente, incluso entre miembros de la misma familia, y afectan a muchas partes del cuerpo. Signos y síntomas característicos pueden incluir defectos de nacimiento tales como cardiopatía congénita, defectos en el paladar, la mayoría relacionados con problemas neuromusculares con problemas de cerramiento (insuficiencia velofaríngea), discapacidades de aprendizaje, diferencias leves en las características faciales, e infecciones recurrentes. Las microeliminaciones en la región cromosómica 22q11.2 están asociadas con un riego de esquizofrenia incrementado de 20 a 30 veces.

Las eliminaciones en el brazo corto del cromosoma 10 están asociadas con un fenotipo similar al síndrome de DiGeorge. La monosomía parcial del cromosoma 10p es rara pero se ha observado en una porción de pacientes que muestran características del síndrome de DiGeorge

En una realización, el método se usa para determinar monosomías parciales incluyendo pero no limitadas a monosomía parcial de los cromosomas 1, 4, 5, 7, 11, 18, 15, 13, 17, 22 y 10, p. ej. monosomía parcial 1q21.11, monosomía parcial 4p16.3, monosomía parcial 5p15.3-p15.2, monosomía parcial 7q11.23, monosomía parcial 11q24.1, monosomía parcial 18p, monosomía parcial del cromosoma 15 (15q11-q13), monosomía parcial 13q, monosomía parcial 17p11.2, monosomía parcial del cromosoma 22 (22q11.2) y monosomía parcial 10p también se pueden determinar usando el método. El método de la invención se puede usar para ayudar en la determinación de estas y otras anormalidades cromosómicas prenatalmente.

Otras monosomías parciales que se pueden be determinar según el método incluyen translocación desequilibrada t(8;11)(p23.2;p15.5); microeliminación 11q23; eliminación 17p11.2; 22q13.3; microeliminación Xp22.3; eliminación 10p14; microeliminación 20p, [del(22)(q11.2q11.23)], eliminaciones 7q11.23 y 7q36; eliminación 1p36; microeliminación 2p; neurofibromatosis tipo 1 (microeliminación 17q11.2), eliminación Yq; microeliminación 4p16.3; microeliminación 1p36.2; eliminación 11q14; microeliminación 19q13.2; Rubinstein-Taybi (microeliminación 16 p13.3); microeliminación 7p21; síndrome de Miller-Dieker (17p13.3); y microeliminación 2q37. Las eliminaciones parciales pueden ser pequeñas eliminaciones de parte de un cromosoma, o pueden ser microeliminaciones de un cromosoma cuando se pueda producir la eliminación de un solo gen. El método de la invención se puede usar para ayudar en la determinación de estas y otras anormalidades cromosómicas prenatalmente.

Se han identificado varios síndromes de duplicación provocados por la duplicación de parte de los brazos cromosómicas (véase OMIN [Online Mendelian Inheritance in Man observado en Internet en ncbi.nlm.nih.gov/omim]). En una realización, el presente método se puede usar para determinar la presencia o ausencia de duplicaciones y/o multiplicaciones de segmentos de uno cualquiera de los cromosomas 1-22, X e Y. Ejemplos no limitativos de síndromes de duplicaciones que se pueden determinar según el presente método incluyen duplicaciones de parte de los cromosomas 8, 15, 12 y 17, que se describen en lo siguiente.

El síndrome de duplicación 8p23.1 es un trastorno genético raro provocado por una duplicación de una región procedente del cromosoma 8 humano. Este síndrome de duplicación tiene una prevalencia estimada de 1 en 64.000 nacimientos y es el recíproco del síndrome de eliminación 8p23.1. La duplicación 8p23.1 está asociada con un fenotipo variable que incluye uno o más de retardo en el habla, retardo en el desarrollo, dismorfismo leve, con frente prominente y cejas arqueadas, y cardiopatía congénita (CHD).

El síndrome de duplicación del cromosoma 15q (Dup15q) es un síndrome clínicamente identificable que resulta de duplicaciones del cromosoma 15q11-13.1. Habitualmente, los niños con Dup15q tienen hipotonía (bajo tono muscular), retardo del crecimiento; pueden nacer con labio y/o paladar leporino o malformaciones del corazón, los riñones u otros órganos; muestran algún grado de retardo/discapacidad cognitivos (retardo mental), retardos en el habla y el lenguaje y trastornos de procesamiento sensorial.

El síndrome de Pallister Killian es un resultado de material del cromosoma nº 12 adicional. Habitualmente, existe una mezcla de células (mosaicismo), algunas con material del nº 12 adicional, y algunas que son normales (46 cromosomas sin el material del nº 12 adicional). Los niños con este síndrome tienen muchos problemas incluyendo retardo mental grave, bajo 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 sanitarios incluyen convulsiones, mala alimentación, rigidez articular, cataratas en la edad adulta, pérdida de audición y defectos cardíacos. Las personas con Pallister Killian tienen una vida reducida.

Los individuos con la afección genética denominada dup(17)(p11.2p11.2) o dup 17p portan información genética adicional (conocida como una 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 conocida con solo una pocas docenas de casos presentados en la bibliografía médica. Los pacientes que tienen esta duplicación tienen a menudo bajo tono muscular, mala alimentación y fallo de crecimiento durante la lactancia, y también presentan desarrollo retardado de los objetivos motores y verbales. Muchos individuos que tienen PTLS tienen dificultad con el procesamiento de la articulación y el lenguaje. Además, los pacientes pueden tener características conductuales similares a las de las 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 región grande en el cromosoma 17p12 que incluye el gen PMP22 provoca la enfermedad de Charcot-Marie Tooth.

Las CNV se han asociado con muertes fetales. Sin embargo, debido a las limitaciones inherentes de la citogenética convencional, se cree que la contribución de las CNV a las muertes fetales está subrepresentada (Harris y cols., PrenatalDiagn 31:932-944 [2011]). Los presentes métodos son útiles para ayudar en la determinación de la presencia de aneuploidías parciales, p. ej. eliminaciones y multiplicaciones de segmentos cromosómicos, y se pueden usar para ayudar a identificar y determinar la presencia o ausencia de CNV que estén asociadas con muertes fetales.

El presente método también puede ayudar a identificar polimorfismos asociados con trastornos genéticos que son complejos, multifactoriales o poligénicos, significando que probablemente estén asociados con los efectos de múltiples genes en combinación con el estilo de vida y factores medioambientales. Trastornos multifactoriales incluyen, por ejemplo, cardiopatía y diabetes. Aunque los trastornos complejos a menudo se agrupan en familias, no tienen un patrón claro de herencia. En un linaje, las enfermedades poligénicas tienden a "correr en familias", pero la herencia no es simple como en las enfermedades mendelianas. Componentes medioambientales fuertes se asocian a menudo con muchos trastornos complejos, p. ej., la presión sanguínea. El presente método se puede usar para identificar polimorfismos que estén asociados con trastornos poligénicos incluyendo, pero no limitados a, asma, enfermedades autoinmunitarias tales como esclerosis múltiple, cánceres, ciliopatías, paladar leporino, diabetes, cardiopatía, hipertensión, enteropatía inflamatoria, retardo mental, trastorno del estado de ánimo, obesidad, error de refracción y esterilidad. En algunas realizaciones, los polimorfismos son SNPs. En otras realizaciones, los polimorfismos son STRs. En otras realizaciones más, los polimorfismos son una combinación de SNPs y STRs.

En una realización, la identificación de las secuencias polimórficas asociadas con trastornos comprende secuenciar al menos una porción del genoma celular correspondiente al segundo genoma en la mezcla de ADNcf. La identificación de secuencias polimórficas aportadas por un primer genoma se realiza al determinar la secuencia en múltiples sitios polimórficos en una primera muestra que contiene moléculas de ADN derivadas esencialmente de solo un segundo genoma, determinar la secuencia en los múltiples sitios polimórficos correspondientes en una segunda muestra que contiene una mezcla de moléculas de ADN derivadas de un primer y un segundo genoma, y comparar las secuencias polimórficas determinadas en ambas muestras identificando de ese modo múltiples polimorfismos 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 el primer genoma, se realiza al determinar la secuencia en múltiples sitios polimórficos en una muestra de capa leucocitaria materna, es decir una muestra que contiene moléculas de ADN derivadas esencialmente de solo un segundo genoma, determinar la secuencia en los múltiples sitios polimórficos correspondientes en una muestra de plasma purificada, es decir una segunda muestra que contiene una mezcla de moléculas de ADNcf derivadas de los genomas fetal y materno, y comparar las secuencias polimórficas determinadas en ambas muestras para identificar múltiples polimorfismos fetales. En una realización, el primer genoma es un genoma fetal y el segundo genoma es un genoma materno. En otra realización, el primer genoma es un genoma de una célula no afectada y el segundo genoma es una genoma procedente de una célula afectada. En algunas realizaciones, las células afectadas y no afectadas se derivan del mismo sujeto. Por ejemplo, la célula afectada puede ser una célula cuyo genoma ha sido alterado por un trastorno.

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

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

Cánceres que se pueden identificar y/o comprobar según el método incluyen tumores sólidos, así como tumores y/o enfermedades malignas hematológicos. Diversos cánceres que se van a tratar incluyen sarcomas, carcinomas y adenocarcinomas no limitados a cáncer de mama, cáncer de pulmón, cáncer colorrectal, cáncer pancreático, cáncer ovárico, cáncer de próstata, carcinoma renal, hepatoma, cáncer cerebral, melanoma, mieloma múltiple, linfoma, linfoma de Hodgkin, linfoma no hodgkiniano, linfomas infantiles y linfomas de origen linfocítico y cutáneo, leucemia, leucemia infantil, leucemia drepanocítica, leucemia linfocítica aguda, leucemia mielocítica aguda, leucemia linfocítica crónica, leucemia mielocítica crónica, leucemia mielógena crónica y leucemia mastocítica, neoplasmas mieloides, neoplasmas mastocíticos, tumor hematológico y tumor linfoide, incluyendo lesiones metastásicas en otros tejidos u órganos distantes de la zona del tumor primario.

Los métodos de la presente divulgación son útiles, por ejemplo, en el contexto del diagnóstico o la determinación de un pronóstico en una afección patológica que se sabe que está asociada con un haplotipo o haplotipos específicos, para determinar nuevos haplotipos, y para detectar asociaciones del haplotipo con sensibilidades a productos farmacéuticos. La asociación de múltiples secuencias polimórficas con múltiples trastornos se puede determinar a partir de la identidad de una sola secuencia polimórfica para cada uno de los múltiples trastornos. Alternativamente, la asociación de múltiples secuencias polimórficas con múltiples trastornos se puede determinar a partir de la identidad de múltiples secuencias polimórficas para cada uno de los múltiples trastornos.

Las técnicas de genotipado convencionales se han limitado a identificar polimorfismos en regiones genómicas cortas de una pocas kilobases, y la identificación de los haplotipos se ha basado en datos familiares y una estimación estadística usando algoritmos informáticos. La secuenciación del genoma completo permite la identificación de haplotipos al identificar directamente los polimorfismos en un genoma. La identificación de los haplotipos según diversas realizaciones no está limitada por la distancia intermedia entre polimorfismos. En algunas realizaciones, un método comprende la secuenciación del genoma completo de ADN celular materno. El ADN celular materno se puede obtener a partir de una muestra biológica carente de ADN genómico fetal. Por ejemplo, el ADN materno se puede obtener a partir de la capa leucocitaria de una sangre materna. Se pueden determinar haplotipos que comprenden una pluralidad de secuencias polimórficas que abarcan cromosomas enteros. En una realización, los haplotipos fetales se comparan con haplotipos asociados a un trastorno conocido, y basándose en una coincidencia del haplotipo fetal con uno cualquiera de los haplotipos asociados al trastorno conocido, indica que el feto tiene el trastorno o que el feto es propenso al trastorno. Los haplotipos fetales también se pueden comparar con haplotipos asociados con sensibilidad o insensibilidad al tratamiento del polimorfismo específico. La comparación de los haplotipos fetales identificados con bases de datos de haplotipos conocidos permite el diagnóstico y/o el pronóstico de un trastorno. Cualquier muestra biológica que comprenda una mezcla de ADNcf fetal y materno se puede usar para determinar la presencia o ausencia del trastorno fetal. Preferiblemente, la muestra biológica se selecciona de sangre, o fracciones de la misma incluyendo

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

En una realización, la divulgación proporciona un método para determinar la presencia o ausencia de múltiples trastornos fetales, que comprende (a) obtener una muestra de sangre materna que comprende una mezcla libre de células de ADN fetal y materno, (b) secuenciar el genoma completo de al menos una porción de la mezcla libre de células de ADN fetal y materno, obteniendo de ese modo una pluralidad de marcadores de secuencia; (c) determinar múltiples polimorfismos fetales en los marcadores de secuencia, y (d) determinar la presencia o ausencia de múltiples trastornos fetales. Ejemplos de múltiples trastornos fetales que se pueden identificar según el presente método incluyen trastornos monogénicos y poligénicos descritos en la presente. El método de la invención se puede usar para ayudar en la determinación de estos y otros trastornos fetales prenatalmente.

En una realización, la divulgación proporciona un método para determinar la presencia o ausencia de múltiples trastornos fetales que comprende identificar múltiples polimorfismos fetales asociados con múltiples haplotipos relacionados con trastornos. En algunas realizaciones, cada uno de los haplotipos comprende al menos 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, p. ej. todos los polimorfismos SNP marcadores, o pueden ser una combinación de polimorfismos, p. ej. SNPs de marcadores y eliminaciones de marcadores. En una realización, los polimorfismos son SNPs marcadores. En otra realización, los polimorfismos son STRs marcadores. En otra realización más, los polimorfismos son una combinación de SNPs marcadores y STRs marcadores. Los polimorfismos marcadores pueden estar en regiones codificantes y/o no codificantes del genoma. La identificación de los polimorfismos se realiza mediante secuenciación del genoma completo usando tecnologías de NGS según se describe anteriormente. El método de la invención se puede usar para ayudar en la determinación de estos y otros trastornos fetales prenatalmente.

La divulgación proporciona un método para identificar variaciones del número de copias (CNV) como polimorfismos de una secuencia de interés en una muestra de prueba que comprende una mezcla de ácidos nucleicos derivados de dos genomas diferentes, y que se sabe o se sospecha que difieren en la cantidad de una o más secuencias de interés. Variaciones del número de copias determinadas mediante el método de la invención incluyen ganancias o pérdidas de cromosomas enteros, alteraciones que implican segmentos cromosómicos muy grandes que son microscópicamente visibles, y una abundancia de variación submicroscópica del número de copias de segmentos de ADN que varían en tamaño de kilobases (kb) a megabases (Mb). El método de la invención se puede usar para ayudar en la identificación de estas y otras CNVs prenatalmente.

Las CNV en el genoma humano influyen significativamente en la diversidad humana y la predisposición a la enfermedad (Redon y cols., Nature 23:444-454 [2006], Shaikhy cols. Genome Res 19:1682-1690 [2009]). Se ha sabido que las CNVs contribuyen a una enfermedad genética a través de diferentes mecanismos, dando como resultado bien un desequilibrio de la dosificación génica o bien una alteración génica en la mayoría de los casos. Además de su correlación directa con trastornos genéticos, se sabe que las CNVs median en cambios fenotípicos que pueden ser perjudiciales. Recientemente, varios estudios han presentado un incremento en la carga de CNVs raras o “de novo” en trastornos complejos tales como autismo, ADHD y esquizofrenia en comparación con controles normales, destacando la patogenicidad potencial de CNVs raras o únicas (Sebat y cols., 316:445 - 449 [2007]; Walsh y cols., Science 320:539 - 543 [2008]). Las CNV surgen de reordenaciones genómicas, principalmente debidas a episodios de eliminación, duplicación, inserción y translocación desequilibrada.

Las realizaciones proporcionan un método para evaluar la variación en el número de copias de una secuencia de interés, p. ej. una secuencia clínicamente importante, en una muestra de prueba que comprende una mezcla de ácidos nucleicos derivados de dos genomas diferentes, y que se sabe o se sospecha que 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 realización, la mezcla de ácido nucleicos se deriva de células normales y cancerosas derivadas de un sujeto que sufre una afección médica, p. ej. cáncer. El método de la invención se puede usar para ayudar en la identificación de estas y otras CNVs prenatalmente.

Se cree que muchos tumores sólidos, tales como el cáncer de mama, progresan desde el inicio hasta la metástasis a través de la acumulación de varias aberraciones genéticas. [Sato y cols., Cancer Res., 50: 7184-7189 [1990]; Jongsmay cols., J ClinPathol: Mol Path 55:305-309 [2002])]. Estas aberraciones genéticas, a medida que se acumulan, pueden conferir ventajas proliferativas, inestabilidad genética y la capacidad concomitante de provocar resistencia a fármacos rápidamente, y potenciar la angiogénesis, la proteólisis y la metástasis. Las aberraciones genéticas pueden afectar bien a "genes supresores de tumores" recesivos o bien a oncogenes de acción dominante. Se cree que las eliminaciones y la recombinación que conducen a pérdida de heterocigosidad (LOH) representan un papel importante en la progresión de tumores al destapar alelos supresores de tumores mutados.

Se ha encontrado ADNcf en la circulación de pacientes diagnosticados de enfermedades malignas incluyendo pero no limitadas a cáncer de pulmón (Pathaky cols. ClinChem 52:1833-1842 [2006]), cáncer de próstata (Schwartzenbach y cols. Clin Cancer Res 15:1032-8 [2009]) y cáncer de mama (Schwartzenbach y cols. disponible en Internet en breast-cancer-research.com/content/11/5/R71 [2009]). La identificación de incapacidades genómicas asociadas con cánceres

que se pueden determinar en el ADNcf circulante en pacientes con cáncer es una herramienta potencial de diagnóstico y pronóstico. En una realización, el método evalúa CNV de una secuencia de interés en una muestra que comprende una mezcla de ácidos nucleicos derivados de un sujeto que se sospecha o se sabe que tiene cáncer, p. ej. carcinoma, sarcoma, linfoma, leucemia, tumores de células germinales y blastoma. En una realización, la muestra es una muestra de plasma derivada (procesos) de sangre periférica y que comprende una mezcla de ADNcf derivados de células normales y cancerosas. En otra realización, la muestra biológica en la que se necesita determinar si está presente una CNV se deriva de una mezcla de células cancerosas y no cancerosas procedentes de otros fluidos biológicos incluyendo pero no limitados a suero, sudor, lágrimas, esputos, orina, esputos, flujo ótico, linfa, saliva, líquido cefalorraquídeo, ravages, suspensión de médula ósea, flujo vaginal, lavado transcervical, líquido cerebral, ascitis, leche, secreciones de los tractos respiratorio, intestinal y genitourinario y muestras de leucoforesis, o en biopsias tisulares, hisopos o lágrimas.

