(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 792 802**

(51) Int. CI.:

C12Q 1/6827 (2008.01)
C12Q 1/6883 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **23.07.2008 E 12173422 (2)**

(97) Fecha y número de publicación de la concesión europea: **04.03.2020 EP 2527471**

(54) Título: **Diagnóstico de cáncer utilizando la secuenciación genómica**

(30) Prioridad:

23.07.2007 US 951438 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
12.11.2020

(73) Titular/es:

THE CHINESE UNIVERSITY OF HONG KONG (100.0%)
Technology Licensing Office, Room 226, Pi Ch'iu Building
Shatin, Hong Kong, CN

(72) Inventor/es:

LO, YUK-MING DENNIS;
CHIU, ROSSA WAI KWUN y
CHAN, KWAN CHEE

(74) Agente/Representante:

PONS ARIÑO, Ángel

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

DESCRIPCIÓN

Diagnóstico de cáncer utilizando la secuenciación genómica

La presente invención se refiere, en líneas generales, al ensayo del diagnóstico de aneuploidía cromosómica fetal determinando desequilibrios entre diferentes secuencias de ácidos nucleicos, y más particularmente, a la identificación de trisomía 21 (síndrome de Down) y otras aneuploidías cromosómicas mediante ensayo de una muestra materna (por ejemplo, sangre).

La aneuploidía cromosómica fetal es el resultado de la presencia de una o varias dosis anómalas de un cromosoma o región cromosómica. Las dosis anómalas pueden ser anormalmente altas, por ejemplo, presencia de un cromosoma 21 o región cromosómica extra, en la trisomía 21, o anormalmente bajas, por ejemplo, ausencia de una copia del cromosoma X, en el síndrome de Turner.

Los métodos de diagnóstico prenatales convencionales de una aneuploidía cromosómica fetal, por ejemplo, trisomía 21, implican el muestreo de material fetal a través de procedimientos invasivos, tales como amniocentesis o muestreo de vellosidades coriónicas, que constituyen un riesgo finito de pérdida fetal. Para estratificar a las gestantes según el riesgo se han utilizado procedimientos no invasivos tales como exploración por ultrasonido y marcadores bioquímicos, antes de los procedimientos de diagnóstico invasivos definitivos. Sin embargo, estos métodos de exploración miden normalmente fenómenos epigenéticos asociados con aneuploidías cromosómicas, por ejemplo, trisomía 21, en lugar de anomalías cromosómicas principales, y por tanto tienen una precisión de diagnóstico limitada y otras desventajas, tales como estar muy influenciadas por la edad gestacional.

El descubrimiento en 1977 de ADN fetal libre circulando en el plasma materno ofrece nuevas posibilidades al diagnóstico prenatal no invasivo (Lo, YMD y Chiu, RWK 2007 *Nat Rev Genet* 8, 71-77). A pesar de que este método se ha aplicado fácilmente al diagnóstico prenatal de trastornos ligados al sexo (Costa, JM et al. 2002 *N Engl J Med* 346, 1502) y a determinados trastornos genéticos sencillos (LO, YMD et al. 1998 *N Engl J Med* 339, 1734-1738), su aplicación a la detección prenatal de aneuploidías cromosómicas fetales ha sido un reto considerable (Lo, YMD y Chiu, RWK 2007, citado anteriormente). En primer lugar, los ácidos nucleicos fetales co-existen en el plasma materno con un alto fondo de ácidos nucleicos de origen materno que, con frecuencia, pueden interferir con el análisis de ácidos nucleicos fetales (Lo, YMD et al. 1998 *Am J Hum Genet* 62, 768-775). En segundo lugar, los ácidos nucleicos fetales circulan en el plasma materno predominantemente en una forma libre, dificultando la obtención de información sobre la dosificación de genes o cromosomas dentro del genoma fetal.

Recientemente se han realizado desarrollos significativos que superan estos retos (Benachi, A & Costa, JM 2007 *Lancet* 369, 440-442). Una estrategia detecta ácidos nucleicos específicos fetales en el plasma materno, superando así el problema de la interferencia con el fondo materno (Lo, YMD y Chiu, RWK 2007, citado anteriormente). La dosificación del cromosoma 21 se dedujo a partir de las proporciones de alelos polimórficos en las moléculas de ADN/ARN derivadas de placenta. Sin embargo, este método es menos preciso cuando las muestras contienen menos cantidad de ácido nucleico diana y solo pueden aplicarse a fetos que son heterocigotos para polimorfismos diana, lo cual es solo un subconjunto de la población si se usa un polimorfismo.

Dhallan et al (Dhallan, R, *et al.* 2007, citado anteriormente Dhallan, R, *et al.* 2007 *Lancet* 369, 474-481) describen una estrategia alternativa de enriquecimiento de la proporción de ADN fetal circulante añadiendo formaldehído al plasma materno. La proporción de secuencias del cromosoma 21 aportadas por el feto en el plasma materno se determinó evaluando la razón de alelos específicos fetales heredados por vía paterna con respecto a alelos específicos no fetales para polimorfismos mononucleotídicos, (SNP, *Single Nucleotide Polymorphisms)* en el cromosoma 21. De manera similar, se calcularon las razones SNP para un cromosoma de referencia. Después, se dedujo un desequilibrio del cromosoma fetal 21 detectando una diferencia estadísticamente significativa entre las razones SNP para el cromosoma 21 y las del cromosoma de referencia, en la que el significante se definió utilizando un valor fijo de p s 0,05. Para garantizar una alta cobertura de la población, se definieron más de 500 SNP por cromosoma. Sin embargo, ha habido controversias con respecto a la eficacia del formaldehído para enriquecer el ADN fetal a una alta proporción (Chung, GTY, et al. 2005 *Clin Chem.* 51, 655-658) y por tanto la reproducibilidad del método requiere evaluarse adicionalmente. Además, como cada feto y madre sería informativo para un número diferente de SNP para cada cromosoma, el poder del ensayo estadístico para la comparación de razones SNP variaría en cada caso (Lo, YMD y Chiu, RWK. 2007 *Lancet* 369, 1997). Además, dado que estas estrategias dependen de la detección de polimorfismos genéticos, se limitan a fetos heterocigotos para estos polimorfismos.

La utilización de la reacción en cadena de la polimerasa (PCR) y la cuantificación de ADN de un locus de cromosoma 21 y de un locus de referencia en cultivos de amniocitos, obtenidos de fetos con trisomía 21 y euploides, Zimmermann et al (2002 Clin Chem. 48, 362-363) pudo diferenciar los dos grupos de fetos basándose en el aumento de 1,5 veces en las secuencias de ADN del cromosoma 21 en el primero de ellos. Dado que una diferencia de 2 veces en la concentración del molde de ADN constituye una diferencia de un solo ciclo de umbral (Ct), la discriminación de una diferencia de 1,5 veces ha sido el límite de la PCR en tiempo real convencional. Para conseguir grados de discriminación cuantitativa más refinados, se requieren estrategias alternativas.

La PCR digital se ha desarrollado para la detección de la desviación de la razón alélica en muestras de ácidos nucleicos (Chang, HW et al. 2002 J Natl Cancer Inst. 94, 1697-1703). La PCR digital es una amplificación basada en técnicas de análisis de ácidos nucleicos que requiere la distribución de un espécimen que contenga ácidos nucleicos en una multitud de muestras distintas donde cada muestra contiene un promedio no mayor de aproximadamente una secuencia diana por muestra. Las dianas específicas de ácidos nucleicos se amplifican con cebadores específicos de secuencia para generar amplicones específicos por PCR digital. Los loci de ácidos nucleicos diana y las especies de o un panel de cebadores específicos de secuencia a incluir en las reacciones se determinan o se seleccionan antes del análisis de ácidos nucleicos.