La secuencia de interés es una secuencia de ácido nucleico que se sabe o se sospecha que representa un papel en el desarrollo y/o la progresión del cáncer. Ejemplos de una secuencia de interés incluyen secuencias de ácidos nucleicos que se amplifican o se eliminan en células cancerosas según se describe en lo siguiente.

Los genes de acción dominante asociados con tumores sólidos humanos ejercen típicamente su efecto mediante la sobreexpresión o la expresión alterada. La amplificación génica es un mecanismo común que conduce a la regulación al alza de la expresión génica. La evidencia procedente de estudios citogenéticos indica que se produce una amplificación significativa en más de 50% de los cánceres de mama humanos. Lo más notablemente, la amplificación del protooncogén receptor de factor de crecimiento epidérmico humano 2 (HER2) situado en el cromosoma 17 (17(17q21-q22)), da como resultado la sobreexpresión de receptores HER2 sobre la superficie celular que conduce a una señalización excesiva y desregulada en cáncer de mama y otras enfermedades malignas (Park y cols., Clinical Breast Cancer 8:392-401 [2008]). Se ha encontrado que una variedad de oncogenes son amplificados en otras enfermedades malignas humanas. Ejemplos de la amplificación de oncogenes celulares en tumores humanos incluyen amplificaciones de: c-myc en la línea celular de leucemia promielocítica HL60, y en líneas celulares de carcinoma de pulmón microcítico, N-myc en neuroblastomas primarios (estadios III y IV), líneas celulares de neuroblastoma, línea celular y tumores primarios de retinoblastoma, y líneas y tumores de carcinoma pulmonar microcítico, L-myc en líneas celulares y tumores de carcinoma pulmonar microcítico, c-myb en leucemia mieloide aguda y en líneas celulares de carcinoma de colon, c-erbb en células de carcinoma epidermoide, y gliomas primarios, c-K-ras-2 en carcinomas primarios de pulmón, colon, vejiga urinaria y recto, N-ras en línea celular de carcinoma mamario (Varmus H., Ann Rev Genetics 18: 553-612 (1984) [citado en Watson y cols., Molecular Biology of the Gene (4ª ed.; Benjamin/Cummings Publishing Co. 1987)].

Las eliminaciones cromosómicas que implican genes supresores de tumores pueden representar un papel importante en el desarrollo y la progresión de tumores sólidos. El gen supresor de tumor de retinoblastoma (Rb-1), situado en el cromosoma 13q14, es el gen supresor de tumores más ampliamente caracterizado. El producto génico Rb-1, una fosfoproteína nuclear de 105 kDa, aparentemente representa un papel importante en la regulación del ciclo celular (Howe y cols., ProcNatlAcadSci (USA) 87:5883-5887 [1990]). La expresión alterada o perdida de la proteína Rb está provocada por la inactivación de ambos alelos génicos bien a través de una mutación puntual o bien de una eliminación cromosómica. Se ha encontrado que la alteración del gen Rb-i está presente no solo en retinoblastomas sino también en otras enfermedades malignas tales como osteosarcomas, cáncer de pulmón microcítico (Rygaard y cols., Cancer Res 50: 5312-5317 [1990)]) y cáncer de mama. Estudios de polimorfismos en la longitud de fragmentos de restricción (RFLP) han indicado que estos tipos de tumores frecuentemente han perdido heterocigosidad en 13q, sugiriendo que uno de los alelos del gen Rb-1 se ha perdido debido a una eliminación cromosómica en bruto (Bowcock y cols., Am J Hum Genet, 46: 12 [1990]). Las anormalidades del cromosoma 1 incluyendo duplicaciones, eliminaciones y translocaciones desequilibradas que implican al cromosoma 6 y otros cromosomas asociados indican que regiones del cromosoma 1, en particular 1q21-1q32 y 1p11-13, podrían alojar oncogenes o genes supresores de tumores que son patogenéticamente importantes para las fases crónica y avanzada de neoplasmas mieloproliferativos (Caramazza y cols., Eur J Hematol84:191-200 [2010]). Los neoplasmas mieloproliferativos también están asociados con eliminaciones del cromosoma 5. La pérdida completa o las eliminaciones intersticiales del cromosoma 5 son la anormalidad cariotípica más común en síndromes mielodisplásticos (MDSs). Los pacientes con MDS del(5q)/5q- aislados tienen un pronóstico más favorable que aquellos con defectos cariotípicos adicionales, que tienden a desarrollar neoplasmas mieloproliferativos (MPNs) y leucemia mieloide aguda. La frecuencia de eliminaciones del cromosoma 5 desequilibradas ha conducido a la idea de que 5q alberga uno o más genes supresores de tumores que tienen papeles fundamentales en el control del crecimiento de células madre/progenitoras hematopoyéticas (HSCs/HPCs). La asignación citogenética de regiones comúnmente eliminadas (CDRs) centradas en 5q31 y 5q32 identificaba posibles genes supresores de tumores, incluyendo la subunidad ribosómica RPS14, el factor de transcripción Egr1/Krox20 y la proteína de remodelación citoesquelética, alfa-catenina (Eisenmanny cols., Oncogene 28:3429-3441 [2009]). Estudios citogenéticos y de alotipado de tumores y líneas celulares tumorales recientes han mostrado que la pérdida alélica desde varias regiones distintas en el cromosoma 3p, incluyendo 3p25, 3p21-22, 3p21.3, 3p12-13 y 3p14, son las anormalidades genómicas más tempranas y más frecuentes implicadas en un amplio espectro de cánceres epiteliales importantes de pulmón, mama, riñón, cabeza y cuello, ovario, cuello uterino, colon, páncreas, esófago, vejiga urinaria y otros órganos. Varios genes supresores de tumores se han asignado a la región del cromosoma 3p, y se cree que eliminaciones intersticiales o hipermetilación del promotor preceden a la pérdida del 3p o todo el cromosoma 3 en el desarrollo de carcinomas (Angeloni D., Briefings Functional Genomics 6:19-39 [2007]).

Los recién nacidos y los niños con síndrome de Down (DS) presentan a menudo leucemia transitoria congénita y tienen un incremento en el riesgo de leucemia mieloide aguda y leucemia linfoblástica aguda. El cromosoma 21, que alberga aproximadamente 300 genes, puede estar implicado en numerosas aberraciones estructurales, p. ej., translocaciones, eliminaciones y amplificaciones, en leucemias, linfomas y tumores sólidos. Por otra parte, se ha identificado que los genes situados en el cromosoma 21 representan un papel importante en la tumorigénesis. Las aberraciones numéricas somáticas así como estructurales en el cromosoma 21 están asociadas con leucemias, y genes específicos incluyendo RUNX1, TMPRSS2 y TFF, que están situados en 21q, representan un papel en la tumorigénesis (Fonatsch C Gene Chromosomes Cancer 49:497-508 [2010]).

La divulgación proporciona un medio para evaluar la asociación entre la amplificación génica y el grado de evolución del tumor. La correlación entre la amplificación y/o la eliminación y el estadio o el grado de un cáncer puede ser importante para el pronóstico debido a que esta información puede contribuir a la definición de un grado tumoral basado genéticamente que predeciría mejor el curso fututo de la enfermedad con tumores más avanzados que tengan el peor pronóstico. Además, la información acerca de episodios de amplificación y/o eliminación tempranos puede ser útil al asociar esos episodios como predictores de una progresión posterior de la enfermedad. La amplificación y las eliminaciones génicas que se identifican mediante el método se pueden asociar con otros parámetros conocidos tales como el grado del tumor, la histología, el índice de marcaje de Brd/Urd, el estado hormonal, la implicación nodular, el tamaño del tumor, la duración de supervivencia y otras propiedades del tumor disponibles de estudios epidemiológicos y bioestadísticos. Por ejemplo, el ADN tumoral a probar mediante el método podría incluir hiperplasia atípica, carcinoma ductal in situ, cáncer en estadio I-III y nódulos linfáticos metastásicos a fin de permitir la identificación de asociaciones entre las amplificaciones y las eliminaciones y el estadio. Las asociaciones realizadas pueden hacer posible una intervención terapéutica eficaz. Por ejemplo, las regiones amplificadas consecuentemente pueden contener un gen sobreexpresado, cuyo producto puede ser atacado terapéuticamente (por ejemplo, la tirosina cinasa receptora de factor de crecimiento, $p185^{HER2}$).

El método se puede usar para identificar episodios de amplificación y/o eliminación que están asociados con resistencia a fármacos al determinar la variación del número de copias de ácidos nucleicos procedentes de cánceres primarios con los de células que se han metastasizado a otras zonas. Si la amplificación y/o eliminación génica es una manifestación de inestabilidad cariotípica que permite un desarrollo rápido de resistencia a fármacos, se esperaría más amplificación y/o eliminación en tumores primarios procedentes de pacientes quimiorresistentes que en tumores de pacientes quimiosensibles. Por ejemplo, si la amplificación de genes específicos es responsable del desarrollo de resistencia a fármaco, se esperaría que las regiones que rodean esos genes se amplificaran consecuentemente en células tumorales procedentes de efusiones pleurales de pacientes quimiorresistentes pero no en los tumores primarios. El descubrimiento entre la amplificación y/o eliminación génica y el desarrollo de resistencia a fármacos puede permitir la identificación de pacientes que se beneficiaran o no de una terapia adyuvante.

En otras realizaciones, el presente método se puede usar para identificar polimorfismos asociados con trastornos de repetición trinucleotídica, que son un conjunto de trastornos genéticos provocados por expansión de repeticiones trinucleotídicas. Las expansiones trinucleotídicas son un subconjunto de repeticiones microsatelitales inestables que se presentan a través de todas las secuencias genómicas. Si la repetición está presente en un gen sano, una mutación dinámica puede incrementar el recuento de repeticiones y dar como resultado un gen defectuoso. En una realización, el método se puede usar para identificar repeticiones trinucleotídicas asociadas con el síndrome del cromosoma X frágil. El brazo largo del cromosoma X de pacientes que sufren síndrome del cromosoma X frágil puede contener de 230 a 4000 CGG, en comparación con de 60 a 230 repeticiones en portadores y de 5 a 54 repeticiones en individuos no afectados. La inestabilidad cromosómica resultante de esta expansión trinucleotídica se presenta clínicamente como retardo mental, rasgos faciales distintivos y macroorquidismo en varones. La segunda enfermedad por tripletes de ADN relacionada, el síndrome de los cromosomas X-E frágiles, también se identificaba en el cromosoma X, pero se encontró que era el resultado de una repetición de CCG expandida. El presente método puede identificar repeticiones trinucleotídicas asociadas con otros trastornos de expansión de repeticiones incluyendo las Categorías I, II y III. Los trastornos de la Categoría I incluyen enfermedad de Huntington (HD) y las ataxias espinocerebelares que son provocadas por una expansión de la repetición CAG en porciones codificantes de proteínas de genes específicos. Las expansiones de la Categoría II tienden a ser más fenotípicamente diversas con expansiones heterogéneas que generalmente son de pequeña magnitud, pero también se encuentran en los exones de genes. La Categoría III incluye síndrome del cromosoma X frágil, distrofia miotónica, dos de las ataxias espinocerebelares, epilepsia mioclónica juvenil y ataxia de Friereich. Estas enfermedades se caracterizan típicamente por expansiones de la repetición mucho mayores que los dos primeros grupos, y las repeticiones están situadas fuera de las regiones codificantes de proteínas de los genes.

En otras realizaciones, el presente método puede identificar repeticiones trinucleotídicas de CAG asociadas con al menos diez trastornos neurológicos que se sabe que están provocados por un incremento en el número de repeticiones de CAG, típicamente en regiones codificantes de proteínas por lo demás no relacionadas. Durante la síntesis de proteínas, las repeticiones de CAG expandidas se traducen en una serie de residuos de glutamina ininterrumpidos que forman lo que se conoce como un tramo de poliglutamina ("polyQ"). Estos tramos de poliglutamina pueden estar sometidos a un incremento de la agregación. Estos trastornos se caracterizan por un modo dominante autosómico de herencia (con la excepción de la atrofia muscular espinobulbar que muestra herencia ligada al cromosoma X), un

comienzo a mitad de vida, un curso progresivo y una correlación del número de repeticiones de CAG con la gravedad de la enfermedad y la edad del comienzo. Los genes causales se expresan ampliamente en todas las enfermedades poliglutamínicas conocidas. Un síntoma común de enfermedades PolyQ se caracteriza por una degeneración progresiva de células nerviosas que habitualmente afecta a personas de cierta edad. Aunque estas enfermedades comparten el mismo codón (CAG) repetido y algunos síntomas, las repeticiones para las diferentes enfermedades poliglutamínicas se presentan en diferentes cromosomas. Ejemplos de trastornos polyQ que se pueden identificar mediante el presente método incluyen sin limitación DRPLA (atrofia dentatorubropalidoluisiana), HD (enfermedad de Huntington), SBMA (atrofia muscular espinobulbar o enfermedad de Kennedy), SCA1 (ataxia espinocerebelar tipo 1), SCA2 (ataxia espinocerebelar tipo 2), SCA3 (ataxia espinocerebelar tipo 3 o enfermedad de Machado-Joseph), SCA6 (ataxia espinocerebelar tipo 6), SCA7 (ataxia espinocerebelar tipo 7), SCA17 (ataxia espinocerebelar tipo 17). Ejemplos de trastornos no polyQ que se pueden identificar mediante el presente método incluyen FRAXA (síndrome del cromosoma X frágil), FXTAS (síndrome de temblores/ataxia asociado al cromosoma X frágil), FRAXE (retardo mental por cromosomas XE frágiles), FRDA (ataxia de Friedreich), DM (distrofia miotónica), SCA8 (ataxia espinocerebelar tipo 8), SCA12 (ataxia espinocerebelar tipo 12).

Además del papel de las CNV en el cáncer, las CNVs se han asociado con un número creciente de enfermedades complejas comunes, incluyendo virus de inmunodeficiencia humana (HIV), enfermedades autoinmunitarias y un espectro de trastornos neuropsiquiátricos.

Hasta la fecha un número de estudios ha presentado una asociación entre CNV en genes implicados en la inflamación y la respuesta inmunitaria y VIH, asma, enfermedad de Crohn y otros trastornos autoinmunitarios (Fanciulli y cols., Clin Genet 77:201-213 [2010]). Por ejemplo, la CNV en *CCL3L1* se ha relacionado con la sensibilidad al VIH/SIDA (*CCL3L1,* eliminación 17q11.2), la artritis reumatoide (*CCL3L1,* eliminación 17q11.2) y la enfermedad de Kawasaki (*CCL3L1,* duplicación 17q11.2); se ha presentado que la CNV en *HBD-2* predispone a enfermedad de Crohn colónica (*HDB-2*, eliminación 8p23.1) y la psoriasis (*HDB-2,* eliminación 8p23.1); se observó que la *CNV* en *FCGR3B* predispone a glomerulonefritis en el lupus eritematoso sistémico (*FCGR3B,* eliminación 1q23, duplicación 1q23), y la vasculitis asociada a anticuerpos citoplásmicos antineutrofílicos (ANCA) (*FCGR3B,* eliminación 1q23), y el incremento en el riesgo de desarrollar artritis reumatoide. Existen al menos dos enfermedades inflamatorias o autoinmunitarias que se ha mostrado que están asociadas con CNV en diferentes locus génicos. Por ejemplo, la enfermedad de Crohn se asocia con un bajo número de copias en *HDB-2,* pero también con un polimorfismo de eliminación común aguas arriba del gen *IGRM* que codifica un miembro de la familia de GTPasas relacionadas con la inmunidad p47. Además de la asociación con el número de copias de *FCGR3B* , también se ha presentado que la sensibilidad a SLE se incrementa significativamente entre sujetos con un número de copias inferior del componente del complemento C4.

Se han presentado asociaciones entre eliminaciones genómicas en los locus *GSTM1* (*GSTM1,* eliminación 1q23) y *GSTT1* (*GSTT1,* eliminación 22q11.2) y un incremento en el riesgo de asma atópica en un número de estudios independientes. En algunas realizaciones, el presente método se puede usar para determinar la presencia o ausencia de una CNV asociada con enfermedades inflamatorias y/o autoinmunitarias. Por ejemplo, el presente método se puede usar para determinar la presencia de una CNV en un paciente que se sospecha que está sufriendo VIH, asma o enfermedad de Crohn. Ejemplos de CNV asociada con estas enfermedades incluyen sin limitación eliminaciones en 17q11.2, 8p23.1, 1q23 y 22q11.2 y duplicaciones en 17q11.2 y 1q23. En algunas realizaciones, el presente método se puede usar para determinar la presencia de CNV en genes incluyendo pero no limitados a *CCL3L1, HBD-2, FCGR3B, GSTM, GSTT1, C4* y *IRGM.*

Se han presentado 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 enfermedad de Parkinson, esclerosis lateral amiotrófica (ALS) y enfermedad de Alzheimer dominante autosómica (Fanciulli y cols., Clin Genet 77:201-213 [2010]). Se han observado anormalidades citogenéticas en pacientes con autismo y trastornos del espectro autista (ASDs) con duplicaciones en 15q11-q13. Según the Autism Genome project Consortium, 154 CNV incluyendo varias CNVs recurrentes, bien en el cromosoma 15q11-q13 o bien en nuevas localizaciones genómicas incluyendo el cromosoma 2p16, 1q21 y en 17p12 en una región asociada con el síndrome de Smith-Magenis que se solapa con ASD. Microeliminaciones o microduplicaciones recurrentes en el cromosoma 16p11.2 han destacado la observación de que se detectan CNVs *de novo* en locus para genes tales como *SHANK3* (eliminación de 22q13.3), neurexina 1 (*NRXN1*, eliminación de 2p16.3) y las neuroglinas (*NLGN4,* eliminación de Xp22.33) que se sabe que regulan la diferenciación sináptica y regulan la liberación de neurotransmisores glutaminérgicos. La esquizofrenia se ha relacionado con múltiples CNVs *de novo*. Las microeliminaciones y las microduplicaciones asociadas con la esquizofrenia contienen una sobrerrepresentación de genes pertenecientes a las rutas de neurodesarrollo y glutaminérgica, sugiriendo que múltiples CNVs que afectan a estos genes pueden contribuir directamente a la patogénesis de la esquizofrenia, p. ej. *ERBB4,* eliminación 2q34, *SLC1A3,* eliminación 5p13.3; *RAPEGF4,* eliminación 2q31.1; *CIT,* eliminación 12.24; y múltiples genes con CNV *de novo* . Las CNVs también se han asociado con otros trastornos neurológicos incluyendo epilepsia (*CHRNA7,* eliminación 15q13.3), enfermedad de Parkinson (*SNCA* duplicación 4q22) y ALS (*SMN1,* eliminación 5q12.2.-q13.3; y eliminación *SMN2*). En algunas realizaciones, el presente método se puede usar para determinar la presencia o ausencia de una CNV asociada con enfermedades del sistema nervioso. Por ejemplo, el presente método se puede usar para determinar la presencia de una CNV en un paciente que se sospecha que está sufriendo autismo, esquizofrenia, epilepsia, enfermedades neurodegenerativas tales como enfermedad de Parkinson, esclerosis lateral

amiotrófica (ALS) o enfermedad de Alzheimer dominante autosómica. El presente método se puede usar para determinar CNV de genes asociados con enfermedades del sistema nervioso incluyendo sin limitación cualquiera de los trastornos del espectro autista (ASD), esquizofrenia y epilepsia, y CNV de genes asociados con trastornos neurodegenerativos tales como enfermedad de Parkinson. Ejemplos de CNV asociadas con estas enfermedades incluyen sin limitación duplicaciones en 15q11-q13, 2p16, 1q21, 17p12, 16p11.2 y 4q22, y eliminaciones en 22q13.3, 2p16.3, Xp22.33, 2q34, 5p13.3, 2q31.1, 12.24, 15q13.3 y 5q12.2. En algunas realizaciones, el presente método se puede usar para determinar la presencia de CNV en genes incluyendo pero no limitados a *SHANK3*, *NLGN4, NRXN1, ERBB4, SLC1A3, RAPGEF4, CIT, CHRNA7, SNCA, SMN1* y SMN2.