Clínicamente, se ha demostrado que es útil para la detección de pérdida de heterocigosidad (LOH, *Lost Of Heterozygosity)* en muestras de ADN tumoral (Zhou, W. et al. 2002 Lancet 359, 219-225). Para el análisis de resultados de PCR digital, estudios previos han adoptado un ensayo de razón de probabilidad secuencial (SPRT, Sequential Probability Ratio Testing) para clasificar los resultados experimentales ya que sugieren la presencia de LOH en una muestra o no (El Karoui et al. 2006 Star Med 25, 3124 a 3133).

En métodos utilizados en los estudios previos, la cantidad de datos obtenidos de la PCR digital es muy pequeña. Por lo tanto, la precisión puede verse comprometida debido a la poca cantidad de puntos de datos y de fluctuaciones estadísticas típicas.

Por tanto, es deseable que los ensayos no invasivos tengan una alta sensibilidad y especificidad para minimizar negativos falsos y positivos falsos, respectivamente. Sin embargo, el ADN fetal está presente a una baja concentración absoluta y representa una pequeña parte de todas las secuencias de ADN en el plasma y suero materno. Por lo tanto, también es deseable tener métodos que permitan la detección no invasiva de aneuploidía cromosómica fetal maximizando la cantidad de información genética que podría deducirse de la cantidad limitada de ácidos nucleicos fetales que existe como una población más pequeña en una muestra biológica que contenga ácidos nucleicos maternos de fondo.

De acuerdo con la presente invención, se proporciona un método implementado por ordenador de análisis de una muestra biológica que incluye moléculas de ADN sin células que provienen de células no malignas y potencialmente de células tumorales asociadas con cáncer, comprendiendo el método: recibir etiquetas secuenciadas obtenidas de una secuenciación aleatorizada de moléculas de ADN contenidas en la muestra biológica, siendo la muestra biológica de plasma, suero, saliva u orina; alinear una pluralidad de las etiquetas secuenciadas a un genoma humano de referencia; determinar una primera cantidad de etiquetas secuenciadas como la alineación con una primera región cromosómica que es parte de un primer cromosoma; determinar una segunda cantidad de etiquetas secuenciadas identificadas como la alineación con una o más segundas regiones cromosómicas; determinar un primer parámetro a partir de la primera cantidad y de la segunda cantidad, en donde el parámetro incluye una proporción entre la primera cantidad y la segunda cantidad, en donde una anomalía en el primer parámetro para la primera región cromosómica se asocia con cáncer: y comparar el primer parámetro con uno o más valores umbrales para determinar una clasificación de si las primeras regiones cromosómica presentan una deleción y/o una amplificación asociada con el cáncer.

Puede obtenerse un mejor entendimiento de la naturaleza y ventajas de la presente invención con referencia a la siguiente descripción detallada y a los dibujos adjuntos.

La FIG. 1 es un flujograma de un método 100 para realizar el diagnóstico prenatal de una aneuploidía cromosómica fetal en una muestra biológica obtenida de una gestante de acuerdo con una realización de la presente invención.

La FIG. 2 es un flujograma de un método 200 para realizar el diagnóstico prenatal de una aneuploidía cromosómica fetal usando secuenciación aleatoria de acuerdo con una realización de la presente invención.

La FIG. 3 A muestra una representación gráfica del porcentaje de secuencias del cromosoma 21 en muestras de plasma materno que implican trisomía 21 o fetos euploides de acuerdo con una realización de la presente invención.

La FIG. 3B muestra una correlación entre concentraciones de ADN fetal fraccional de plasma materno determinada por secuenciación masiva en paralelo y PCR digital de microfluidos de acuerdo con una realización de la presente invención.

La FIG. 4A muestra una representación gráfica del porcentaje de secuencias alineadas por cromosoma de acuerdo con una realización de la presente invención.

La FIG. 4B muestra una representación gráfica de diferencia (%) en porcentaje por cromosoma entre el caso de trisomía 21 y el caso euploide mostrado en la FIG. 4A.

La FIG. 5 muestra una correlación entre el grado de sobre-representación en secuencias del cromosoma 21 y las

concentraciones de ADN fetal fraccional en plasma materno que implica fetos con trisomía 21 de acuerdo con una realización de la presente invención.

La FIG. 6 muestra una tabla de una parte del genoma humano que se analizó de acuerdo con una realización de la presente invención. T21 indica una muestra obtenida de una gestante que lleva un feto con trisomía 21.

La FIG. 7 muestra una tabla de diversas secuencias necesarias para diferenciar fetos euploides de trisomía 21 de acuerdo con una realización de la presente invención.

La FIG. 8A muestra una tabla de las diez primeras posiciones de partida de etiquetas secuenciadas alineadas con el cromosoma 21 de acuerdo con una realización de la presente invención.

La FIG. 8B muestra una tabla de las diez primeras posiciones de partida de etiquetas secuenciadas alineadas con el cromosoma 22 de acuerdo con una realización de la presente invención.

La FIG. 9 muestra un diagrama de bloque de un equipo informático a modo de ejemplo que puede usarse con el sistema y los métodos de acuerdo con realizaciones de la presente invención.

La expresión “muestra biológica”, como se usa en el presente documento, se refiere a cualquier muestra extraída de un sujeto {por ejemplo, un ser humano, tal como una gestante) y contiene una o más moléculas de ácido nucleico de interés.

La expresión *“ácido nucleico"* o *“polinucleótido"* se refiere a un ácido desoxirribonucleico (ADN) o ácido ribonucleico (ARN) y a un polímero de los mismos en forma mono- o bicatenaria. Salvo que se limite específicamente, la expresión incluye ácidos nucleicos que contienen análogos conocidos de nucleótidos naturales que tienen propiedades de unión similares a las del ácido nucleico de referencia y que se metabolizan de manera similar a la de los nucleótidos de origen natural. Salvo que se indique de otra manera, una secuencia de ácido nucleico particular también abarca implícitamente variantes conservativamente modificadas de las mismas (por ejemplo, sustituciones de codones degenerados), alelos, ortólogos, SNP y secuencias complementarias así como la secuencia explícitamente indicada. Específicamente, las sustituciones de codones degenerados pueden conseguirse generando secuencias en las que la tercera posición de uno o más codones seleccionados (o todos) se sustituye con restos de bases y/o desoxiinosina mixtos (Batzer *et al., Nucleic Acid Res.* 19:5081 (1991); Ohtsuka *et al, J. Biol. Chem.* 260:2605-2608 (1985); y Rossolini *et al, Mol Cell Probes* 8:91-98 (1994)). El término ácido nucleico se usa indistintamente con gen, ADNc, ARNm, ARN pequeño no codificante, micro ARN (miRNA), ARN de interacción con Piwi y ARN corto en horquilla (shRNA) codificado por un gen o locus.

El término *“gen"* significa el segmento de ADN implicado en la producción de una cadena polipeptídica. Este puede incluir regiones anteriores y posteriores a la región codificante (líder y remolque) así como a secuencias de intervención (intrones) entre segmentos codificantes individuales (exones).

El término *“reacción",* como se usa en el presente documento, se refiere a cualquier proceso que implique una acción química, enzimática o física que sea indicativa de la presencia o ausencia de una secuencia de polinucleótidos de interés particular. Un ejemplo de una *“reacción"* es una reacción de amplificación, tal como una reacción en cadena de la polimerasa (PCR). Otro ejemplo de una *“reacción"* es una reacción de secuenciación, por síntesis o por ligamiento. Una *“reacción informativa"* es una que indica la presencia de una o más secuencias de polinucleótidos de interés particular, y en un caso en la que solo está presente una secuencia de interés. El término “bien" como se usa en el presente documento se refiere a una reacción en un lugar predeterminado dentro de una estructura confinada, por ejemplo, un vial, célula o cámara bien conformados en una matriz PCR.

La expresión *“secuencia de ácido nucleico clínicamente relevante",* como se usa en el presente documento, puede referirse a una secuencia de polinucleótidos correspondiente a un segmento de una secuencia genómica más grande cuyo posible desequilibrio va a someterse a ensayo o a la propia secuencia genómica más grande. Un ejemplo es la secuencia del cromosoma 21. Otros ejemplos incluyen el cromosoma 18, 13, X e Y. Incluso otros ejemplos incluyen secuencias genéticas mutadas o polimorfismos genéticos o variaciones en el número de copias que un feto puede heredar de uno de sus padres o de ambos. Incluso otros ejemplos incluyen secuencias que están mutadas, delecionadas o amplificadas en un tumor maligno, por ejemplo, secuencias en las que se produce pérdida de heterocigosidad o duplicación génica. En algunas realizaciones, pueden usarse secuencias múltiples de ácido nucleico clínicamente relevantes, o marcadores equivalentemente múltiples de la secuencia de ácido nucleico clínicamente relevante para proporcionar datos para detectar el desequilibrio. Por ejemplo, pueden usarse datos de cinco secuencias no consecutivas en el cromosoma 21 de una manera aditiva para la determinación de un posible desequilibrio del cromosoma 21, reduciendo eficazmente la necesidad de que el volumen de la muestra sea de 1/5.

La expresión *“secuencia de ácido nucleico de fondo",* como se usa en el presente documento, se refiere a una secuencia de ácido nucleico cuya razón normal con respecto a la secuencia de ácido nucleico clínicamente relevante se conoce, por ejemplo, una razón de 1 a 1. Como un ejemplo, la secuencia de ácido nucleico de fondo y la secuencia de ácido nucleico clínicamente relevante son dos alelos del mismo cromosoma que se diferencian debido

a la heterocigosidad. En otro ejemplo, la secuencia de ácido nucleico de fondo es un alelo que es heterocigoto con respecto a otro alelo que es la secuencia de ácido nucleico clínicamente relevante. Además, algo de cada una de la secuencia de ácido nucleico de fondo y la secuencia de ácido nucleico clínicamente relevante puede proceder de diferentes individuos.

La expresión *"secuencia de ácido nucleico de referencia"* como se usa en el presente documento se refiere a una secuencia de ácido nucleico cuya concentración promedio por reacción se conoce o se ha medido de manera equivalente.

La expresión *"secuencia de ácido nucleico sobre-representada"* como se usa en el presente documento se refiere a la secuencia de ácido nucleico entre dos secuencias de interés (por ejemplo una secuencia clínicamente relevante y una secuencia de fondo) que es más abundante que la otra secuencia en una muestra biológica.

La expresión *"basándose en"* como se usa en el presente documento significa "basándose al menos en parte en" y se refiere a un valor (o resultado) que se usa en la determinación de otro valor, tal como se produce en la relación de una entrada de un método y la salida de ese método. El término "deriva" como se usa en el presente documento también se refiere a la relación de una entrada de un método y la salida de ese método, tal como se produce cuando la derivación es el cálculo de una fórmula.

La expresión "*datos cuantitativos*" como se usa en el presente documento significa datos que se obtienen de una o más reacciones y que proporcionan uno o más valores numéricos. Por ejemplo, el número de pocillos que presenta un marcador fluorescente para una secuencia particular sería un dato cuantitativo.

El término *"parámetro"* como se usa en el presente documento significa un valor numérico que caracteriza un conjunto de datos cuantitativo y/o una relación numérica entre conjuntos de datos cuantitativos. Por ejemplo, un parámetro es una razón (o función de una razón) entre una primera cantidad de una primera secuencia de ácido nucleico y una segunda cantidad de una segunda secuencia de ácido nucleico.

La expresión *"valor de límite"* como se usa en el presente documento significa un valor numérico cuyo valor se usa para arbitrar entre dos o más estados (por ejemplo, enfermo y no enfermo) de clasificación de una muestra biológica. Por ejemplo, si un parámetro es mayor que el valor de límite, se realiza una primera clasificación del dato cuantitativo (por ejemplo, estado enfermo); o si el parámetro es menor que el valor de límite, se realiza una clasificación diferente de los datos cuantitativos (por ejemplo, estado no enfermo).

La expresión "desequilibrio" como se usa en el presente documento significa cualquier desviación significativa definida por al menos un valor de límite en una cantidad de la secuencia de ácido nucleico clínicamente relevante de una cantidad de referencia. Por ejemplo, la cantidad de referencia podría ser una razón de 3/5 y por tanto podría producirse un desequilibrio si la razón medida es de 1:1.

La expresión "aneuploidía cromosómica", como se usa en el presente documento, significa una variación en la cantidad cuantitativa de un cromosoma a partir de la de un genoma diploide. La variación puede ser una ganancia o una pérdida. Esto puede implicar todo un cromosoma o una región de un cromosoma.

La expresión "secuenciación aleatoria", como se usa en el presente documento, se refiere a una secuenciación mediante la cual los fragmentos de ácido nucleico secuenciados no se han identificado o seleccionado específicamente antes del procedimiento de secuenciación. No se precisan cebadores específicos de secuencia para los locus de genes específicos diana. Los grupos de ácidos nucleicos secuenciados varían entre muestras e incluso entre análisis para la misma muestra. Las identidades de los ácidos nucleicos secuenciados solo se revelan a partir del resultado de secuenciación generado. En algunas realizaciones de la presente invención, la secuenciación aleatoria puede comenzar con procedimientos para enriquecer una muestra biológica con poblaciones particulares de moléculas de ácido nucleico que comparten determinadas características comunes. En una realización, cada uno de los fragmentos en la muestra biológica tiene una misma probabilidad de ser secuenciado.

La expresión *"fracción del genoma humano"* o *"parte del genoma humano"* como se usa en el presente documento se refiere a menos del 100 % de las secuencias de nucleótidos en el genoma humano que comprenden unos 3 billones de pares de bases de nucleótidos. En el contexto de secuenciación, esto se refiere a menos de una cobertura de 1 vez de las secuencias de nucleótidos en el genoma humano. El término puede expresarse como un porcentaje o número absoluto de nucleótidos/pares de bases. Como un ejemplo de uso, el término puede usarse para hacer referencia a la cantidad real de secuenciación realizada. Las realizaciones pueden determinar el valor mínimo requerido para la fracción secuenciada del genoma humano para obtener un diagnóstico preciso. Como otro ejemplo de su uso, el término puede referirse a la cantidad de datos secuenciados usados para derivar un parámetro o cantidad para una clasificación de enfermedad.

La expresión *"etiqueta secuenciada",* como se usa en el presente documento, se refiere a una cadena de nucleótidos secuenciada desde cualquier parte o toda una molécula de ácido nucleico. Por ejemplo, una etiqueta secuenciada puede ser una cadena corta de nucleótidos secuenciada a partir de un fragmento de ácido nucleico,

una cadena corta de nucleótidos en ambos extremos de un fragmento de ácido nucleico, o la secuenciación de todo el fragmento de ácido nucleico que existe en la muestra biológica. Un fragmento de ácido nucleico es cualquier parte de una molécula de ácido nucleico más grande. Un fragmento (por ejemplo, un gen) puede existir independientemente (es decir, no conectado) con respecto a las otras partes de la molécula de ácido nucleico más grande.

Descripción detallada

Las realizaciones de la presente invención proporcionan métodos, sistemas y aparatos para determinar si existe un aumento o una disminución (estado enfermo) de una región cromosómica clínicamente relevante en comparación con un estado no enfermo. Esta determinación puede realizarse usando un parámetro de una cantidad de una región cromosómica clínicamente relevante en relación con otras regiones cromosómicas no clínicamente relevantes (regiones de fondo) dentro de una muestra biológica. Las moléculas de ácido nucleico de la muestra biológica se secuencian, de tal manera que una fracción del genoma se secuencia, y la cantidad se puede determinar a partir de los resultados de la secuenciación. Uno o más valores de límite se seleccionan para determinar si existe un cambio en comparación con una cantidad de referencia (por ejemplo un desequilibrio), por ejemplo, con respecto a la razón de cantidades de dos regiones (o conjuntos de regiones) cromosómicas

El cambio detectado en la cantidad de referencia puede ser cualquier desviación (positiva o negativa) en la relación de la secuencia de ácido nucleico clínicamente relevante con las otras secuencias no clínicamente relevantes. Por lo tanto, el estado de referencia puede ser cualquier razón u otra cantidad (por ejemplo distinta de una correspondencia 1-1) y un estado medido que significa un cambio puede ser cualquier razón u otra cantidad que difiera de la cantidad de referencia determinada por uno o más valores de límite.