La asociación entre rasgos metabólicos y cardiovasculares, tales como hipercolesterolemia familiar (FH), aterosclerosis y arteriopatía coronaria, y se han presentado CNVs en un número de estudios (Fanciulli y cols., Clin Genet 77:201-213 [2010]). Por ejemplo, se han observado reordenaciones, principalmente eliminaciones, de la línea germinal en el gen *LDLR* (*LDLR,* eliminación/duplicación 19p13.2) en algunos pacientes con FH que no portan otras mutaciones de *LDLR*. Otro ejemplo es el gen *LPA* que codifica apolipoproteína(a) (apo(a)) cuya concentración en plasma está asociada con riesgo de arteriopatía coronaria, infarto de miocardio (MI) y apoplejía. Las concentraciones en plasma de la lipoproteína Lp(a) que contiene apo(a) varían por encima de 1000 veces entre individuos y 90% de esta variabilidad está determinada genéticamente en el locus *LPA*, siendo la concentración en plasma y el tamaño de la isoforma Lp(a) proporcionales a un número altamente variable de secuencias de repetición 'Kringle 4' (intervalo 5-50). Estos datos indican que CNV en al menos dos genes se pueden asociar con riesgo cardiovascular. El presente método se puede usar en grandes estudios para buscar específicamente asociaciones de CNV con trastornos cardiovasculares. En algunas realizaciones, el presente método se puede usar para determinar la presencia o ausencia de una CNV asociada con una enfermedad metabólica o cardiovascular. Por ejemplo, el presente método se puede usar para determinar la presencia de una CNV en un paciente que se sospecha que está sufriendo hipercolesterolemia familiar. El presente método se puede usar para determinar CNV de genes asociados con una enfermedad metabólica o cardiovascular, p. ej. hipercolesterolemia. Ejemplos de CNV asociadas con estas enfermedades incluyen sin limitación eliminación/duplicación 19p13.2 del gen LDLR, y multiplicaciones en el gen LPA.

SECUENCIACIÓN

En diversas realizaciones, el método descrito en la presente emplea tecnología de secuenciación de última generación (NGS) en la que plantillas de ADN o moléculas de ADN individuales amplificadas clonalmente se secuencian de un modo masivamente paralelo dentro de una celdilla de flujo (p. ej., según se describe en Volkerding y cols. ClinChem 55:641-658 [2009]; Metzker M Nature Rev 11:31-46 [2010]). Además de la información de secuencias de alto rendimiento, la NGS proporciona información cuantitativa digital, ya que cada lectura de secuencia es un “marcador de secuencia” contable que representa una plantilla de ADN clonal individual o una sola molécula de ADN. Las tecnologías de secuenciación de NGS incluyen pirosecuenciación, secuenciación mediante síntesis con terminadores de colorantes reversibles, secuenciación mediante ligación de sondas oligonucleotídicas y secuenciación en tiempo real.

En diversas realizaciones, se pueden analizar muestras que no están amplificadas, o solo se amplifican parcialmente (amplificación dirigida). En algunos casos, los métodos para determinar la fracción fetal se pueden efectuar sin requerir ningún tipo de amplificación dirigida.

La amplificación del genoma completo que se produce como parte del procedimiento de secuenciación proporciona suficientes copias que pueden ser cubiertas por un número creciente de ciclos de secuenciación para proporcionar una cobertura cada vez mejor.

En realizaciones preferidas, la muestra que comprende la mezcla de moléculas de ADN derivadas de dos genomas diferentes se enriquece inespecíficamente con respecto a secuencias del genoma completo antes de la secuenciación del genoma completo, es decir la amplificación del genoma completo se realiza antes de la secuenciación.

El enriquecimiento inespecífico de un ADN de muestra se puede referir a la amplificación del genoma completo de los fragmentos de ADN genómico de la muestra que se pueden usar para incrementar el nivel de ADN de muestra antes de identificar polimorfismos mediante secuenciación. El enriquecimiento inespecífico puede ser el enriquecimiento selectivo de uno de los dos genomas presentes en la muestra. Por ejemplo, el enriquecimiento inespecífico puede ser selectivo del genoma fetal en una muestra materna, que se puede obtener mediante métodos conocidos para incrementar la proporción relativa de ADN fetal a materno en una muestra. Alternativamente, el enriquecimiento inespecífico puede ser la amplificación no selectiva de ambos genomas presentes en la muestra. Por ejemplo, la amplificación inespecífica puede ser de ADN fetal y materno en una muestra que comprende una mezcla de ADN procedente de los genomas fetal y materno. Métodos para la amplificación del genoma completo se conocen en la técnica. La PCR cebada con oligonucleótidos degenerados (DOP), la técnica de PCR con extensión con cebador (PEP) y la amplificación con múltiples desplazamientos (MDA) son ejemplos de métodos de amplificación del genoma completo. En algunas realizaciones, la muestra que contiene la mezcla de ADNcf procedente de diferentes genomas no está enriquecida con respecto al ADNcf de los genomas presentes en la mezcla. En otras realizaciones, la muestra

que comprende la mezcla de ADNcf procedente de diferentes genomas está enriquecida inespecíficamente con respecto a uno cualquiera de los genomas presentes en la muestra.

En otras realizaciones, el ADNcf en la muestra está enriquecido específicamente. El enriquecimiento específico se refiere al enriquecimiento de una muestra genómica con respecto a secuencias específicas, p. ej. una secuencia diana polimórfica, que se seleccionan para la amplificación antes de secuenciar la muestra de ADN. Sin embargo, una ventaja de las realizaciones divulgadas es que no es necesaria la amplificación dirigida. Polimórfica

Algunas de las tecnologías de secuenciación están disponibles comercialmente, 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), que se describen posteriormente. Además de la secuenciación de moléculas individuales realizada usando la secuenciación por síntesis de Helicos Biosciences, otras tecnologías de secuenciación de moléculas individuales son abarcadas por el método divulgado e incluyen la tecnología SMRT™ de Pacific Biosciences, la tecnología Ion Torrent™ y la secuenciación en nanoporos que es desarrollada, por ejemplo, por Oxford Nanopore Technologies.

Aunque el método de Sanger automatizado se considera una tecnología de 'primera generación', la secuenciación de Sanger incluyendo la secuenciación de Sanger automatizada también puede ser empleada por el método divulgado. Métodos de secuenciación adicionales que comprenden el uso del desarrollo de tecnologías de obtención de imágenes de ácidos nucleicos, p. ej. microscopía de fuerza atómica (AFM) o microscopía electrónica de transmisión (TEM), también son abarcados por el método divulgado. Tecnologías de secuenciación ejemplares se describen posteriormente.

En una realización, la tecnología de secuenciación de ADN que se usa en los métodos divulgados es la secuenciación cierta de moléculas individuales (tSMS) de Helicos (p. ej., según se describe en Harris T.D. y cols., Science 320:106-109 [2008]). En la técnica de tSMS, una muestra de ADN se escinde en dos cadenas de aproximadamente 100 a 200 nucleótidos, y una secuencia de poliA se añade al extremo 3' de cada cadena de ADN. Cada cadena de ADN se marca mediante la adición de un nucleótido de adenosina marcado fluorescentemente. A continuación, las cadenas de ADN se hibridan a una celdilla de flujo, que contiene millones de sitios de captura de oligo-T que están inmovilizados en la superficie de la celdilla de flujo. Las plantillas pueden estar en una densidad de aproximadamente 100 millones de plantillas/cm^2. A continuación, la celdilla de flujo se carga en un instrumento, p. ej., un secuenciador HeliScope™, y un láser ilumina la superficie de la celdilla de flujo, revelando la posición de cada plantilla. Una cámara CCD puede asignar la posición de las plantillas sobre la superficie de la celdilla de flujo. A continuación, el marcador fluorescente de las plantillas se escinde y se elimina por lavado. La reacción de secuenciación comienza al introducir una ADN polimerasa y un nucleótido marcado fluorescentemente. El ácido nucleico oligo-T sirve como un cebador. La polimerasa incorpora los nucleótidos marcados al cebador de un modo dirigido por la plantilla. La polimerasa y los nucleótidos no incorporados se retiran. Las plantillas que tienen una incorporación dirigida del nucleótido marcado fluorescentemente se disciernen al obtener imágenes de la superficie de la celdilla de flujo. Después de la obtención de imágenes, una etapa de escisión retira el marcador fluorescente, y el procedimiento se repite con otros nucleótidos marcados fluorescentemente hasta que se alcanza una longitud de lectura deseada. La información de la secuencia se recoge con cada etapa de adición de nucleótido. La secuenciación del genoma completo mediante tecnologías de secuenciación de moléculas individuales excluye la amplificación basada en PCR en la preparación de bibliotecas de secuenciación, y lo directo de la preparación de muestras permite la medida directa de la muestra, en lugar de la medida de copias de esa muestra.

En una realización, la tecnología de secuenciación de ADN que se usa en los métodos divulgados es la secuenciación 454 (Roche) (p. ej. según se describe en Margulies, M. y cols. 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 los fragmentos se enroman. A continuación se ligan adaptadores oligonucleotídicos a los extremos de los fragmentos. Los adaptadores sirven como cebadores para la amplificación y la secuenciación de los fragmentos. Los fragmentos se pueden unir a microesferas de captura de ADN, p. ej., microesferas revestidas con estreptavidina que usan, p. ej., Adaptador B, que contiene marcador de biotina 5'. Los fragmentos unidos a las microesferas se amplifican por PCR dentro de gotículas de una emulsión de aceite-agua. El resultado es múltiples copias de fragmentos de ADN amplificados clonalmente sobre cada microesfera. En la segunda etapa, las microesferas se capturan en pocillos (tamaño de picolitros). La pirosecuenciación se realiza sobre cada fragmento de ADN en paralelo. La adición de uno o más nucleótidos genera una señal luminosa que se registra mediante una cámara CCD en un instrumento de secuenciación. La intensidad de la señal es proporcional al número de nucleótidos incorporados. La pirosecuenciación hace uso de pirofosfato (PPi) que se libera tras la adición del nucleótido. El PPi se convierte en ATP mediante ATP sulfurilasa en presencia de 5' fosfosultato de adenosina. La luciferasa usa ATP para convertir luciferina en oxiluciferina, y esta reacción genera luz que se discierne y analiza.

En una realización, la tecnología de secuenciación de ADN que se usa en los métodos divulgados es la tecnología SOLiD™ (Applied Biosystems). En la secuenciación por ligación SOLiD™, ADN genómico se corta en fragmentos, y se unen adaptadores a los extremos 5' y 3' de los fragmentos para generar una biblioteca de fragmentos. Alternativamente, se pueden introducir adaptadores internos al ligar adaptadores a los extremos 5' y 3' de los

fragmentos, circularizar los fragmentos, digerir el fragmento circularizado para generar un adaptador interno y unir adaptadores a los extremos 5' y 3' de los fragmentos resultantes para generar una biblioteca emparejada. Posteriormente, se preparan poblaciones de microesferas clonales en microrreactores que contienen microesferas, cebadores, plantilla y componentes de PCR. Después de la PCR, las plantillas se desnaturalizan y las microesferas se enriquecen para separar las microesferas con plantillas extendidas. Las plantillas sobre las microesferas seleccionadas se someten a una modificación en 3' que permite la unión a un portaobjetos de vidrio. La secuencia se puede determinar mediante hibridación y ligación secuenciales de oligonucleótidos parcialmente aleatorios con una base (o par de bases) determinada central que se identifica mediante un fluoróforo específico. Después de que se registre un color, el oligonucleótido ligado se escinde y se retira y el procedimiento se repite a continuación.

En una realización, la tecnología de secuenciación de ADN que se usa en los métodos divulgados es la tecnología de secuenciación en tiempo real de moléculas individuales (SMRT™) de Pacific Biosciences. En la secuenciación SMRT, se obtienen imágenes de la incorporación continua de nucleótidos marcados con colorante durante la síntesis de ADN. Moléculas de ADN polimerasa individuales se unen a la superficie inferior de identificadores de longitud de onda en modo cero (identificadores ZMW) individuales que obtienen información de secuencia mientras se están incorporando nucleótidos fosfoligados a la cadena cebadora en crecimiento. Un ZMW es una estructura de confinamiento que permite la observación de la incorporación de un solo nucleótido mediante ADN polimerasa frente al fondo de nucleótidos fluorescentes que se difunden rápidamente en una salida del ZMW (en microsegundos). Lleva varios milisegundos incorporar un nucleótido en una cadena en crecimiento. Durante este tiempo, el marcador fluorescente se excita y produce una señal fluorescente, y la marca fluorescente se escinde. La identificación de la fluorescencia correspondiente del colorante indica qué base se incorporaba. El procedimiento se repite.

En una realización, la tecnología de secuenciación de ADN que se usa en los métodos divulgados es la secuenciación en nanoporos (p. ej. según se describe en Soni GV and Meller A. ClinChem 53: 1996-2001 [2007]). Las técnicas de análisis de ADN por secuenciación en nanoporos están siendo desarrolladas industrialmente por un número de compañías, incluyendo Oxford Nanopore Technologies (Oxford, Reino Unido). La secuenciación en nanoporos es una tecnología de secuenciación de moléculas individuales por la que una sola molécula de ADN se secuencia directamente a medida que pasa a través de un nanoporo. Un nanoporo es un orificio pequeño, del orden de 1 nanómetro de diámetro. La inmersión de un nanoporo en un fluido conductor y la aplicación de un potencial (voltaje) a su través da como resultado una ligera corriente eléctrica debido 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 a través de un nanoporo, cada nucleótido sobre la molécula de ADN obstruye el nanoporo hasta un grado diferente, cambiando la magnitud de la corriente a través del nanoporo en diferentes grados. Así, este cambio en la corriente a medida que la molécula de ADN pasa a través del nanoporo representa una lectura de la secuencia de ADN.

En una realización, la tecnología de secuenciación de ADN que se usa en los métodos divulgados es la serie de transistores con efecto de campo sensibles a productos químicos (chemFET) (p. ej., según 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, moléculas de ADN se pueden introducir en cámaras de reacción y las moléculas de plantilla se pueden hibridar a un cebador de secuenciación unido a una polimerasa. La incorporación de uno o más trifosfatos a una nueva cadena de ácido nucleico en el extremo 3' del cebador de secuenciación se puede discernir mediante un cambio en la corriente mediante un chemFET. Una serie puede tener múltiples sensores de chemFET. En otro ejemplo, ácidos nucleicos individuales se pueden conectar a microesferas, y los ácidos nucleicos se pueden amplificar sobre la microesfera, y las microesferas individuales se pueden transferir a cámaras de reacción individuales en una serie de chemFET, teniendo cada cámara un sensor chemFET, y los ácidos nucleicos se pueden secuenciar.

En una realización, la tecnología de secuenciación de ADN que se usa en los métodos divulgados es el método de Halcyon Molecular que usa microscopía electrónica de transmisión (TEM). El método, denominado nanotransferencia rápida con posicionamiento de moléculas individuales (IMPRNT), comprende utilizar obtención de imágenes por microscopía de transmisión de resolución de un solo átomo de ADN de alto peso molecular (150 kb o más) marcado selectivamente con marcadores de átomos pesados y disponer estas moléculas sobre películas ultradelgadas en series paralelas ultradensas (3 nm de cadena a cadena) con un espaciamiento de base a base consiguiente. El microscopio electrónico se usa para obtener imágenes de las moléculas sobre las películas para determinar la posición de los marcadores de átomos pesados y para extraer información de la secuencia de bases a partir del ADN. El método se describe adicionalmente en la publicación de patente PCT WO 2009/046445. El método permite la secuenciación de genomas humanos completos en menos de diez minutos.