La región cromosómica clínicamente relevante (denominada también secuencia de ácido nucleico clínicamente relevante) y la secuencia de ácido nucleico de fondo pueden proceder de un primer tipo de células y de uno o más segundos tipos de células. Por ejemplo, en una muestra biológica, tal como plasma materno, están presentes secuencias de ácido nucleico fetal, originadas de células fetales/placentarias, que contiene un fondo de secuencias de ácido nucleico materno originado de células maternas. En una realización, el valor de límite se determina basándose al menos en parte en un porcentaje del primer tipo de células en una muestra biológica. Obsérvese que el porcentaje de secuencias fetales en una muestra puede determinarse mediante cualquier loci derivado del feto y no limitarse a medir las secuencias de ácido nucleico clínicamente relevantes. En otra realización, el valor de límite se determina, al menos en parte, sobre el porcentaje de secuencias tumorales en una muestra biológica, tal como plasma, suero, saliva u orina, que contiene un fondo de secuencias de ácido nucleico derivadas de células no malignas dentro del organismo.

I. MÉTODO GENERAL

La FIG. 1 es un diagrama de flujo de un método 100 para realizar el diagnóstico prenatal de una aneuploidía cromosómica fetal en una muestra biológica obtenida de una gestante de acuerdo con una realización de la presente invención.

En la etapa 110, se recibe una muestra biológica de la gestante. La muestra biológica puede ser plasma, orina, suero, o cualquier otra muestra adecuada. La muestra contiene moléculas de ácido nucleico del feto y de la gestante. Por ejemplo, las moléculas de ácido nucleico pueden ser fragmentos de cromosomas.

En la etapa 120, al menos una parte de una pluralidad de las moléculas de ácido nucleico contenida en la muestra biológica se secuencia. La parte secuenciada representa una fracción del genoma humano. En una realización, las moléculas de ácido nucleico son fragmentos de cromosomas respectivos. Puede secuenciarse un extremo (por ejemplo, 35 pares de bases (pb)), ambos extremos, o todo el fragmento. Puede secuenciarse todas las moléculas de ácido nucleico en la muestra o solo un subconjunto. Este subconjunto se puede seleccionar aleatoriamente, como se describirá más adelante con más detalle.

En una realización, la secuenciación se realiza usando secuenciación masiva en paralelo. La secuenciación masiva en paralelo, tal como la que puede obtenerse en la plataforma 454 (Roche) (Margulies, M. *et al.* 2005 *Nature* 437, 376-380), Illumina Analyzer Genome (o plataforma Solexa) o SOLiD System (Applied Biosystems) o la tecnología de secuenciación de ADN Helicos True Single Molecule (Harris TD et al. 2008 Science, 320, 106-109), la tecnología de molécula sencilla, en tiempo real (SMRT™) de Biosciences Pacific, y secuenciación por nanoporos (Soni GV y Meller A. 2007 Clin Chem 53: 1996-2001), permiten la secuenciación de muchas moléculas de ácido nucleico aisladas de un espécimen a altos grados de formación de complejos múltiples de una manera paralela (Dear Brief Funct Genomic Proteomic 2003; 1: 397-416). Cada una de estas plataformas secuencian moléculas sencillas clonalmente expandidas o incluso no amplificadas de fragmentos de ácido nucleico.

Como a partir de cada muestra en cada proceso se genera un alto número de lecturas de secuenciación, en el orden de cientos de miles a millones, o incluso posiblemente, de cientos de millones o billones, las lecturas secuenciadas resultantes forman un perfil representativo de la mezcla de especies de ácido nucleico en el espécimen original. Por

ejemplo, el haplotipo, trascriptoma y perfiles de metilación de las lecturas secuenciadas se asemejan a las del espécimen original (Brenner et al Nat. Biotech 2000; 18: 630-634; Taylor et al Cancer Res. 2007; 67: 85118518). Debido al gran muestreo de secuencias de cada espécimen, el número de secuencias idénticas, tales como las generadas a partir de la secuenciación de un grupo de ácido nucleico a diversas veces de cobertura o alta redundancia, es también una buena representación cuantitativa del recuento de una especie de ácido nucleico particular o locus en la muestra original.

En la etapa 130, basándose en la secuenciación (por ejemplo, datos procedentes de la secuenciación), se determina una primera cantidad de un primer cromosoma (por ejemplo, el cromosoma clínicamente relevante). La primera cantidad se determina a partir de secuencias identificadas como originadas a partir del primer cromosoma. Por ejemplo, después puede utilizarse un procedimiento bioinformático para localizar cada una de estas secuencias de ADN en el genoma humano. Es posible que una proporción de dichas secuencias se descarte de un análisis posterior debido a que están presentes en las regiones repetidas del genoma humano, o en regiones sometidas a variaciones inter-individuales, por ejemplo, variaciones en el número de copias. Por tanto, puede determinarse una cantidad del cromosoma de interés y uno o más cromosomas distintos.

En la etapa 140, basándose en la secuenciación, se determina una segunda cantidad de uno o más segundos cromosomas a partir de secuencias identificadas como originadas a partir de uno de los segundos cromosomas. En una realización, los segundos cromosomas son todos los otros cromosomas además del primero (es decir, el sometido a ensayo). En otra realización, el segundo cromosoma es precisamente un solo cromosoma distinto.

Existen diversas maneras de determinar las cantidades de los cromosomas, incluyendo, pero sin limitación, recuento del número de etiquetas secuenciadas, del número de nucleótidos (pares de bases) secuenciados o las longitudes acumuladas de nucleótidos secuenciados (pares de bases) que se originan a partir de uno o más cromosomas o regiones cromosómicas particulares.

En otra realización, pueden imponerse normas sobre los resultados de la secuenciación para determinar los recuentos conseguidos. En un aspecto, puede obtenerse una cantidad basándose en una proporción del resultado secuenciado. Por ejemplo, después del análisis bioinformático, podría seleccionarse el resultado de secuenciación correspondiente a fragmentos de ácido nucleico de un intervalo de tamaño especificado. Son ejemplos de los intervalos de tamaño aproximadamente < 300 pb, < 200 pb o < 100 pb.

En la etapa 150, se determina un parámetro a partir de la primera y segunda cantidad. El parámetro puede ser, por ejemplo, una simple razón de la primera cantidad con respecto a la segunda cantidad, o de la primera cantidad con respecto a la segunda cantidad más la primera cantidad. En un aspecto, cada cantidad podría ser un argumento con respecto a una función o funciones distintas, en el que después una razón puede sustraerse de estas funciones distintas. Un experto en la técnica apreciará el número de diferentes parámetros adecuados.

En una realización, un parámetro (por ejemplo, una representación fraccional) de un cromosoma posiblemente implicado en una aneuploidía cromosómica, por ejemplo, cromosoma 21 o cromosoma 18 o cromosoma 13, puede después calcularse a partir de los resultados del procedimiento bioinformático. La representación fraccional puede obtenerse basándose en una cantidad de todas las secuencias (por ejemplo, alguna medición de todos los cromosomas incluyendo el cromosoma clínicamente relevante) o un subconjunto particular de cromosomas (por ejemplo, precisamente el otro cromosoma sometido a ensayo).

En la etapa 150, el parámetro se compara con uno o más valores de límite. Los valores de límite pueden determinarse a partir de cualquiera de las diversas formas adecuadas. Dichas formas incluyen el método de probabilidad Bayesiano, el ensayo de razón de probabilidad secuencial (SPRT), el descubrimiento de falsos, el intervalo de confianza, las características operativas receptoras (ROC).

En una realización, el parámetro (por ejemplo, la representación fraccional del cromosoma clínicamente relevante) se compara después con un intervalo de referencia establecido en gestantes que llevan fetos normales (es decir, euploides). Es posible que, en algunas variantes del procedimiento, el intervalo de referencia (es decir, los valores de límite) se ajusten de acuerdo con la concentración fraccional del ADN fetal (f) en una muestra de plasma materno particular. El valor de f puede determinarse a partir del conjunto de datos de secuenciación, utilizando, por ejemplo, secuencias mapeables con respecto al cromosoma Y si el feto es de sexo masculino. El valor de f también puede determinarse en un análisis distinto, utilizando, por ejemplo, marcadores epigenéticos fetales (Chan KCA et al 2006 Clin Chem. 52, 2211-8) o a partir de análisis de polimorfismos mononucleotídicos.

En la etapa 160, basándose en la comparación, se determina una clasificación de si existe una aneuploidía cromosómica fetal para el primer cromosoma. En una realización, la clasificación es un sí o no definitivo. En otra realización, una clasificación puede ser no clasificable o dudosa. En otra realización, la clasificación puede ser una puntuación que debe interpretar, por ejemplo, un médico, a última instancia.

II SECUENCIACIÓN, ALINEAMIENTO Y CANTIDADES DE DETERMINACIÓN

Como se ha mencionado anteriormente, solo se secuencia una fracción del genoma. En un aspecto, incluso cuando se secuencia un grupo de ácidos nucleicos en un espécimen a < 100 % de cobertura genómica en lugar de a diversas veces de cobertura, y entre la proporción de moléculas de ácido nucleico capturadas, la mayoría de cada especie de ácido nucleico solo se secuencia una vez. Además, el desequilibrio de dosificación de un cromosoma o regiones cromosómicas particulares puede determinarse cuantitativamente. En otras palabras, el desequilibrio de dosificación del cromosoma o de las regiones cromosómicas se deduce del porcentaje de representación de dicho locus entre otras etiquetas mapeables secuenciadas del espécimen.

Esto discrepa de situaciones en las que el mismo grupo de ácidos nucleicos se secuencia varias veces para conseguir alta redundancia o diversas veces de cobertura mediante lo cual cada especie de ácido nucleico se secuencia varias veces En dichas situaciones, el número de veces que una especie de ácido nucleico particular se ha secuenciado con respecto al de otra especie de ácido nucleico se correlaciona con concentraciones relativas en la muestra original. El coste de secuenciación aumenta con el número de veces de cobertura necesarios para conseguir una representación exacta de la especie de ácido nucleico.

En un ejemplo, una proporción de dichas secuencias sería a partir del cromosoma implicado en una aneuploidía tal como el cromosoma 21 en este ejemplo ilustrativo. Incluso otras secuencias a partir de dicho ejercicio de secuenciación derivarían de otros cromosomas. Teniendo en cuenta el tamaño relativo del cromosoma 21 en comparación con otros cromosomas, podría obtenerse una frecuencia normalizada, dentro de un intervalo de referencia, de secuencias específicas del cromosoma 21 a partir de dicho ejercicio de secuenciación. Si el feto tiene trisomía 21, entonces la frecuencia normalizada de las secuencias derivadas del cromosoma 21, a partir de dicho ejercicio de secuenciación, aumentará, permitiendo por tanto la detección de la trisomía 21. El grado de cambio en la frecuencia normalizada dependerá de la concentración fraccional de los ácidos nucleicos fetales en la muestra analizada.

En una realización, se utilizó el Illumina Genome Analyzer para secuenciar un solo extremo de ADN genómico humano y muestras de ADN de plasma humano. Las secuencias del Illumina Genome Analyzer expanden clonalmente moléculas de ADN sencillas capturadas en una superficie sólida denominada célula de flujo. Cada célula de flujo tiene 8 carriles para la secuenciación de 8 especímenes individuales o conjunto de especímenes. Cada carril puede generar ~ 200 Mb de secuencias que es solo una fracción de los 3 billones de pares de bases de secuencias en el genoma humano. Cada muestra de ADN genómico o ADN plasmático se secuenció usando un carril de una célula de flujo. Las etiquetas de secuencias cortas generadas se alinearon con la secuencia del genoma de referencia humano y se anotó el origen cromosómico. El número total de etiquetas secuenciadas individuales alineadas con cada cromosoma se tabuló y se comparó con el tamaño relativo de cada cromosoma como se espera a partir del genoma humano de referencia o especímenes representativos sin enfermedad. Después, se identificaron las ganancias o pérdidas cromosómicas.

La estrategia descrita es solo un ejemplo de la estrategia de dosificación de gen/cromosoma descrita en el presente documento. Como alternativa, podría realizarse secuenciación de extremos emparejados. En lugar de comparar la longitud de los fragmentos secuenciados a partir de lo esperado en el genoma de referencia, como describen Campbell et al (Nat. Genet 2008; 40: 722-729), se contó el número de etiquetas secuenciadas alineadas y se clasificó de acuerdo con la localización cromosómica. Las ganancias o pérdidas de regiones cromosómicas o cromosomas completos se determinaron comparando los recuentos de etiquetas con el tamaño de cromosoma esperado en el genoma de referencia o con el de un espécimen representativo sin enfermedad. Como la secuenciación de extremos emparejados solo permite deducir el tamaño del fragmento de ácido nucleico original, un ejemplo es enfocarse en el recuento del número de etiquetas secuenciadas emparejadas correspondientes a los fragmentos de ácido nucleico de un tamaño específico, tal como < 300 pb, < 200 pb o < 100 pb.

En otra realización, la fracción del conjunto de ácido nucleico que se secuencia en un proceso se sub-selecciona adicionalmente antes de la secuenciación. Por ejemplo, podrían utilizarse técnicas basadas en hibridación, tales como, matriz de oligonucleótidos, para subseleccionar primero secuencias de ácido nucleico de determinados cromosomas, por ejemplo, un cromosoma posiblemente aneuploide y otro cromosoma (o cromosomas) no implicados en la aneuploidía ensayada. Otro ejemplo es que una determinada subpoblación de secuencias de ácido nucleico del conjunto de muestras se subseleccione o se enriquezca antes de la secuenciación. Por ejemplo, como se ha indicado anteriormente, se ha publicado que las moléculas de ADN fetal en plasma materno comprenden fragmentos más cortos que las moléculas de ADN de fondo materno (Chan et al Clin Chem 2004; 50: 88-92). Por tanto, para fraccionar las secuencias de ácido nucleico en la muestra de acuerdo con el tamaño molecular, pueden usarse uno o más métodos conocidos por los expertos en la técnica, por ejemplo, por electroforesis en gel o columnas de exclusión por tamaño o estrategias basadas en microfluidos. Incluso, como alternativa, en el ejemplo de analizar ADN fetal libre en plasma materno, la parte de ácido nucleico fetal podría enriquecerse mediante un método que suprima el fondo materno, tal como mediante la adición de formaldehído (Dhallan et al JAMA 2004; 291: 1114-9). En una realización, una parte o subconjunto del grupo de ácido nucleico preseleccionado se secuencia aleatoriamente.

En la presente solicitud podrían utilizarse, de manera similar, otras estrategias de secuenciación de moléculas sencillas, tales como la plataforma 454 de Roche, la plataforma SOLID de Applied Biosystems, la tecnología de

secuenciación de ADN de Helicos True Single Molecule, la tecnología de molécula sencilla, en tiempo real (SMRT™) de Biosciences Pacific, y la secuenciación por nanoporos.