En una realización, la tecnología de secuenciación de ADN es la secuenciación monomolecular de Ion Torrent, que junta tecnología de semiconductores con una química de secuenciación simple para traducir directamente información químicamente codificada (A, C, G, T) en información digital (0, 1) sobre un chip de semiconductor. En la naturaleza, cuando un nucleótido se incorpora en una cadena de ADN mediante una polimerasa, se libera un ion hidrógeno como un subproducto. Ion Torrent usa una serie de alta densidad de pocillos microlabrados para realizar este procedimiento bioquímico de un modo masivamente paralelo. Cada pocillo contiene una molécula de ADN diferente. Bajo los pocillos hay una capa sensible a iones y bajo esta un sensor iónico. Cuando un nucleótido, por ejemplo una C, se añade a una plantilla de ADN y a continuación se incorpora en una cadena de ADN, se liberará un ion hidrógeno. La carga

procedente de ese ion cambiará el pH de la solución, lo que se puede identificar mediante el sensor iónico de Ion Torrent. El secuenciador – esencialmente el pH-metro en estado sólido más pequeño del mundo – designa la base, pasando directamente de información química a información digital. A continuación, el secuenciador de Ion personal Genome Machine (PGM™) inunda secuencialmente el chip con un nucleótido tras otro. Si el siguiente nucleótido que inunda el chip no es una coincidencia. No se registrará cambio de voltaje y no designará ninguna base. Si hay dos bases idénticas en la cadena de ADN, el voltaje se doblará, y el chip registrará dos bases idénticas designadas. La identificación directa permite el registro de incorporación de nucleótidos en segundos.

En algunas realizaciones, los métodos emplean PCR o una técnica relacionada para amplificar secuencias nucleotídicas de muestra antes de identificarlas o asignarlas. Sin embargo, las técnicas algorítmicas divulgadas en la presente generalmente no requieren amplificación, particularmente amplificación dirigida de polimorfismos usada para estimar la fracción genómica.

Ciertas realizaciones emplean PCR digital y secuenciación por hibridación. La reacción en cadena de la polimerasa digital (PCR digital o dPCR) se puede usar para identificar y cuantificar directamente ácidos nucleicos en una muestra. La PCR digital se puede realizar en una emulsión. Los ácidos nucleicos individuales se separan, p. ej., en un dispositivo de cámara microfluídica, y cada ácido nucleico se amplifica individualmente mediante PCR. Los ácidos nucleicos se pueden separar de modo que haya un promedio de aproximadamente 0,5 ácidos nucleicos/pocillo, o no más de un ácido nucleico/pocillo. Se pueden usar diferentes sondas para distinguir alelos fetales y alelos maternos. Los alelos se pueden enumerar 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 donde cada una de la pluralidad de sondas polinucleotídicas se puede fijar opcionalmente a un sustrato. El sustrato podría ser una superficie plana que comprendiera una serie de secuencias nucleotídicas conocidas. El patrón de hibridación a la serie se puede usar para determinar las secuencias polinucleotídicas presentes en la muestra. En otras realizaciones, cada sonda se fija a una microesfera, p. ej., una microesfera magnética o similares. La hibridación a las microesferas se puede identificar y usar para identificar la pluralidad de secuencias polinucleotídicas dentro de la muestra.

En una realización, el método emplea secuenciación masivamente paralela de millones de fragmentos de ADN usando la secuenciación por síntesis de Illumina y la química de secuenciación basada en terminadores reversibles (p. ej. según se describe en Bentley y cols., Nature 6:53-59 [2009]). El ADN de plantilla puede ser ADN genómico, p. ej. ADNcf. En algunas realizaciones, se usa como la plantilla ADN genómico procedente de células aisladas, y se fragmenta en longitudes de varios cientos de pares de bases. En otras realizaciones, se usa ADNcf como la plantilla, y no se requiere fragmentación ya que el ADNcf existe como fragmentos cortos. Por ejemplo, el ADNcf fetal circula en la corriente sanguínea como fragmentos de <300 pb, y se ha estimado que el ADNcf materno circula como fragmentos de entre aproximadamente 0,5 y 1 Kb (Li y cols., ClinChem, 50: 1002-1011 (2004)). La tecnología de secuenciación de Illumina se basa en el empalme de ADN genómico fragmentado a una superficie plana ópticamente transparente sobre la que se unen anclajes oligonucleotídicos. Los extremos del ADN de plantilla se reparan para generar extremos romos fosforilados en 5', y la actividad de polimerasa del fragmento de Klenow se usa para añadir una sola base A al extremo 3' de los fragmentos de ADN fosforilados romos. Esta adición prepara los fragmentos de ADN para el enlace la ligación a adaptadores oligonucleotídicos, que tienen un saliente de una sola base T en su extremo 3' para incrementar la eficacia de ligación. Los oligonucleótidos adaptadores son complementarios a los anclajes de la celdilla de flujo. Bajo condiciones de dilución limitativa, ADN de plantilla monocatenario modificado con adaptador se añade a la celdilla de flujo y se inmoviliza mediante hibridación a los anclajes. Los fragmentos de ADN empalmados se extienden y se amplifican en puente para crear una celdilla de flujo de secuenciación de densidad ultraalta con cientos de millones de agregados, que contienen cada uno -1.000 copias de la misma plantilla. En una realización, el ADN genómico, p. ej. ADNcf, fragmentado aleatoriamente se amplifica usando PCR antes de que se someta a amplificación de los agregados. Alternativamente, se usa una preparación de biblioteca genómica libre de amplificación, y el ADN genómico, p. ej. ADNcf, fragmentado aleatoriamente se enriquece usando la amplificación de agregados sola (Kozarewa y cols., Nature Methods 6:291-295 [2009]). Las plantillas se secuencian usando una tecnología de secuenciación por síntesis de ADN tetracromática robusta que emplea terminadores reversibles con colorantes fluorescentes retirables. La identificación por fluorescencia de alta sensibilidad se consigue usando excitación laser y óptica de reflexión interna total. Las lecturas de secuencia corta de aproximadamente 20-40 pb, p. ej. 36 pb, se alinean frente a un genoma de referencia con repeticiones enmascaradas y las diferencias genéticas se designan usando un programa de canalización de análisis de datos desarrollado especialmente. Después de la terminación de la primera lectura, las plantillas se pueden regenerar in situ para permitir una segunda lectura desde el extremo opuesto de los fragmentos. Así, se usa según el método una secuenciación bien de un solo extremo o de extremos apareados de los fragmentos de ADN. Se realiza una secuenciación parcial de fragmentos de ADN presentes en la muestra, y se cuentan marcadores de secuencia que comprenden lecturas de longitud predeterminada, p. ej. 36 pb, que se asignan a un genoma de referencia conocido.

La longitud de la lectura de secuencia está asociada con la tecnología de secuenciación particular. Los métodos de NGS proporcionan lecturas de secuencia que varían en tamaño de decenas a cientos de pares de bases. En algunas realizaciones del método descrito en la presente, las lecturas de secuencia son de 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, 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 permitan lecturas de un solo extremo de más de 500 pb que permitan lecturas de más de aproximadamente 1000 pb cuando se generan lecturas de extremos apareados. En una realización, las lecturas de secuencia tienen 36 pb. Otros métodos se secuenciación que se pueden emplear mediante los métodos divulgados incluyen los métodos de secuenciación monomolecular que pueden secuenciar moléculas de ácidos nucleicos >5000 pb. La cantidad masiva de resultados de secuencias se transfiere mediante una canalización de análisis que transforma los resultados de obtención de imágenes primarios procedentes del secuenciador en cadenas de bases. Un paquete de algoritmos integrados realiza las etapas de transformación de datos primarios centrales: análisis de imágenes, puntuación de intensidad, designación de bases y alineamiento.

ASIGNACIÓN

Se pueden usar diversos métodos informáticos para asignar cada secuencia identificada a un “bin”, p. ej., al identificar todas las secuencias de la muestra que se asignan a un gen, un cromosoma, un alelo u otra estructura particular. Existe un número de algoritmos informáticos para alinear secuencias, incluyendo sin limitación BLAST (Altschul y cols., 1990), BLITZ (MPsrch) (Sturrock & Collins, 1993), FASTA (Person & Lipman, 1988), BOWTIE (Langmead y cols., Genome Biology 10:R25.1-R25.10 [2009]) o ELAND (Illumina, Inc., San Diego, CA, EE. UU. de A.). En algunas realizaciones, las secuencias de los compartimentos se encuentran en bases de datos de ácidos nucleicos conocidas por los expertos en la técnica, incluyendo sin limitación GenBank, dbEST, dbSTS, EMBL (the European Molecular Biology Laboratory) y el DDBJ (the DNA Databank of Japan). BLAST o herramientas similares se pueden usar para buscar las secuencias identificadas frente a la base de datos de secuencias, y los aciertos de búsqueda se pueden usar para clasificar las secuencias identificadas en compartimentos apropiados.

APARATO

Los análisis de los datos de secuenciación y los diagnósticos derivados de los mismos se realizan típicamente usando un equipo informático que funciona según algoritmos y programas definidos. Por lo tanto, ciertas realizaciones emplean procedimientos que implican datos almacenados en o transferidos a través de uno o más sistemas informáticos u otros sistemas de procesamiento. Realizaciones de la invención también se refieren a un aparato para realizar estas operaciones. Este aparato puede estar construido especialmente para los propósitos requeridos, o puede ser un ordenador (o un grupo de ordenadores) de uso general activado o reconfigurado selectivamente mediante un programa informático y/o una estructura de datos almacenada en el ordenador. En algunas realizaciones, un grupo de procesadores realiza algunas o todas las operaciones analíticas citadas colaborativamente (p. ej., a través de una computación en red o en la nube) y/o en paralelo. Un procesador o grupo de procesadores para realizar los métodos descritos en la presente pueden ser de diversos tipos incluyendo microcontroladores y microprocesadores tales como dispositivos programables (p. ej., CPLDs y FPGAs) y otros dispositivos tales como ASICs de matrices de puertas, procesadores de señales digitales y/o microprocesadores de uso general.

Además, ciertas realizaciones se refieren a medios legibles informáticamente o productos de programas informáticos tangibles y/o no transitorios que incluyen las instrucciones del programa y/o datos (incluyendo estructuras de datos) para realizar diversas operaciones ejecutadas informáticamente. Ejemplos de medios legibles informáticamente incluyen, pero no se limitan a, dispositivos de memoria semiconductores, medios magnéticos tales como unidades de disco, cinta magnética, medios ópticos tales como CDs, medios magnetoópticos y dispositivos físicos que están especialmente configurados para almacenar y ejecutar las instrucciones del programa, tales como dispositivos de memoria de solo lectura (ROM) y memoria de acceso aleatorio (RAM). Los medios legibles informáticamente pueden estar controlados directamente por un usuario final o los medios pueden estar controlados indirectamente por el usuario final. Ejemplos de medios controlados directamente incluyen los medios situados en las instalaciones del usuario y/o medios que no son compartidos con otras entidades. Ejemplos de medios controlados indirectamente incluyen medios que son accesibles indirectamente para el usuario a través de una red externa y/o a través de un servicio que proporciona recursos compartidos, tal como la "nube". Ejemplos de instrucciones del programa incluyen tanto el código de la máquina, tal como el producido por un compilador, y archivos que contienen un código de nivel superior que puede ser ejecutado por el ordenador usando un interpretador.

En una realización, se divulga un producto de programa informático para generar un resultado 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 prueba. El producto informático puede contener instrucciones para realizar uno cualquiera o más de los métodos descritos anteriormente para determinar una fracción de ácidos nucleicos procedente de un organismo particular. Según se explica, el producto informático puede incluir un medio legible informáticamente no transitorio y/o tangible que tenga una lógica (p. ej., instrucciones) ejecutable o recopilable informáticamente registrada sobre el mismo para permitir que un procesador determine la fracción

genómica y, en algunos casos, si una aneuploidía u otra afección está presente o ausente en el genoma. En un ejemplo, el producto informático comprende un medio legible informáticamente que tiene una lógica (p. ej., instrucciones) ejecutable o recopilable informáticamente registrada sobre 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 moléculas de ácido nucleico procedentes de una muestra biológica materna, en donde dichos datos de secuenciación comprenden secuencias en los locus de uno o más polimorfismos; lógica asistida informáticamente para analizar secuencias para determinar recuentos alélicos 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 resultados que indiquen la fracción fetal de ácidos nucleicos en la muestra.

La información de secuencias procedente de la muestra considerada se puede asignar a secuencias de referencia polimórficas según se describe. Además, la información de secuencias asignada se puede usar para generar recuentos alélicos y/o determinar casos de cigosidad para los polimorfismos. Esta información se puede usar para determinar la fracción fetal. En diversas realizaciones, las secuencias de referencia polimórficas se almacenan en una base de datos tal como una base de datos relacional o de objetos, por ejemplo. Se debe entender que no es práctico, o incluso posible en la mayoría de los casos, que un ser humano sin ayuda realice una cualquiera o todas estas operaciones informáticas. Por ejemplo, asignar una sola lectura de 30 pb procedente de una muestra a una base de datos de secuencias de referencia polimórficas llevaría potencialmente un período prohibitivamente prolongado sin la ayuda de un aparato informático. Por supuesto, el problema se complica debido a que las designaciones fiables requieren asignar miles (p. ej., al menos aproximadamente 10.000) o incluso millones de lecturas a uno o más cromosomas.

En ciertas realizaciones, los métodos divulgados hacen uso de una lista almacenada u otra colección organizada de datos relativos a polimorfismos de referencia para el organismo que produce las secuencias de ácido nucleico que se van a analizar. Según se explica anteriormente, las secuencias procedentes de la muestra considerada se pueden alinear o asignar de otro modo a los polimorfismos almacenados. Los polimorfismos individuales son típicamente secuencias de una longitud suficiente para asignar inequívocamente a secuencias identificadas procedentes de la muestra de ácidos nucleicos. Típicamente, los polimorfismos van en grupos, uno para cada alelo. En diversas realizaciones, los polimorfismos de referencia están almacenados en una base de datos que contiene características de los polimorfismos además de sus secuencias. Esta colección de información acerca de los polimorfismos se puede almacenar en una base de datos relacional o de objetos, por ejemplo.

La Figura 10 ilustra un sistema informático típico que, cuando está configurado o diseñado apropiadamente, puede servir como un aparato de análisis de esta 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 incluyendo almacenamiento primario 206 (típicamente una memoria de acceso aleatorio, o RAM), almacenamiento secundario 204 (típicamente una memoria de solo lectura, o ROM). La CPU 202 puede ser de diversos tipos incluyendo microcontroladores y microprocesadores tales como dispositivos programables (p. ej., CPLDs y FPGAs) y dispositivos no programables tales como ASICs de matrices de puertas o microprocesadores de uso general. Como se sabe bien en la técnica, el almacenamiento primario 204 actúa para transferir datos e instrucciones a la CPU y el almacenamiento primario 206 se usa típicamente para transferir datos e instrucciones de un modo bidireccional. Ambos dispositivos de almacenamiento primario pueden incluir cualesquiera medios legibles informáticamente adecuados tales como los descritos anteriormente. Un dispositivo de almacenamiento masivo 208 también está acoplado bidireccionalmente a la CPU 202 y proporciona una capacidad de almacenamiento de datos adicional y puede incluir cualquiera de los medios legibles informáticamente descritos anteriormente. El dispositivo de almacenamiento masivo 208 se puede usar para almacenar programas, datos y similares y es típicamente un medio de almacenamiento secundario tal como un disco duro. Se apreciará que la información retenida dentro del dispositivo de almacenamiento masivo 208, en casos apropiados, se puede incorporar de un modo estándar como parte del almacenamiento primario 206 como memoria virtual. Un dispositivo de almacenamiento masivo específico tal como un CD-ROM 214 también pueden transmitir datos unidireccionalmente a la CPU.

La CPU 202 también se acopla a una interfaz 210 que conecta con uno o más dispositivos de entrada/salida tales como monitores de video, ratones, teclados, micrófonos, pantallas sensibles al tacto, lectores de tarjetas de transductor, tabletas, lápices táctiles, reconocedores de voz o de escritura a mano, u otros dispositivos de entrada bien conocidos tales como, por supuesto, otros ordenadores. Finalmente, la CPU 202 se puede acoplar opcionalmente a un dispositivo externo tal como una base de datos o un ordenador o una red de telecomunicaciones que use una conexión externa según se muestra generalmente en 212. Con esta conexión, se contempla que la CPU pueda recibir información de la red o pueda enviar información a la red en el transcurso de la realización de las etapas del método descritas en la presente.

Los datos de secuencia u otros pueden ser introducidos en un ordenador por un usuario bien directamente o bien indirectamente. En una realización, el sistema informático 200 está acoplado directamente a una herramienta de secuenciación que lee y/o analiza secuencias de ácidos nucleicos amplificados. Las secuencias u otra información procedente de estas herramientas se proporcionan a través de una interfaz 212 para el análisis mediante el sistema 200. Alternativamente, 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 un almacenamiento primario 206 o un almacenamiento masivo 208 introduce en la memoria intermedia o almacena, al menos temporalmente, secuencias de los ácidos nucleicos. Además, el dispositivo de memoria puede almacenar números indicadores 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 asignados. Estos programas/rutinas pueden incluir programas para realizar análisis estadísticos, etc.

En un ejemplo, un usuario introduce una muestra en un aparato de secuenciación. Los datos son recogidos y/o analizados por el aparato de secuenciación que está conectado a un ordenador. Un programa del ordenador permite la recogida y/o el análisis de datos. Los datos se pueden almacenar, exhibir (a través de un monitor u otro dispositivo similar) y/o enviar a otra localización. Según se indica, el ordenador puede estar conectado a Internet que se usa para transmitir datos a un dispositivo manual utilizado por un usuario remoto (p. ej., un médico, científico o analista). Se entiende que los datos se pueden almacenar y/o analizar antes de la transmisión. En algunas realizaciones, se recogen datos brutos y se envían a un usuario (o aparato) remoto que analizará y/o almacenará los datos. La transmisión se puede producir a través de Internet, pero también se puede producir a través de un satélite u otra conexión. Alternativamente, los datos se pueden almacenar en un medio legible informáticamente (p. ej., un CD o un dispositivo de almacenamiento de memoria semiconductor) y el medio se puede enviar a un usuario final (p. ej., a través del correo). El usuario remoto puede estar en la misma u otra localización geográfica incluyendo, pero no limitada a, un edificio, una ciudad, un estado, un país o un continente.

En algunas realizaciones, los métodos de la invención comprenden además recoger datos referentes a una pluralidad de secuencias polinucleotídicas y enviar los datos a un ordenador. Por ejemplo, el ordenador puede estar conectado a un equipo de laboratorio, p. ej., 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 datos aplicables reunidos por el dispositivo de laboratorio. Los datos se pueden almacenar en un ordenador en cualquier etapa, p. ej., mientras se recogen en tiempo real, antes del envío, durante o junto con el envío o después del envío. Los datos se pueden almacenar en un medio legible informáticamente que se puede extraer del ordenador. Los datos recogidos o almacenados se pueden transmitir desde el ordenador a una localización remota, p. ej., a través de una red local o una red de área amplia tal como Internet.