III. DETERMINACIÓN DE CANTIDADES DE CROMOSOMAS A PARTIR DE RESULTADOS DE SECUENCIACIÓN

Después de una secuenciación masiva en paralelo, se realizó un análisis bioinformático para localizar el origen cromosómico de las etiquetas secuenciadas. Después de este procedimiento, las etiquetas identificadas como originadas a partir del cromosoma posiblemente aneuploide, es decir, el cromosoma 21 en este estudio, se compararon cuantitativamente con todas las etiquetas secuenciadas o con etiquetas que se originaban a partir de uno o más cromosomas no implicados en la aneuploidía. La relación entre el resultado de la secuenciación del cromosoma 21 y otros cromosomas que no son el cromosoma 21, para un espécimen de ensayo, se comparó con valores de límite deducidos con los métodos descritos en la sección anterior, para determinar si el espécimen se obtuvo de una gestante portadora de un feto euploide o con trisomía 21.

Diversas cantidades diferentes que incluyen, pero sin limitación, lo siguiente podrían deducirse de las etiquetas secuenciadas. Por ejemplo, las diversas etiquetas secuenciadas, es decir, recuento absoluto, alineadas con un cromosoma particular podría compararse con el recuento absoluto de etiquetas secuenciadas alineadas con otros cromosomas. Como alternativa, el recuento fraccional de la cantidad de etiquetas secuenciadas del cromosoma 21 con referencia a todas o algunas otras etiquetas secuenciadas podría compararse con la de uno o más cromosomas no aneuploides. En el presente experimento, dado que se secuenciaron 36 pb de cada fragmento de ADN, el número de nucleótidos secuenciado a partir de un cromosoma particular podría deducirse fácilmente de 36 pb multiplicado por el recuento de etiquetas secuenciadas.

Adicionalmente, cada espécimen de plasma materno se secuenció únicamente usando una célula de flujo que solo podría secuenciar una fracción del genoma humano, por estadística, la mayoría de especies de fragmento de ADN plasmático materno podrían solamente haberse secuenciado para generar un recuento de etiqueta secuenciada. En otras palabras, los fragmentos de ácido nucleico presentes en el espécimen de plasma materno se secuenciaron menos de 1 vez de cobertura. Por tanto, el número total de nucleótidos secuenciados para cualquier cromosoma particular correspondería principalmente a la cantidad, proporción o longitud de la parte de dicho cromosoma que se ha secuenciado. Así pues, la determinación cuantitativa de la representación del cromosoma posiblemente aneuploide podría derivar de una fracción del número o longitud equivalente de nucleótidos secuenciados de ese cromosoma con referencia a una cantidad derivada de manera similar de otros cromosomas.

IV. ENRIQUECIMIENTO CON GRUPOS DE ÁCIDOS NUCLEICOS PARA SECUENCIACIÓN

Como anteriormente se ha mencionado y establecido en la siguiente sección de ejemplos, solo se necesita secuenciar una parte del genoma humano para diferenciar casos de trisomía 21 de euploides. Por tanto, sería posible y rentable enriquecer el grupo de ácidos nucleicos a secuenciar antes de la secuenciación aleatoria de una fracción del grupo enriquecido. Por ejemplo, las moléculas de ADN fetales en el plasma materno comprenden fragmentos más cortos que las moléculas de ADN de fondo materno (Chan et al Clin Chem 2004; 50: 88-92). Por lo tanto, para fraccionar las secuencias de ácido nucleico en la muestra de acuerdo con el tamaño de la molécula, pueden usarse uno o más métodos conocidos por los expertos en la técnica, por ejemplo, mediante estrategias de electroforesis en gel o columnas de exclusión por tamaño o basadas en microfluidos.

Incluso, como alternativa, en el ejemplo de analizar ADN fetal libre en plasma materno, la parte de ácido nucleico fetal podría enriquecerse mediante un método que suprima el fondo materno, tal como por la adición de formaldehído (Dhallan et al JAMA 2004; 291: 1114-9). La proporción de secuencias derivadas fetales podría enriquecerse en el grupo de ácido nucleico comprendido por fragmentos más cortos. De acuerdo con la FIG. 7, el número de etiquetas secuenciadas necesario para diferenciar casos euploides de trisomía 21 podría reducirse a medida que aumenta la concentración de ADN fetal fraccional.

Como alternativa, secuencias originadas a partir de un cromosoma posiblemente aneuploide y uno o más cromosomas no implicados en la aneuploidía podrían enriquecerse mediante técnicas de hibridación, por ejemplo, sobre micromatrices de oligonucleótidos. Los grupos de ácidos nucleicos enriquecidos podrían después someterse a secuenciación aleatoria. Esto permitiría reducir costes de secuenciación.

V. SECUENCIACIÓN ALEATORIA

La FIG. 2 es un flujograma de un método 200 para realizar el diagnóstico prenatal de una aneuploidía cromosómica fetal usando secuenciación aleatoria de acuerdo con una realización de la presente invención. En un aspecto para la estrategia de secuenciación masiva en paralelo, pueden generarse datos representativos de todos los cromosomas al mismo tiempo. El origen de un fragmento particular no se selecciona con antelación. La secuenciación se realiza aleatoriamente y después puede realizarse una búsqueda en la base de datos para ver cuando se origina un fragmento particular. Esto difiere de situaciones en las que se amplifica un fragmento específico del cromosoma 21 y otro del cromosoma 1.

En la etapa 210, se recibe una muestra biológica de una gestante. En la etapa 220, el número N de secuencias a analizar se calcula para una precisión deseada. En una realización, primero se identifica un porcentaje de ADN fetal en la muestra biológica. Esto puede realizarse mediante cualquier medio adecuado bien conocido por un experto en la técnica. La identificación puede ser simplemente la lectura de un valor que se mide mediante otra entidad. En esta realización, el cálculo del número N de secuencias a analizar se basa en el porcentaje. Por ejemplo, el número de secuencias necesario para analizar podría aumentarse cuando el porcentaje de ADN fetal disminuye, y podría disminuir cuando el ADN fetal aumenta. El número N puede ser un número fijo o un número relativo, tal como un porcentaje. En otra realización, podría secuenciarse un número N que se sabe que es adecuado para un diagnóstico de enfermedad preciso. El número N podría ser constante incluso en gestantes con concentraciones de ADN fetal que están en el límite inferior del intervalo normal.

En la etapa 230, al menos N de una pluralidad de las moléculas de ácido nucleico contenidas en la muestra biológica se secuencian aleatoriamente. Una característica de esta estrategia descrita es que los ácidos nucleicos a secuenciar no se identifican o seleccionan específicamente antes del análisis de la muestra, es decir, secuenciación. Para la secuenciación, no se requieren cebadores específicos de secuencia para dirigir loci génicos específicos. Los conjuntos de ácidos nucleicos secuenciados varían de una muestra a otra e incluso de un análisis a otro para la misma muestra. Además, a partir de las descripciones indicadas a continuación (FIG. 6) la cantidad de resultados de secuenciación necesaria para el diagnóstico de casos podría variar entre los especímenes sometidos a ensayo y la población de referencia. Estos aspectos difieren notablemente con la mayoría de las estrategias de diagnóstico molecular, tales como las basadas en hibridación *in situ* por fluorescencia, PCR de fluorescencia cuantitativa, PCR cuantitativa en tiempo real, PCR digital, hibridación genómica comparativa, hibridación genómica comparativa de micromatriz, etc., en las que el loci génico a dirigir requiere determinación previa, siendo por tanto necesario el uso de conjuntos de cebadores o sondas específicos de locus o paneles de los mismos.

En una realización, la secuenciación aleatoria se realiza en fragmentos de ADN que están presentes en el plasma de una gestante, y se obtienen secuencias genómicas que originalmente podrían provenir del feto o de la madre. La secuenciación aleatoria implica el muestreo (secuenciación) de una parte aleatoria de las moléculas de ácido nucleico presentes en la muestra biológica. Como la secuenciación es aleatoria, en cada análisis puede secuenciarse un subconjunto (fracción) diferente de las moléculas de ácido nucleico (y por tanto el genoma). Las realizaciones se llevarán a cabo incluso cuando este subconjunto varíe de una muestra a otra y de un análisis a otro, lo que puede producirse incluso usando la misma muestra. Ejemplos de la fracción son aproximadamente el 0,1 %, 0. 5 %, 1 %, 5 %, 10 %, 20 % o 30 % del genoma. En otras realizaciones, la fracción es al menos uno cualquiera de estos valores.

Las etapas 240-270 restantes pueden continuar de una manera similar al método 100.

VI. SELECCIÓN POST-SECUENCIACIÓN DE GRUPOS DE ETIQUETAS SECUENCIADAS

Como se describe en los siguientes ejemplos II y III, un subconjunto de los datos secuenciados es suficiente para diferenciar los casos de trisomía 21 de casos euploides. El subconjunto de datos secuenciados podría ser la proporción de etiquetas secuenciadas que pasan determinados parámetros de calidad. Por ejemplo, en el ejemplo II, se utilizaron las etiquetas secuenciadas que están únicamente alineadas con el genoma humano de referencia oculto con repetición. Como alternativa, se puede secuenciar un grupo representativo de fragmentos de ácido nucleico de todos los cromosomas pero centrándose en la comparación entre datos relevantes con el cromosoma posiblemente aneuploide y datos relevantes con un número de cromosomas no aneuploides.

Incluso, como alternativa, durante el análisis post-secuenciación, podría sub-seleccionarse un subconjunto de los resultados de secuenciación que incluye etiquetas secuenciadas generadas de los fragmentos de ácido nucleico correspondientes a una ventana de tamaño específico en el espécimen original. Por ejemplo, usando el Illumina Genome Analyzer, podría usarse secuenciación de extremos emparejados que se refiere a la secuenciación de dos extremos de fragmentos de ácidos nucleico. Los datos secuenciados de cada extremo emparejado se alinean después con la secuencia genómica humana de referencia. Después, podría deducirse la distancia o el número de nucleótidos que se extiende entre los dos extremos. La longitud completa del fragmento de ácido nucleico original también podría deducirse. Como alternativa, las plataformas de secuenciación, tales como la plataforma 454 y posiblemente algunas de las técnicas de secuenciación de una molécula sencilla, pueden secuenciar la longitud completa de fragmentos cortos de ácidos nucleicos, por ejemplo de 200 pb. De esta manera, la longitud real del fragmento de ácido nucleico se sabría de un modo inmediato a partir de los datos secuenciados.

Dicho análisis de extremos emparejados también es posible usando otras plataformas de secuenciación, por ejemplo, el sistema de Applied Biosystems SOLiD. Para la plataforma 454 de Roche, debido a su longitud de lectura aumentada en comparación con otros sistemas de secuenciación masiva en paralelo, también es posible determinar la longitud de un fragmento a partir de su secuencia completa.

La ventaja de enfocar los análisis de datos sobre el subconjunto de etiquetas secuenciadas corresponde a fragmentos de ácido nucleico cortos en el espécimen de plasma materno original debido a que el subconjunto de datos podría enriquecerse eficazmente con secuencias de ADN derivadas del feto. Esto es porque las moléculas de

ADN fetal en el plasma materno comprenden fragmentos más cortos que las moléculas de ADN de fondo materno (Chan et al Clin Chem 2004; 50: 88-92). De acuerdo con la FIG. 7, el número de etiquetas secuenciadas necesario para diferenciar casos euploides de trisomía 21 se reduciría a medida que aumenta la concentración de ADN fetal fraccional.

La selección post-secuenciación de subconjuntos de grupos de ácido nucleico es diferente de otras estrategias de enriquecimiento de ácidos nucleicos que se realizan antes del análisis de especímenes, tales como el uso de electroforesis en gel o columnas de exclusión por tamaño para la selección de ácidos nucleicos de tamaños particulares, que requieren la separación física del grupo enriquecido del grupo de ácidos nucleicos de fondo. Los procedimientos físicos introducirían más etapas experimentales y pueden ser propensos a problemas tales como contaminación. La selección post-secuenciación por ordenador de subconjuntos de resultados de secuenciación permitiría también variar la selección dependiendo de la sensibilidad y especificidad requerida para la determinación de la enfermedad.

Las estrategias bioinformáticas, informatizadas y estadísticas usadas para determinar si se obtiene un espécimen de plasma materno de una gestante que concibe un feto con trisomía 21 o euploide podrían recopilarse en un producto de programa informatizado usado para determinar parámetros de resultados de secuenciación. El funcionamiento del programa informático implicaría la determinación de una cantidad cuantitativa a partir del cromosoma posiblemente aneuploide así como una o más cantidades de uno o más del resto de cromosomas. Un parámetro podría determinarse y compararse con valores de límite apropiados para determinar si existe una aneuploidía cromosómica fetal para el cromosoma posiblemente aneuploide.

Ejemplos

Se ofrecen los siguientes ejemplos para ilustrar, pero no para limitar la invención reivindicada.

1. DIAGNÓSTICO PRENATAL DE TRISOMIA 21 FETAL

En este estudio participaron ocho gestantes. Todas las gestantes estaban en el 1er o en el 2° trimestre de gestación y con un embarazo de feto único. Cuatro de ellas gestaban un feto con trisomía 21 y las otras cuatro gestaban un feto euploide. Se extrajeron veinte mililitros de sangre venosa periférica de cada mujer. El plasma materno se recogió después de centrifugar a 1600 x g durante 10 minutos y se centrifugó adicionalmente a 16000 x g durante 10 minutos. Después, se extrajo ADN de 5-10 ml de cada muestra plasmática. Después, el ADN plasmático materno se usó para secuenciación masiva en paralelo mediante el Analyzer Illumina Genome de acuerdo con las instrucciones del fabricante. Los técnicos que realizaban la secuenciación desconocían los diagnósticos fetales durante los análisis de secuenciación y datos de secuencias.

En resumen, se usaron aproximadamente 50 ng de ADN plasmático materno para preparar una biblioteca de ADN. Es posible comenzar con menores cantidades, tales como 15 ng o 10 ng de ADN plasmático materno. Los fragmentos de ADN plasmático materno se enromaron, se ligaron a adaptadores Solexa y se seleccionaron fragmentos de 150-300 pb por purificación en gel. Como alternativa, los fragmentos de ADN plasmático materno enromados y ligados a adaptadores podrían pasarse a través de columnas (por ejemplo AMPure, Agencourt) para eliminar adaptadores no ligados sin selección por tamaño antes de la generación de grupos. El ADN ligado a adaptador se hibridó a la superficie de células de flujo y se generaron grupos de ADN usando la estación de grupos Illumina, seguido de 36 ciclos de secuenciación en el Analyzer Illumina Genome. El ADN de cada espécimen de plasma materno se secuenció por una célula de flujo. Las lecturas secuenciadas se compilaron usando Solexa Analysis Pipeline. Después, todas las lecturas se alinearon con la secuencia genómica humana de referencia oculta con repeticón, ensamblaje NCBI 36 (números de entrada a GenBank: NC_000001 a NC_000024), usando la aplicación Eland.

En este estudio, para reducir la complejidad de los análisis de los datos, solo se consideraron secuencias adicionales que se habían mapeado en una sola localización en la referencia del genoma humano oculto con repetición. Como alternativa, podrían usarse otros subconjuntos, o todo el conjunto, de, los datos secuenciados. Se contó el número total de secuencias únicamente mapeables para cada espécimen. El número de secuencias únicamente alineadas con el cromosoma 21 se expresó como una proporción con respecto al recuento total de secuencias alineadas para cada espécimen. Como el plasma materno contiene ADN fetal entre un fondo de ADN de origen materno, el feto con trisomía 21 contribuiría a originar etiquetas extra secuenciadas del cromosoma 21 debido a la presencia de una copia extra del cromosoma 21 en el genoma fetal. Por tanto, el porcentaje de secuencias de cromosoma 21 en el plasma materno de una gestante portadora de un feto con trisomía 21 sería ser mayor que el de una gestante con un feto euploide. El análisis no requiere la selección de secuencias específicas de feto. Tampoco requiere la separación física previa de ácidos nucleicos fetales de maternos. Tampoco requiere la necesidad de diferenciar o identificar secuencias fetales de maternas después de la secuenciación.