En un aspecto, la divulgación proporciona además un sistema capaz de realizar un análisis cuantitativo de secuenciación de nucleótidos con una precisión de al menos 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% o al menos 99%. La secuenciación de nucleótidos puede comprender secuenciación de Sanger, secuenciación masivamente paralela, hibridación u otras técnicas como las descritas en la presente. El sistema puede comprender diversos componentes, p. ej., equipo de laboratorio y sistemas informáticos, y se puede configurar para llevar a cabo los métodos de la invención.

En algunas realizaciones, las instrucciones del aparato y/o de programación pueden incluir además instrucciones para registrar automáticamente información relativa al método tal como la fracción de ADN fetal y opcionalmente la presencia o ausencia de una aneuploidía cromosómica fetal en una historia clínica de un paciente para un sujeto humano que proporciona la muestra de ensayo materna. La historia clínica del paciente puede ser mantenida, por ejemplo, por un laboratorio, una consulta médica, un hospital, un seguro médico restringido, una compañía aseguradora o un cibersitio de historias clínicas personales. Además, basándose en los resultados del análisis ejecutado por el procesador, el método puede implicar además prescribir, iniciar y/o alterar el tratamiento de un sujeto humano del que se recoge la muestra de sangre materna. Esto puede implicar realizar una o más pruebas o análisis adicionales sobre muestras adicionales recogidas del sujeto.

Ejemplo

Fracción Fetal Predicha a partir de las Variaciones Secuenciadas: Caso 2

Para demostrar que el presente método se puede usar para estimar fiablemente la fracción fetal en una muestra materna, se creó una muestra 'materna' artificial, y se identificaron las variaciones de bases en todos los locus de los cromosomas 1 y 7 para predecir la fracción del genoma de menor contribución.

El ADNcf que se aísla de una embarazada es una mezcla de ADNcf materno y fetal, correspondiendo el nivel de ADNcf fetal a una mediana de ~ 10% del ADNcf total (Lo y cols., 2010, "Maternal Plasma DNA seqeucning reveals the genome-wide genetic and mutational profile of the fetus", Prenatal Diagnosis, 2, 1-12). Para crear la muestra materna artificial, se usó ADN genómico (ADNg) obtenido de una madre y su hijo (ADNs de la madre y el hijo NA10924 y NA10925; TheCoriell Institute for Medical Research, Camden, NJ) para crear la muestra de genomas mixtos. Se cortaron cinco microgramos de cada ADNg de la madre y el hijo en fragmentos de aproximadamente 200 pb, y se determinó la concentración de cada uno. Se creó una muestra artificial que contenía 10% de ADN procedente del hijo y 90% de ADN procedente de la madre para imitar una muestra de sangre materna, que se cree que contiene típicamente 2-40% de ADNcf fetal, dependiendo de la edad gestacional [Lun y cols, 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 carriles de la celdilla de flujo usando el IlluminaHiSeq 2000. Se generaron aproximadamente 800 millones de lecturas de secuencias 49-meras.

Los ~ 800 millones de lecturas se alinearon al genoma de referencia humano de repeticiones enmascaradas (construcción hg19) usando el algoritmo GSNAP (http://research-pub.gene.com/gmap/), permitiendo una discordancia y no permitiendo inserciones ni eliminaciones. Las secuencias que se asignaban a múltiples localizaciones en el genoma se ignoraron. Todas las otras lecturas asignadas se contaron como marcadores de secuencia, y solo los locus a los que se asignaban los marcadores de secuencia 40 y 100 se consideraban para un análisis adicional, es decir solo se consideraban las bases que tenían cobertura de los marcadores 40 y 100.

Para cada locus de bases, se contó el número de marcadores que se asignaban a cada una de las cuatro bases. Los locus que tenían más de dos bases posibles se eliminaban, y solo se usaban los marcadores que se asignaban a locus monoalélicos y bialélicos para predecir la fracción fetal artificial. El número total de marcadores que se asignaban a cada locus de bases representaba la cobertura (D) en ese locus. En esta muestra materna simulada, se espera que la contribución del alelo principal de la madre (B) refleje la porción de 90% de los marcadores y la contribución del alelo secundario del hijo (A) refleje la porción de 10% de los marcadores.

Las Figuras 11 A y B muestran histogramas del número de observaciones (frecuencia) de bases variantes en los cromosomas 1 y 7, respectivamente, para los porcentajes de alelo secundario (A/D) para los cromosomas 1 y 7. El porcentaje de alelo secundario es el porcentaje del número total de alelos en un locus dado. Por ejemplo, para un locus dado en el que haya 8 presencias de alelo secundario A y 56 presencias de alelo principal B, entonces el porcentaje de alelo secundario es 8%. Los datos muestran que el mayor número de presencias (frecuencia) para el alelo secundario se observa cuando el alelo secundario está presente en 5%, lo que representa la mitad de la fracción fetal. Según esto, los datos predecían que la muestra contenía una fracción fetal de 10%, lo que corresponde a lo que se usaba para crear la muestra materna artificial.

Las Figuras 12A y B muestran la distribución de la frecuencia alélica a lo largo de los cromosomas 1 y 7, respectivamente. Ambas gráficas muestran que el número máximo de alelos variantes a lo largo de los cromosomas se produce a una frecuencia del alelo secundario de 5% y una frecuencia del alelo principal de 95%. Algunos de los puntos de datos restantes representan locus bialélicos presentes en el genoma de la madre, mientras que otros representar el ruido de la metodología de secuenciación. La porción central de cada gráfica en la que no se representan alelos variantes coincide con los centrómeros del cromosoma, que se sabe que son regiones de los cromosomas ricas en repeticiones, a las que se asignan marcadores en más de un locus y por lo tanto se excluyen del análisis. En otras regiones, por ejemplo regiones que flanquean el centrómero y regiones correspondientes a telómeros, los alelos variantes están sobrerrepresentados. La sobrerrepresentación de estas regiones se puede atribuir a la metodología de secuenciación con lo que algunas regiones se secuencian a niveles mayores que otras.

Por lo tanto, el presente método se puede usar para predecir la fracción fetal. El método es particularmente útil ya que no requiere la identificación de secuencias elegidas como diana, p. ej. SNPs, y cualquier variación en cualquier posición de cualquier cromosoma puede servir para predecir el porcentaje de fracción fetal.

Otras Realizaciones

Aunque lo anterior ha descrito generalmente la divulgación según procedimientos y aparatos específicos, la presente divulgación tiene un intervalo de aplicabilidad mucho más amplio. En particular, la presente divulgación se ha descrito en cuanto a detectar la fracción de ADN fetal en una muestra de ADN tomada de una embarazada, pero no se limita de ese modo, ya que los conceptos y los métodos presentados aquí también se pueden aplicar en otros contextos tales como la detección de las cantidades relativas de tipos de ADN en una muestra que tenga ADN originario de dos o más genomas diferentes. Por supuesto, los expertos normales en la técnica identificarán otras variaciones, modificaciones y alternativas.

Por ejemplo, aunque la mayoría de los ejemplos y las aplicaciones descritos en la presente tratan de la estimación de la fracción fetal de ADN en una muestra de ADN tomada de una embarazada, la divulgación no está limitada de ese modo. Más generalmente, diversas realizaciones divulgan métodos para evaluar cantidades relativas de ácido nucleico procedente de dos genomas diferentes en una muestra de prueba que contiene una mezcla de ácidos nucleicos procedentes de los dos genomas diferentes, y que se sabe o se sospecha que 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 aquí trata de muestras tomadas de una mujer embarazada, la divulgación no se limita de ese modo. Por ejemplo, el individuo que proporciona una muestra que se va a probar puede ser un organismo que comprende secuencias polinucleotídicas, p. ej., una planta, un insecto tal como una mosca, o un animal. En algunas realizaciones, el sujeto es un mamífero, p. ej., un ratón, una rata, un perro, un mono o un ser humano. Según se indica, el sujeto puede ser una embarazada. El sujeto podría ser un individuo con una enfermedad tal como un cáncer, o podría estar infectado con un cuerpo extraño tal como un microorganismo, p. ej.,

un virus. La muestra puede comprender un líquido corporal procedente del sujeto, p. ej., sangre, plasma, suero, esputos, saliva, orina, excremento, pus, linfa, moco o similares. Por ejemplo, la muestra puede ser una muestra de plasma materno que contiene una mezcla de ADN libre de células materno y fetal. Generalmente, los métodos divulgados pueden implicar secuenciar ADN procedente de una muestra; asignar las lecturas de secuencia a polimorfismos; clasificar los polimorfismos basándose en la cigosidad; y estimar la fracción de ADN procedente de una fuente secundaria en la muestra.

Apéndice 1 Lista de Secuencias de Bases 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

LISTA DE SECUENCIAS

<110> Verinata Health, Inc.

<120> RESOLUCIÓN DE FRACCIONES GENÓMICAS USANDO RECUENTOS DE POLIMORFISMOS

<130> P073615EP

<140> EP 12716939.9

<141> 2012-04-12

<140> PCT/US2012/033391

<141> 2012-04-12

<150> 61/474,362

<151> 2011-04-12

<160> 32

<170> PatentIn versión 3.5

<210> 1

<211> 111

<212> ADN

<213> Homo sapiens

<400> 1

cacatgcaca gccagcaacc ctgtcagcag gagttcccac cagtttcttt ctgagaacat 60

ctgttcaggt ttctctccat ctctatttac tcaggtcaca ggaccttggg g 111

<210> 2

<211> 111

<212> ADN

<213> Homo sapiens

<400> 2

cacatgcaca gccagcaacc ctgtcagcag gagttcccac cagtttcttt ctgagaacat 60

ctgttcaggt ttctctccat ctctgtttac tcaggtcaca ggaccttggg g 111

<210> 3

<211> 126

<212> ADN

<213> Homo sapiens

<400> 3

tgaggaagtg aggctcagag ggtaagaaac tttgtcacag agctggtggt gagggtggag 60

attttacact ccctgcctcc cacaccagtt tctccagagt ggaaagactt tcatctcgca 120

ctggca 126

<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 ttcctcttt 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 ttcctcttt 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 ttcctccccc cccatcaacc tcttttgtgc ctccccctcc 60

gtgtaccacc ttctctgtca ccaaccctgg cctcacaact ctctcctttg ccac 114

<210> 28

<211> 114

<212> ADN

<213> Homo sapiens

<400> 28

cagtggaccc tgctgcacct ttcctccccc 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. Un método para estimar la fracción de ADN fetal en ADN obtenido de un líquido corporal de una embarazada, comprendiendo el método:

(a) alinear o asignar de otro modo secuencias de segmentos de ADN derivadas de la secuenciación del ADN en el líquido a uno o más polimorfismos indicados en una secuencia de referencia, en donde el alineamiento o la asignación de otro modo se realiza usando un aparato informático programado para asignar secuencias de ácido nucleico al uno o más polimorfismos indicados;

(b) determinar frecuencias alélicas de las secuencias de segmentos de ADN asignadas para al menos uno de los polimorfismos indicados;

(c) clasificar el al menos un polimorfismo indicado basándose en una combinación de la cigosidad de la embarazada y la cigosidad del feto; y

(d) estimar la fracción de ADN fetal en el ADN obtenido de la embarazada usando las frecuencias alélicas determinadas en (b) junto con la clasificación de cigosidades procedente de (c),

en donde (b)-(d) se realizan en uno o más procesadores que funcionan bajo las instrucciones del programa para la determinación, la clasificación y la estimación; y

en donde la clasificación en (c) clasifica el al menos un polimorfismo indicado en una de las siguientes combinaciones: (i) la embarazada es homocigótica y el feto es homocigótico, (ii) la embarazada es homocigótica y el feto es heterocigótico, (iii) la embarazada es heterocigótica y el feto es homocigótico, y (iv) la embarazada es heterocigótica y el feto es heterocigótico.

2. El método según la reivindicación 1:

(a) que comprende además dejar de considerar cualquier polimorfismo clasificado en la combinación (i) o la combinación (iv);

(b) que comprende además filtrar el al menos un polimorfismo indicado para dejar de considerar cualquier polimorfismo que tenga una frecuencia del alelo secundario mayor que un umbral definido;

(c) que comprende además filtrar el al menos un polimorfismo indicado para dejar de considerar cualquier polimorfismo que tenga una frecuencia del alelo secundario menor que un umbral definido;

(d) en donde la clasificación del al menos un polimorfismo indicado comprende aplicar un umbral a la frecuencia alélica determinada en (b);

(e) en donde la clasificación del al menos un polimorfismo indicado comprende aplicar los datos de frecuencia alélica procedentes de (b), obtenidos para una pluralidad de polimorfismos, a un a modelo mixto, opcionalmente en donde el modelo mixto emplea momentos factoriales;

(f) en donde el ADN obtenido de un líquido corporal de una embarazada es ADN libre de células obtenido del plasma de la embarazada;

(g) en donde la asignación de los segmentos de ADN obtenidos a partir de la sangre de la embarazada comprende asignar informáticamente dichos segmentos a una base de datos de polimorfismos;

(h) que comprende además ejecutar las instrucciones del programa en el uno o más procesadores para registrar automáticamente la fracción de ADN fetal que se estima en (d) en una historia clínica del paciente, almacenada en un medio legible informáticamente, para la embarazada;

(i) que comprende además, basándose en la estimación de la fracción de ADN fetal, prescribir, iniciar y/o alterar el tratamiento de un sujeto humano del que se tomaba la muestra de prueba materna; o

(j) que comprende además, basándose en la estimación de la fracción de ADN fetal, encargar y/o realizar una o más pruebas adicionales.

3. El método según la reivindicación 1, que comprende además, antes de la etapa (a), secuenciar el ADN con un secuenciador de ácido nucleicos bajo condiciones que produzcan secuencias de segmentos de ADN que contienen uno o más polimorfismos, opcionalmente en donde la secuenciación se efectúa sin amplificar selectivamente ninguno de los uno o más polimorfismos indicados.

4. El método según la reivindicación 3, que comprende además, antes de secuenciar el ADN, extraer ADN de la muestra bajo condiciones que extraigan ADN tanto de un genoma materno como de un genoma fetal presentes en el líquido corporal.

5. El método según la reivindicación 4, que comprende además, antes de extraer ADN de la muestra, recibir una muestra del líquido corporal.

6. El método según una cualquiera de las reivindicaciones 3-5, que comprende la secuenciación por síntesis.

7. El método según una cualquiera de las reivindicaciones 3-5, que comprende la secuenciación por hibridación.

8. El método según la reivindicación 7, en donde la hibridación comprende poner en contacto una pluralidad de secuencias polinucleotídicas con una pluralidad de sondas polinucleotídicas, en donde cada una de la pluralidad de sondas polinucleotídicas se fija a un sustrato, en donde el sustrato es una superficie plana que comprende una serie de secuencias nucleotídicas conocidas.

9. El método según la reivindicación 8, en donde el patrón de hibridación a la serie se usa para determinar las secuencias polinucleotídicas presentes en la muestra.

10. El método según una cualquiera de las reivindicaciones 1-9, en donde la secuencia de referencia es una lista almacenada u otra colección organizada de datos relativos a polimorfismos de referencia para la embarazada, opcionalmente en donde la secuencia de referencia es una base de datos de secuencias, por ejemplo una base de datos de secuencias alélicas.

11. Un aparato para estimar la fracción de ADN fetal en ADN obtenido de un líquido corporal de una embarazada, comprendiendo el aparato:

(a) un secuenciador configurado para (i) recibir ADN extraído de una muestra del líquido corporal que comprende ADN tanto de un genoma materno como de un genoma fetal, y (ii) secuenciar el ADN extraído bajo condiciones que produzcan secuencias de segmentos de ADN que contienen uno o más polimorfismos indicados; y

(b) un aparato informático configurado para dar instrucciones a uno o más procesadores para

alinear o asignar de otro modo secuencias de ácido nucleico al uno o más polimorfismos indicados en una secuencia de referencia,

determinar frecuencias alélicas de las secuencias de segmentos de ADN asignadas para al menos uno de los polimorfismos indicados,

clasificar el al menos un polimorfismo indicado basándose en una combinación de la cigosidad de la embarazada y la cigosidad del feto, y

estimar la fracción de ADN fetal en el ADN obtenido de la embarazada usando las frecuencias alélicas junto con la clasificación de cigosidades; y

en donde el aparato informático está configurado además para dar instrucciones al uno o más procesadores para clasificar el al menos un polimorfismo indicado en una de las siguientes combinaciones: (i) la embarazada es homocigótica y el feto es homocigótico, (ii) la embarazada es homocigótica y el feto es heterocigótico, (iii) la embarazada es heterocigótica y el feto es homocigótico, y (iv) la embarazada es heterocigótica y el feto es heterocigótico.

12. El aparato según la reivindicación 11:

(a) que comprende además una herramienta para extraer ADN de la muestra bajo condiciones que extraigan ADN tanto del genoma materno como del genoma fetal;

(b) en donde el aparato informático está configurado además para dar instrucciones al uno o más procesadores para dejar de considerar cualquier polimorfismo clasificado en la combinación (i) o la combinación (iv);

(c) en donde el aparato informático está configurado además para dar instrucciones al uno o más procesadores para dejar de considerar cualquier polimorfismo que tenga una frecuencia del alelo secundario mayor que un umbral definido;

(d) en donde el aparato informático está configurado además para dar instrucciones al uno o más procesadores para filtrar el uno o más polimorfismos indicados para dejar de considerar cualquier polimorfismo que tenga una frecuencia del alelo secundario menor que un umbral definido;

(e) en donde el aparato informático está configurado además para dar instrucciones al uno o más procesadores para clasificar el al menos un polimorfismo indicado al aplicar un umbral a la frecuencia alélica;

(f) en donde el aparato informático está configurado además para dar instrucciones al uno o más procesadores para clasificar el al menos un polimorfismo indicado al aplicar los datos de frecuencia alélica, obtenidos para una pluralidad de polimorfismos, a un modelo mixto, opcionalmente en donde el modelo mixto emplea momentos factoriales;

(g) que comprende además un aparato para extraer ADN libre de células obtenido del plasma de la embarazada para la secuenciación en el secuenciador;

(h) que comprende además una base de datos de polimorfismos, en donde el aparato informático está configurado además para dar instrucciones al uno o más procesadores para asignar los segmentos de ADN obtenidos de la sangre de la embarazada al asignar informáticamente dichos segmentos a la base de datos de polimorfismos; o

(i) en donde el aparato informático está configurado además para dar instrucciones al uno o más procesadores para registrar automáticamente la fracción de fetal de ADN en una historia clínica del paciente, almacenada en un medio legible informáticamente, para la embarazada, opcionalmente en donde dicha historia clínica del paciente es mantenida por un laboratorio, una consulta médica, un hospital, un seguro médico restringido, una compañía de seguros o un cibersitio de historias clínicas personales.

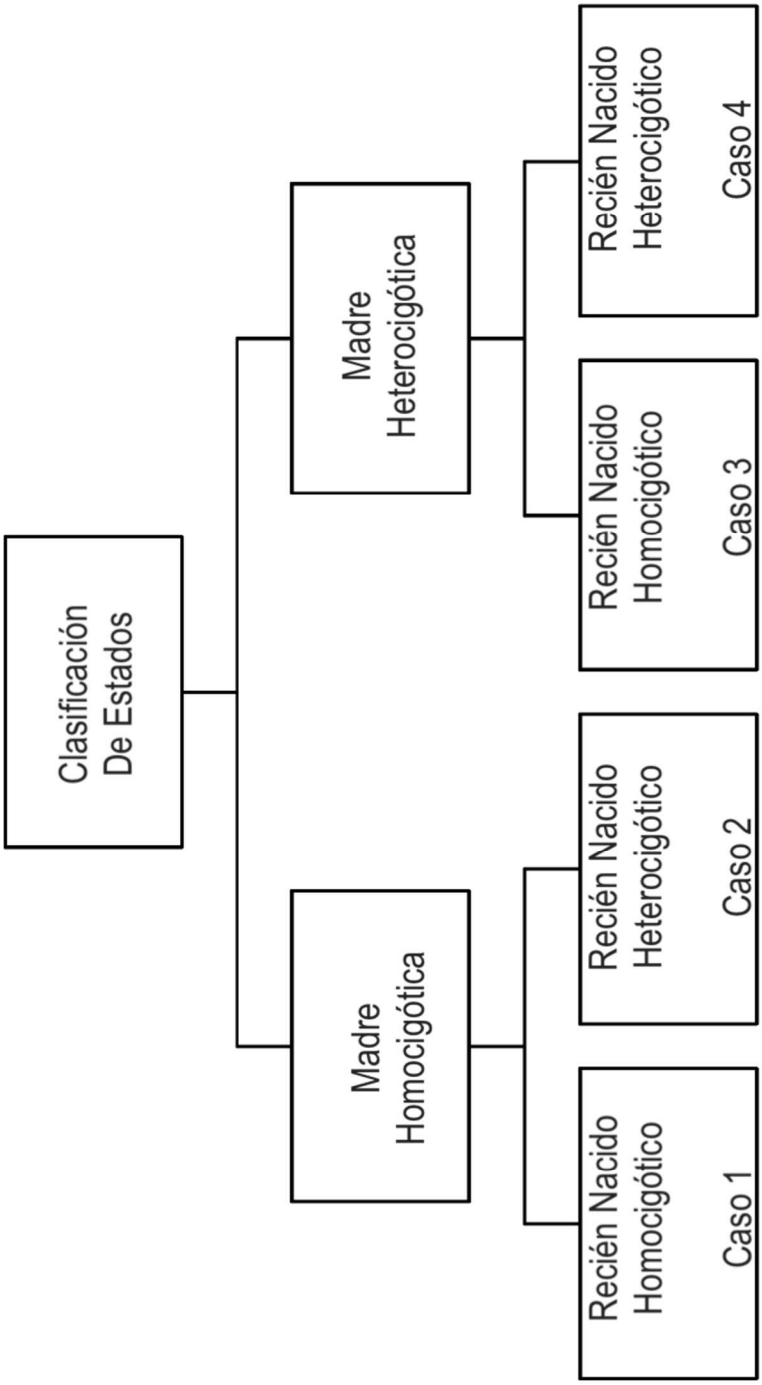
Clasificación De Estados
Madre Homocigótica
Madre Heterocigótica
Recién Nacido Homocigótico Caso 1
Recién Nacido Heterocigótico Caso 2
Recién Nacido Homocigótico Caso 3
Recién Nacido Heterocigótico Caso 4

FIG. 1

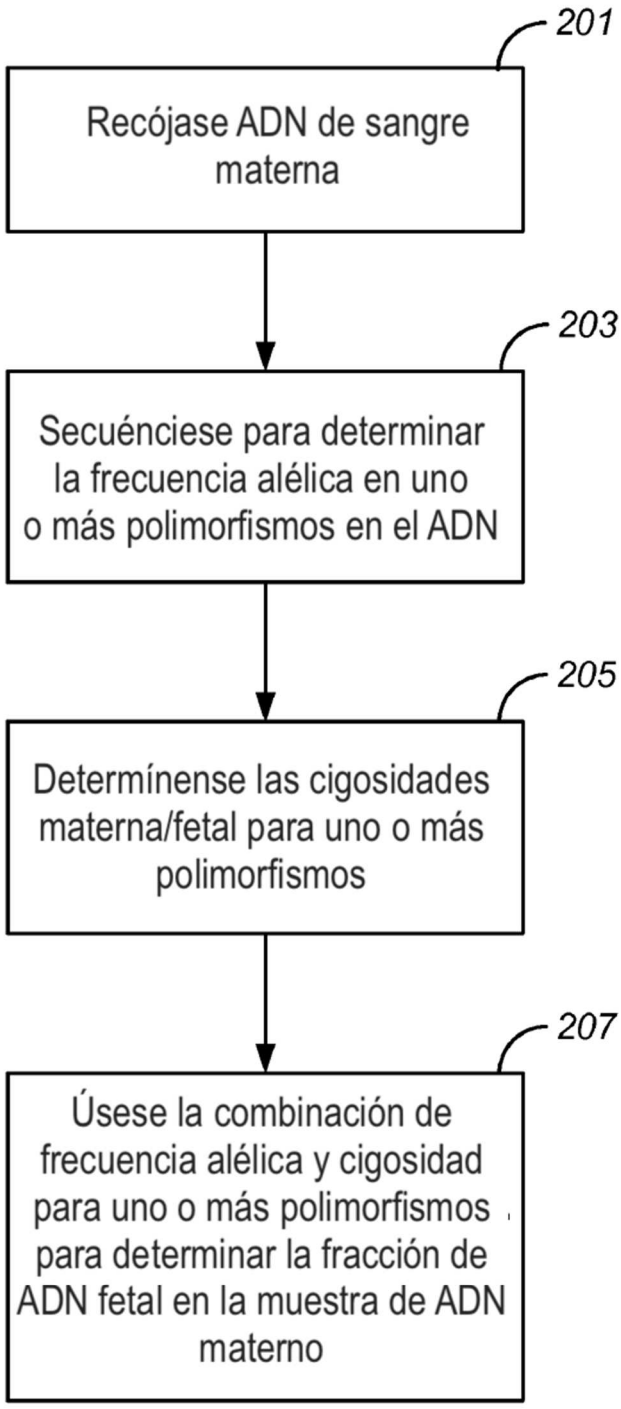
201
Recójase ADN de sangre materna
203
Secuénciese para determinar la frecuencia alélica en uno o más polimorfismos en el ADN
205
Determínense las cigosidades materna/fetal para uno o más polimorfismos
207
Úsese la combinación de frecuencia alélica y cigosidad para uno o más polimorfismos para determinar la fracción de ADN fetal en la muestra de ADN materno

FIG. 2

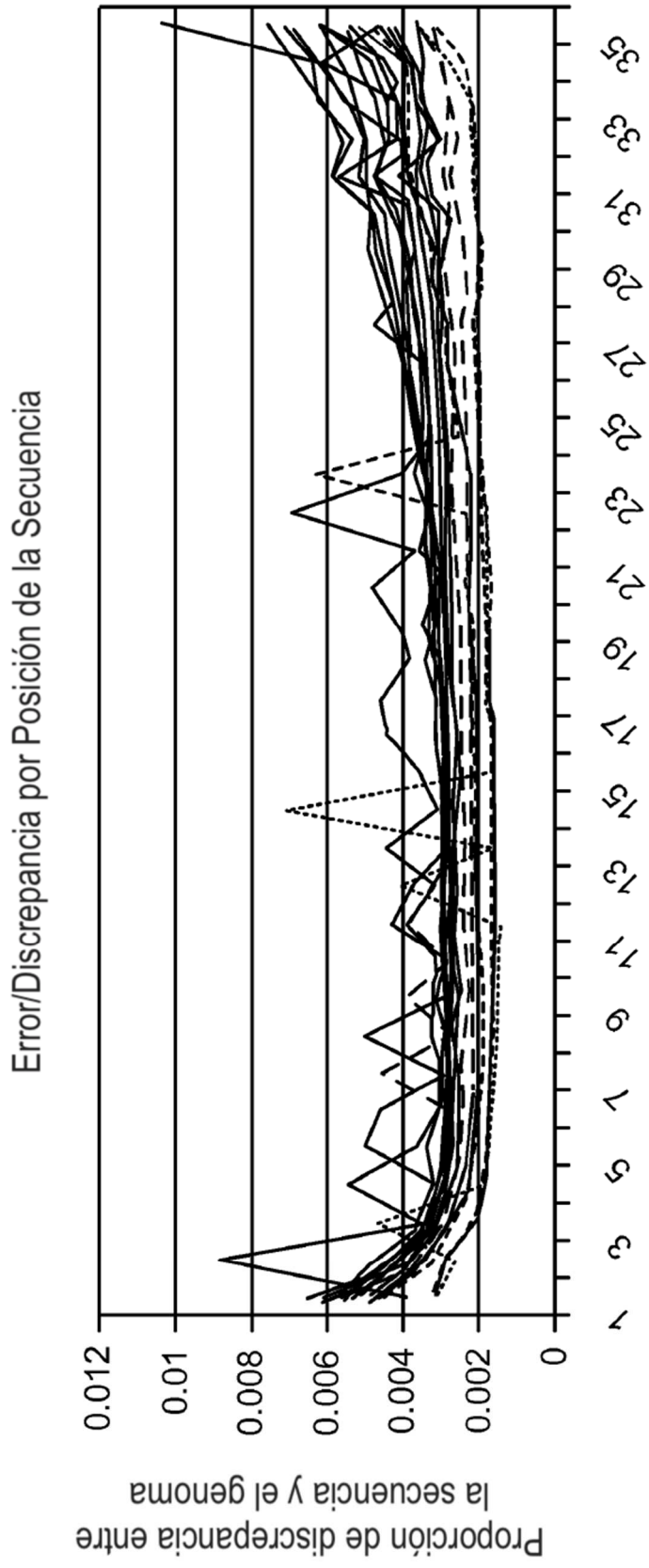
Error/Discrepancia por Posición de la Secuencia
Proporción de discrepancia entre la secuencia y el genoma
0.012
0.01
0.008
0.006
0.004
0.002
0
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
Posición de la Base

FIG. 3

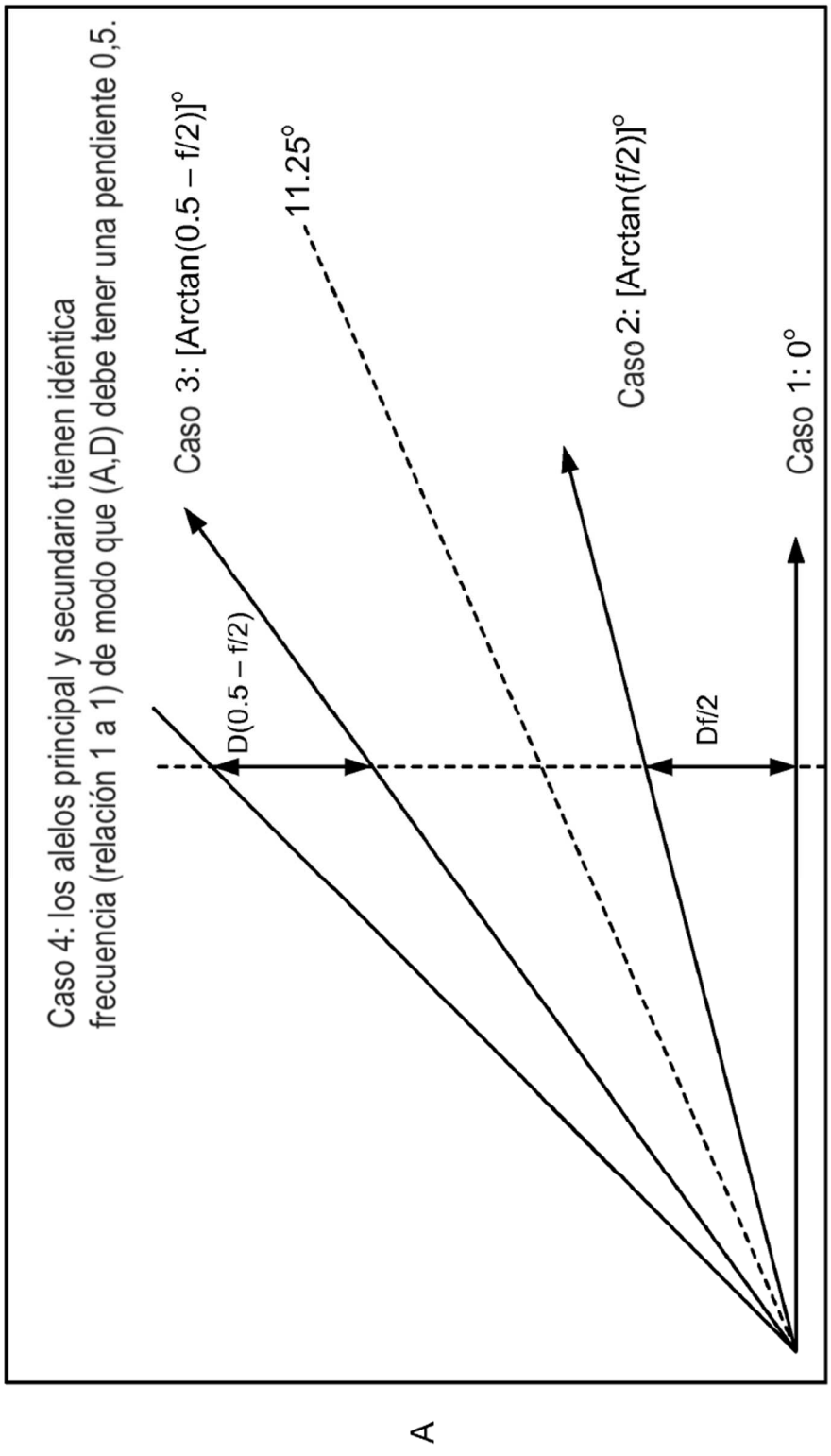
Caso 4: los alelos principal y secundario tienen idéntica frecuencia (relación 1 a 1) de modo que (A,D) debe tener una pendiente 0,5.
Caso 3: [Arctan(0.5 – f/2)]°
11.25°
Caso 2: [Arctan(f/2)]°
Caso 1: 0°
D(0.5 – f/2)
Df/2
D=A+B
A

FIG. 4

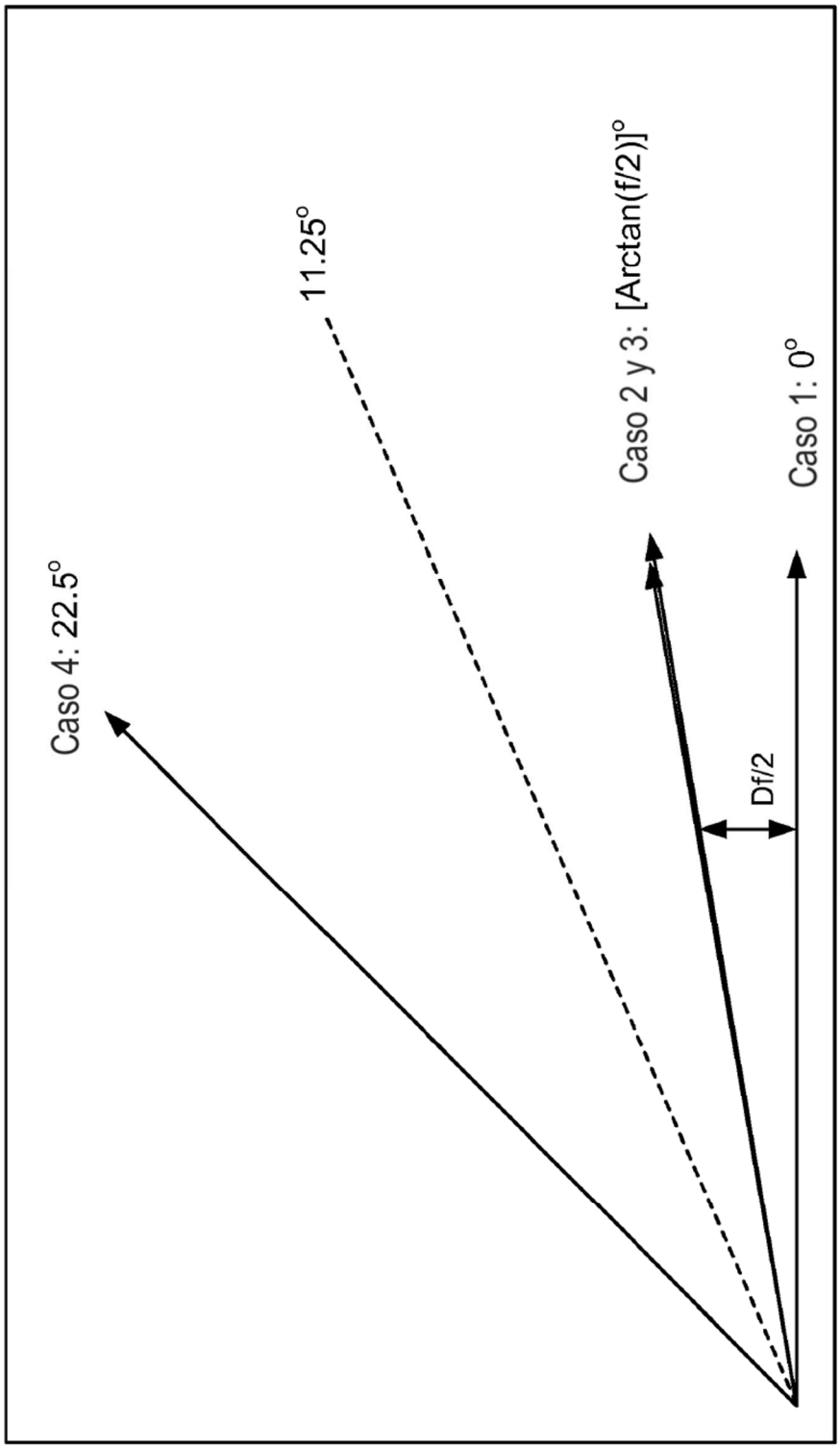

FIG. 5

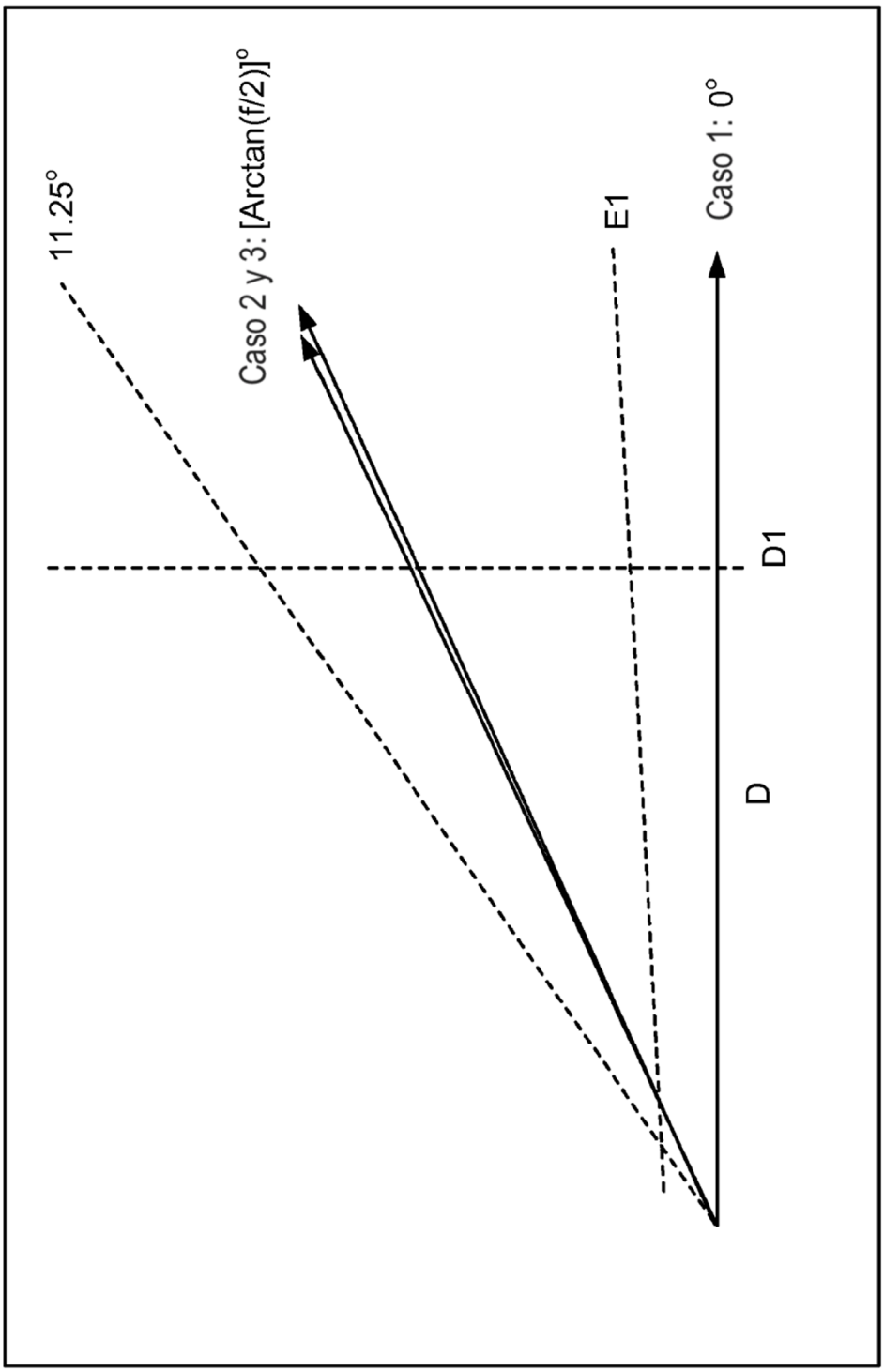

FIG. 6

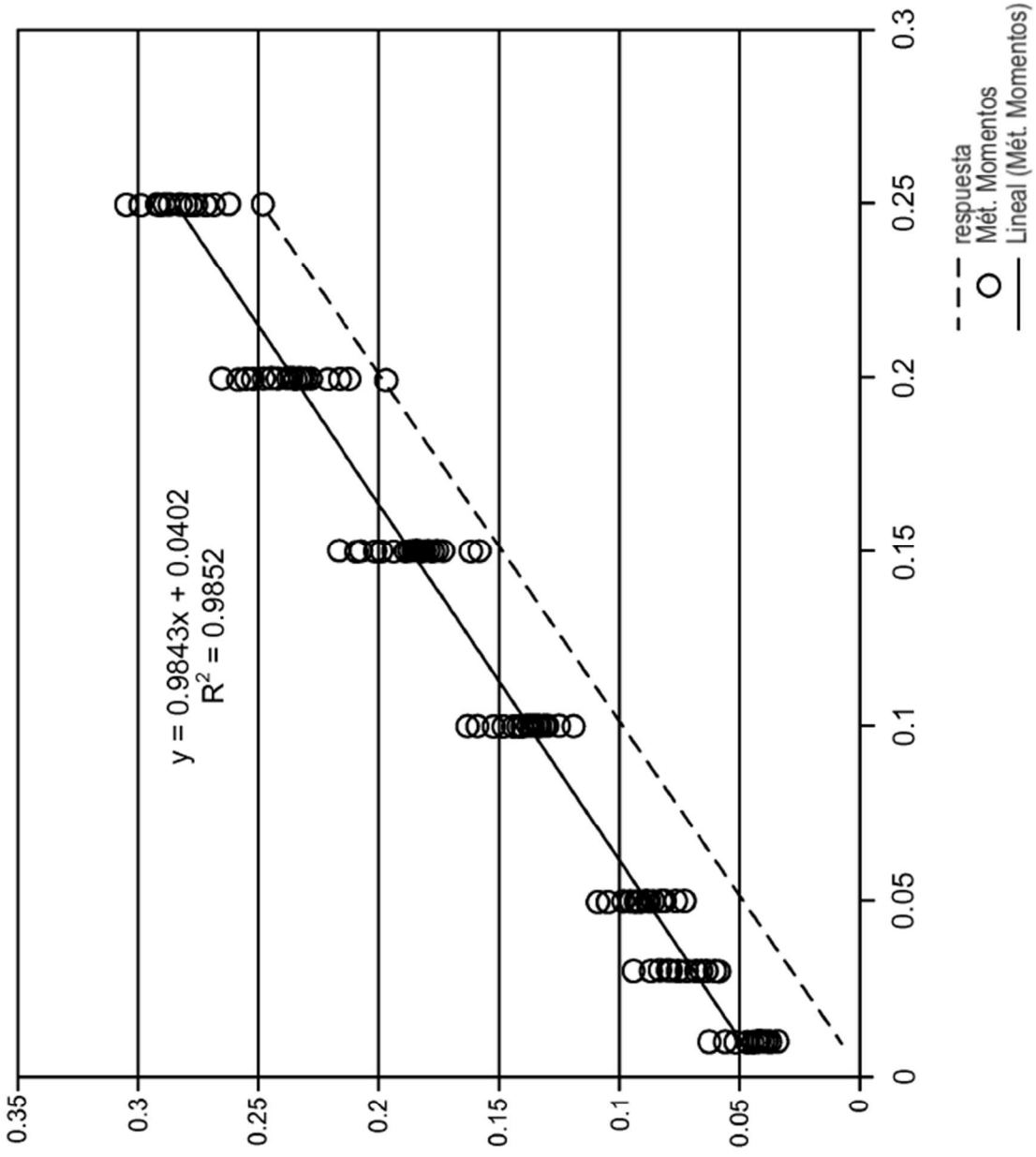
y = 0.9843x + 0.0402
R² = 0.9852
respuesta
Mét. Momentos
Lineal (Mét. Momentos)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.05
0.1
0.15
0.2
0.25
0.3

FIG. 7

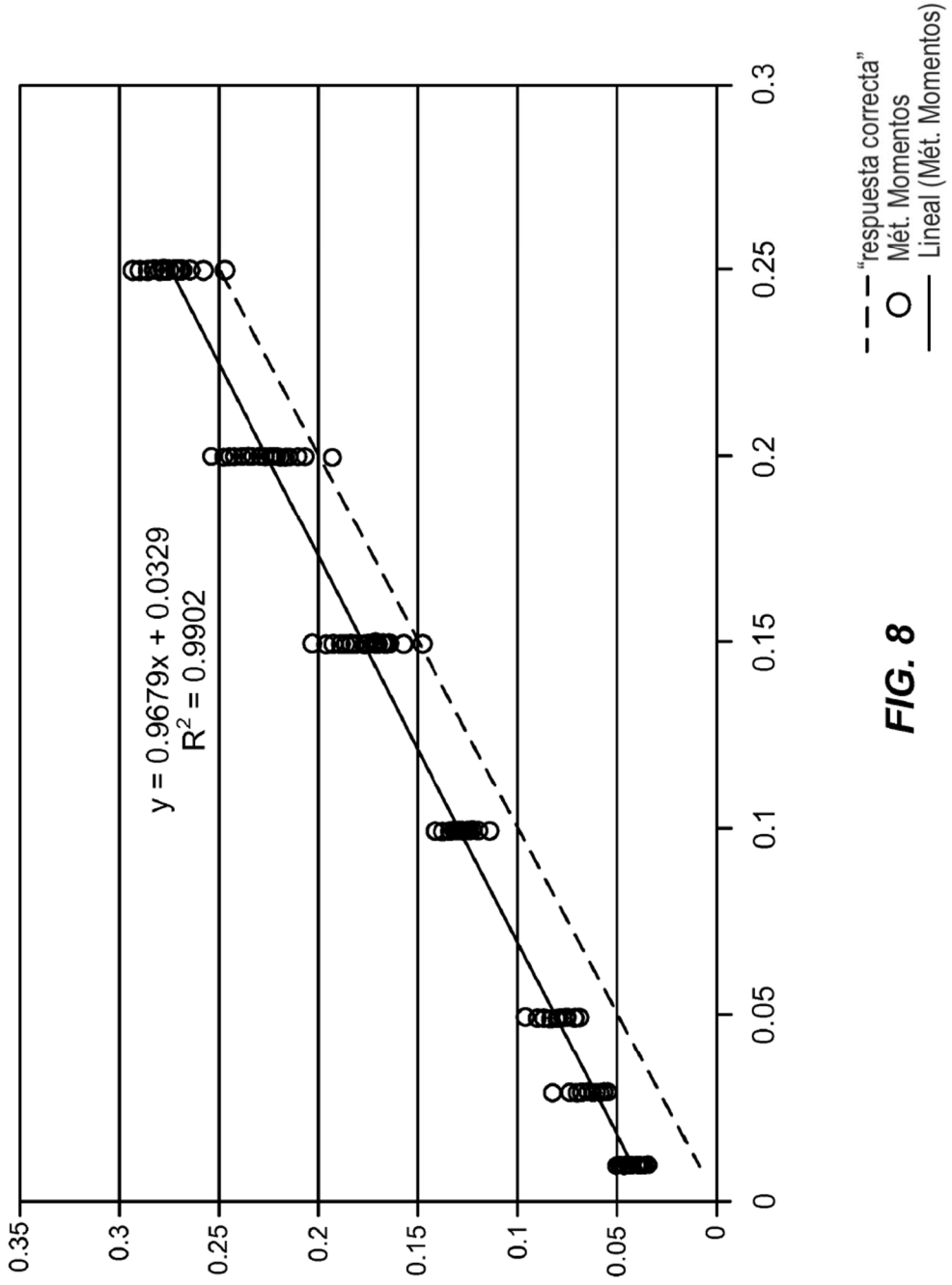
y = 0.9679x + 0.0329
R2 = 0.9902
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.05
0.1
0.15
0.2
0.25
0.3
"respuesta correcta"
Mét. Momentos
Lineal (Mét. Momentos)

FIG. 8

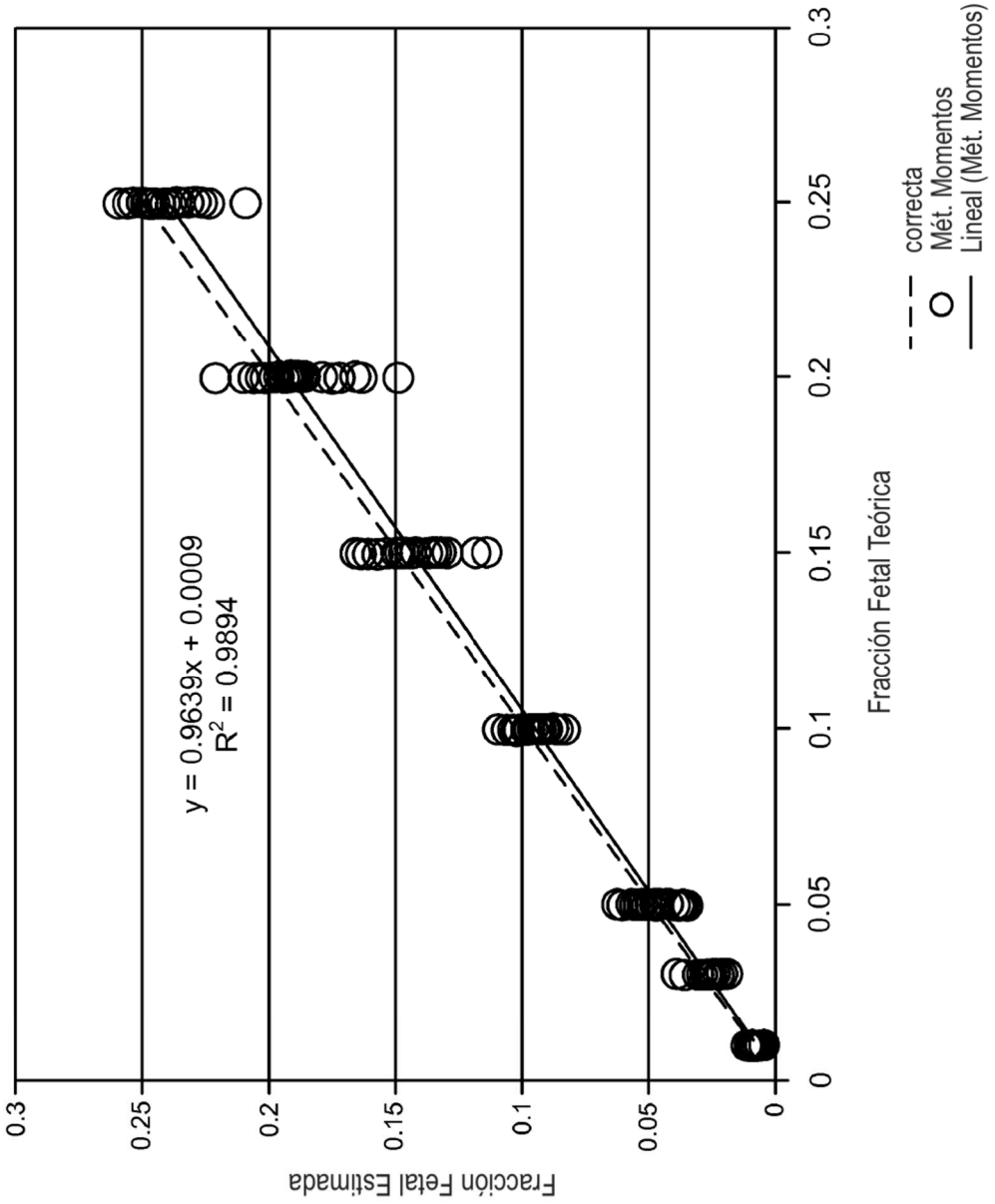
Fracción Fetal Estimada
0.3
0.25
0.2
0.15
0.1
0.05
0
y = 0.9639x + 0.0009
R2 = 0.9894
Fracción Fetal Teórica
0
0.05
0.1
0.15
0.2
0.25
0.3
correcta
Mét. Momentos
Lineal (Mét. Momentos)

FIG. 9

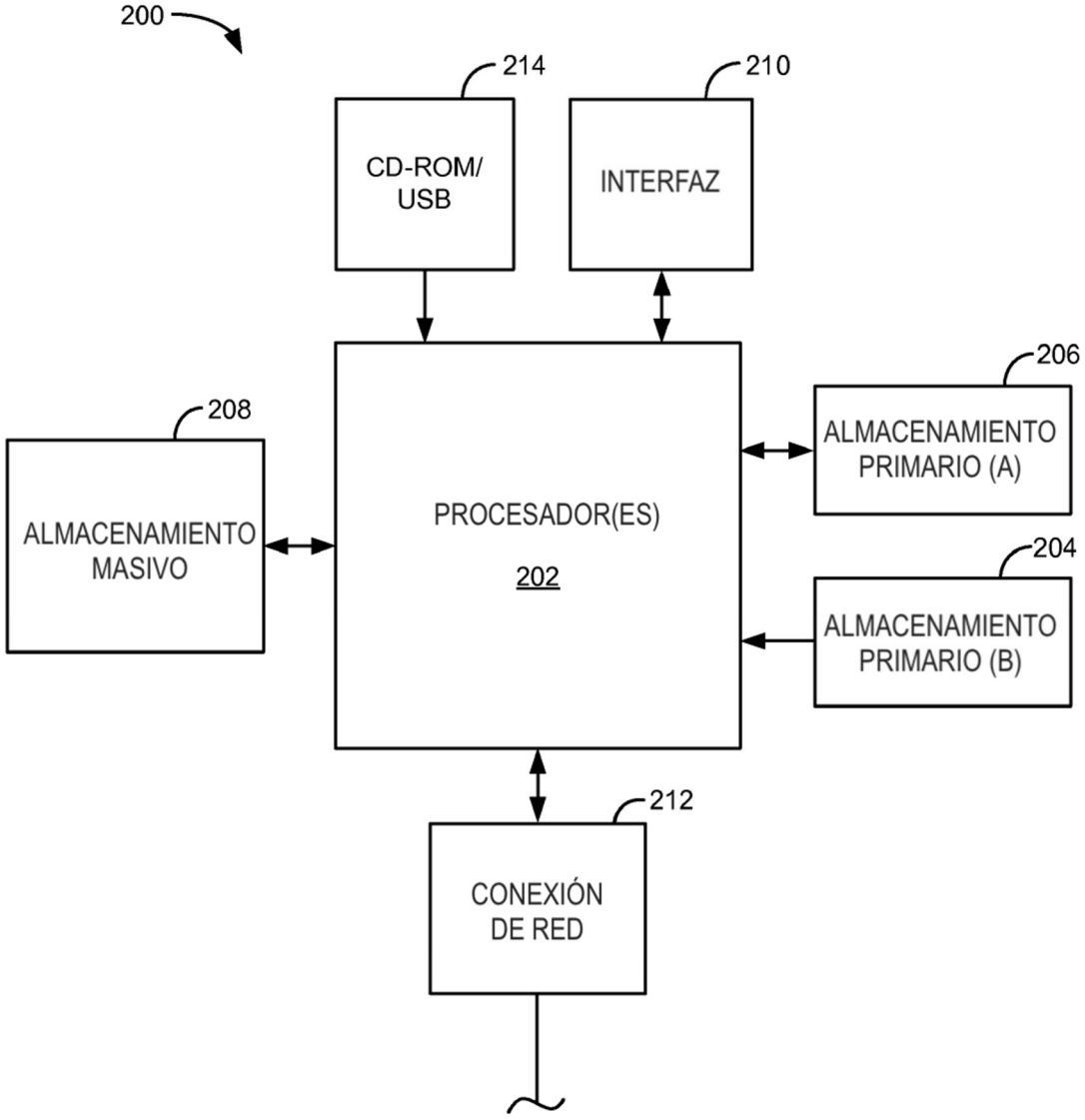
200
214
CD-ROM/
USB
210
INTERFAZ
206
ALMACENAMIENTO
PRIMARIO (A)
208
ALMACENAMIENTO
MASIVO
PROCESADOR(ES)
202
204
ALMACENAMIENTO
PRIMARIO (B)
212
CONEXIÓN
DE RED

FIG. 10

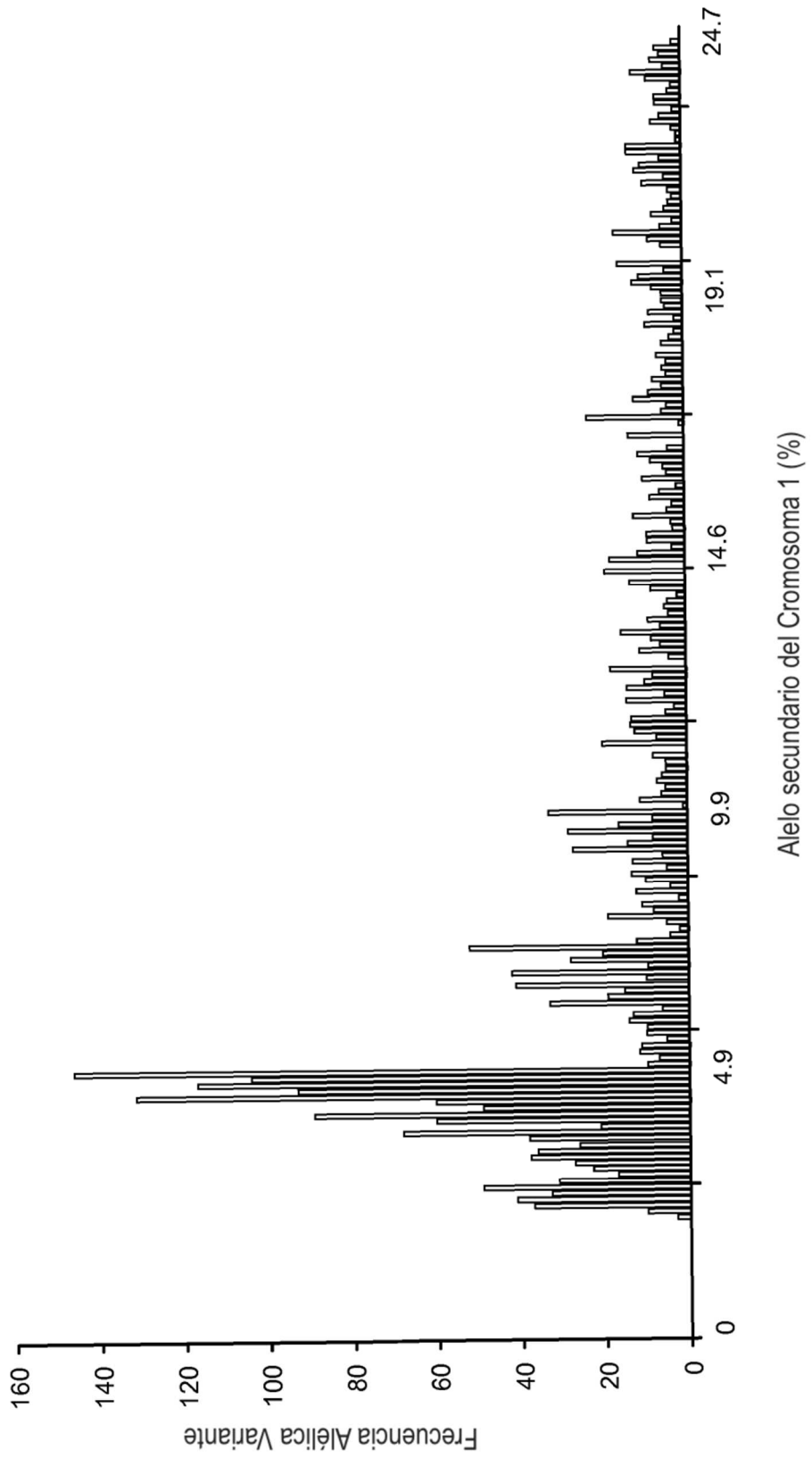
Frecuencia Alélica Variante
160
140
120
100
80
60
40
20
0
0
4.9
9.9
14.6
19.1
24.7
Alelo secundario del Cromosoma 1 (%)

FIG. 11A

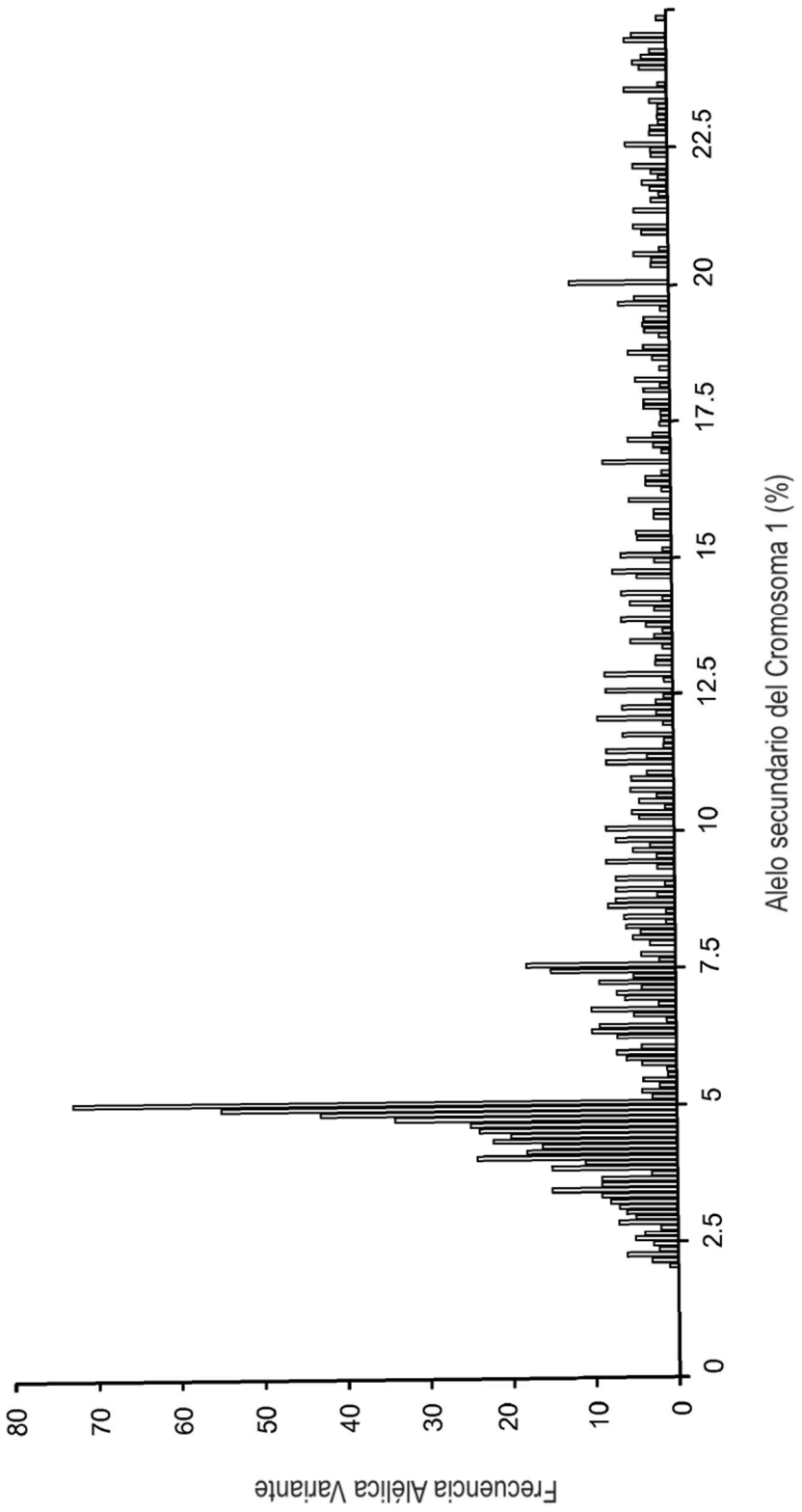
Frecuencia Alélica Variante
0
10
20
30
40
50
60
70
80
0
2.5
5
7.5
10
12.5
15
17.5
20
22.5
Alelo secundario del Cromosoma 1 (%)

FIG. 11B

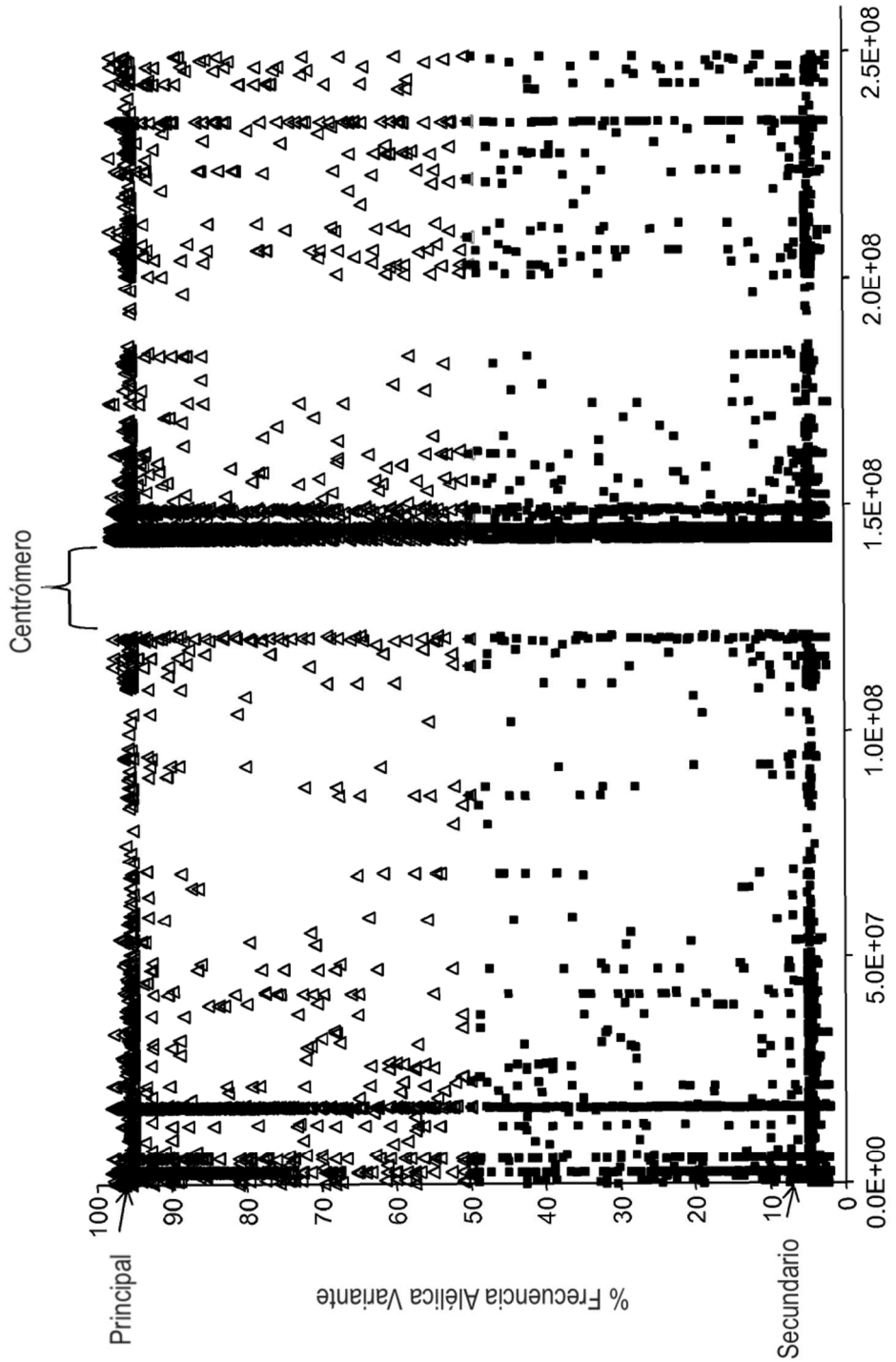
Centrómero
Principal
Secundario
% Frecuencia Alélica Variante
100
90
80
70
60
50
40
30
20
10
0
0.0E+00
5.0E+07
1.0E+08
1.5E+08
2.0E+08
2.5E+08
Posición en el Cromosoma 1

FIG. 12A

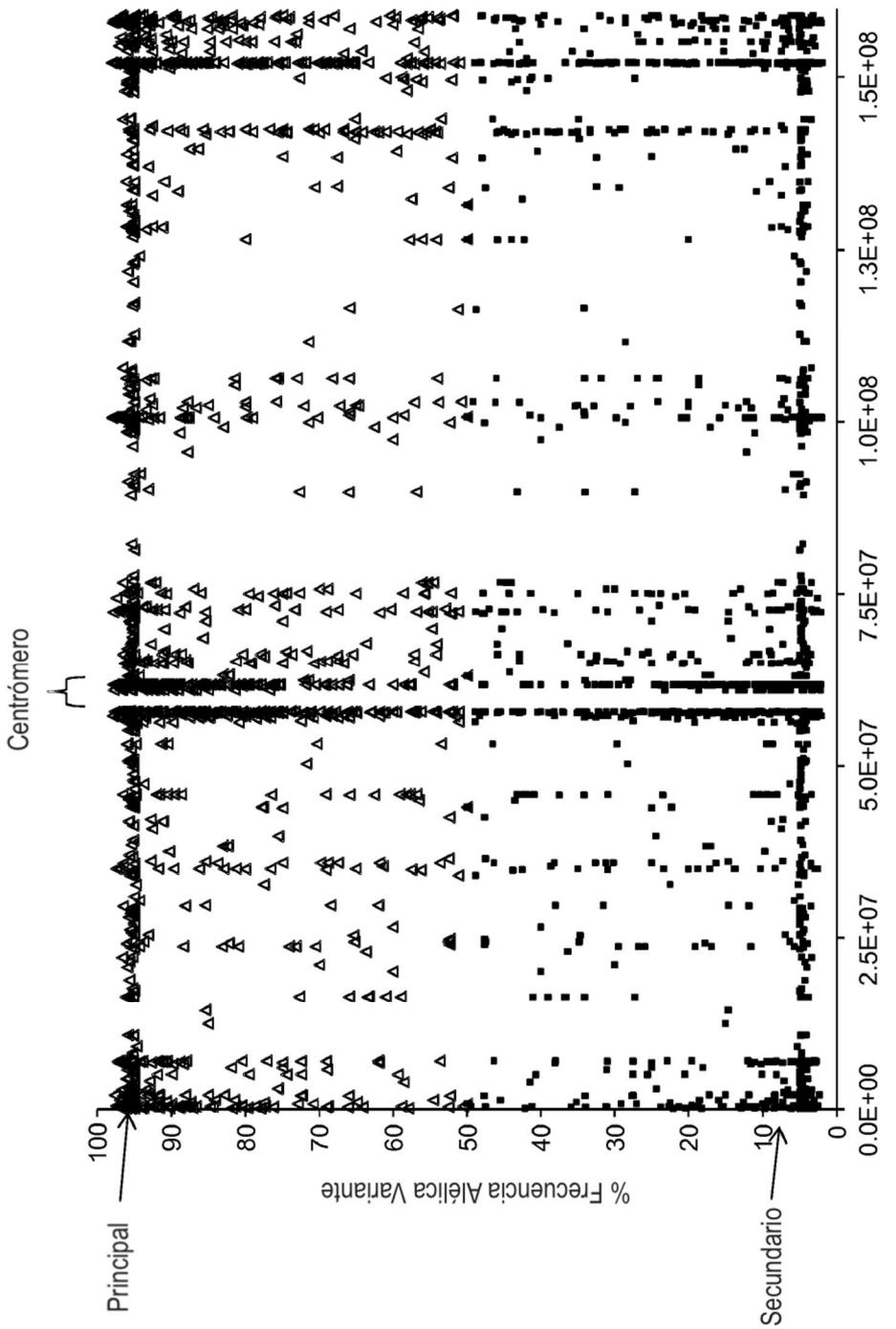
Centrómero
Principal
Secundario
% Frecuencia Alélica Variante
100
90
80
70
60
50
40
30
20
10
0
0.0E+00
2.5E+07
5.0E+07
7.5E+07
1.0E+08
1.3E+08
1.5E+08
Posición en el Cromosoma 7

FIG. 12B