La FIG. 3A muestra el porcentaje de secuencias mapeadas en el cromosoma 21 (porcentaje de representación del cromosoma 21) para cada una de las 8 muestras de ADN de plasma materno. El porcentaje de representación del cromosoma 21 fue significativamente superior en el plasma materno de embarazos con trisomía 21 que en el de

embarazos euploides. Estos datos sugieren que podría conseguirse un diagnóstico prenatal no invasivo de aneuploidía fetal determinando el porcentaje de representación del cromosoma aneuploide en comparación con el de una población de referencia. Como alternativa, la sobre-representación del cromosoma 21 podría detectarse comparando el porcentaje de representación del cromosoma 21 obtenido experimentalmente con el porcentaje de representación de las secuencias de cromosoma 21 esperadas para un genoma humano euploide. Esto podría realizarse ocultando o no las regiones repetidas en el genoma humano.

Cinco de las ocho gestantes eran portadoras de un feto de sexo masculino. Las secuencias mapeadas en el cromosoma Y serían específicas de feto. El porcentaje de secuencias mapeadas en el cromosoma Y se usó para calcular la concentración de ADN fetal fraccional en el espécimen de plasma materno original. Además, la concentración de ADN fetal fraccional también se determinó usando PCR digital de microfluidos que implica los genes parálogos de *proteínas de dedos de cinc,* asociados al cromosoma X *(ZFX)* y al cromosoma Y *(ZFY).*

La FIG.3B muestra la correlación de las concentraciones de ADN fetal fraccional deducida por el porcentaje de representación del cromosoma Y por secuenciación y determinada por PCR digital de microfluidos para *ZFY/ZFX.* Hubo una correlación positiva entre las concentraciones de ADN fetal fraccional en el plasma materno determinada por estos dos métodos. El coeficiente de correlación (r) fue de 0,917 en el análisis de correlación de Pearson.

En la FIG. 4A se muestran los porcentajes de secuencias de ADN plasmático materno alineados con cada uno de los 24 cromosomas (22 autosomas y cromosomas X e Y) para dos casos representativos. Una gestante portaba un feto con trisomía 21 y la otra un feto euploide. El porcentaje de representación de las secuencias mapeadas en el cromosoma 21 es mayor en la gestante portadora de un feto con trisomía 21 cuando se compara con el de la gestante portadora de un feto normal.

Las diferencias (%) del porcentaje de representación por cromosoma entre los especímenes de ADN de plasma materno de los dos casos anteriores se muestran en la FIG. 4B. El porcentaje de diferencia para un cromosoma particular se calcula usando la siguiente fórmula:

$$\text{Porcentaje de diferencia (\%)} = (P_{21} - P_E)/P_E \times 100\ \%,$$

donde

P_{21} = porcentaje de secuencias de ADN plasmático alineado con el cromosoma particular en la gestante portadora de un feto con trisomía 21 y;

P_E = porcentaje de secuencias de ADN plasmático alineado con el cromosoma particular en la gestante portadora de un feto euploide.

Como se muestra en la FIGURA 4B, hay una sobre-representación de secuencias de cromosoma 21 del 11 % en el plasma de la gestante portadora de un feto con trisomía 21 cuando se compara con la gestante portadora en un feto euploide. Para las secuencias alineadas con otros cromosomas, las diferencias entre los dos casos estaban dentro de 5 %. Como el porcentaje de representación para el cromosoma 21 está aumentado en la trisomía 21 en comparación con las muestras de plasma materno euploides, la diferencia (%) podría referirse alternativamente al grado de sobre-representación en las secuencias del cromosoma 21. Además de las diferencias (%) y diferencias absolutas entre la representación del porcentaje de cromosoma 21, también podrían calcularse razones de los recuentos procedentes de muestras de ensayo y de referencia y serían indicativas del grado de sobrerepresentación del cromosoma 21 en trisomía 21 en comparación con muestras euploides.

Para las cuatro mujeres gestantes, cada una de ellas portadora de un feto euploide, se alineó una media de 1,345 % de sus secuencias de ADN plasmático con el cromosoma 21. En las cuatro gestantes portadoras de un feto con trisomía 21, tres de sus fetos eran de sexo masculino. El porcentaje de representación del cromosoma 21 se calculó para cada uno de estos tres casos. La diferencia (%) en el porcentaje de representación del cromosoma 21 para cada uno de estos tres casos de trisomía 21 del porcentaje de representación medio del cromosoma 21 derivado de valores de los cuatro casos euploides se determinó como se ha descrito anteriormente. En otras palabras, la media de los cuatro casos portadores de un feto euploide se usó como la referencia en este cálculo. Las concentraciones de ADN fetal fraccional de estos tres casos de fetos de sexo masculino con trisomía 21 se dedujeron de su porcentaje de representación respectivo de secuencias del cromosoma Y.

En la FIG. 5 se muestra la correlación entre el grado de sobre-representación para secuencias del cromosoma 21 y las concentraciones de ADN fetal fraccional. Hubo una correlación positiva significativa entre los dos parámetros. El coeficiente de correlación (r) fue de 0,898 en el análisis de correlación de Pearson. Estos resultados indican que el grado de sobre-representación de secuencias del cromosoma 21 en plasma materno está relacionado con la concentración de ADN fetal fraccional en la muestra de plasma materno. Por lo tanto, para identificar embarazos que conlleven fetos con trisomía 21, podrían determinarse los valores de límite en el grado de sobre-representación relevante para secuencias del cromosoma 21 con las concentraciones de ADN fetal fraccional.

La determinación de la concentración de ADN fetal fraccional en plasma materno también puede realizarse de manera distinta al proceso de secuenciación. Por ejemplo, la concentración de ADN del cromosoma Y podría predeterminarse usando PCR en tiempo real, PCR de microfluidos o espectrometría de masas. Por ejemplo, en la FIG. 3B se ha demostrado que hay buena correlación entre las concentraciones de ADN fetal estimadas basándose en el recuento de cromosoma Y generado durante el proceso de secuenciación y la razón *ZFY/ZFX* generada externa al proceso de secuenciación. De hecho, la concentración de ADN fetal podría determinarse usando loci distintos del cromosoma Y y aplicables a fetos de sexo femenino. Por ejemplo, Chan et al mostraron que las secuencias *FASSF1A* metiladas derivadas de feto se detectarían en el plasma de gestantes en el fondo de secuencias *FASSF1A* no metiladas derivadas de la madre (Chan et al, Clin Chem 2006; 52:2211-8). La concentración de ADN fetal fraccional puede por tanto determinarse dividiendo la cantidad de secuencias *FASSF1A* metiladas entre la cantidad de secuencias *FASSF1A* totales (metiladas y no metiladas).

Se espera que el plasma materno se prefiera sobre el suero materno para la realización práctica de la invención ya que el ADN se libera de células sanguíneas maternas durante la coagulación sanguínea. Por tanto, si se usa suero, se espera que la concentración de ADN fetal fraccional será más baja en plasma materno que en suero materno. En otras palabras, si se usa suero materno, se espera que se necesitan más secuencias para generar un diagnóstico de aneuploidía cromosómica fetal, cuando se compara con una muestra de plasma obtenida al mismo tiempo de la misma gestante.

Incluso otra forma alternativa de determinar la concentración de ADN fetal fraccional sería a través de la cuantificación de diferencias polimórficas entre la gestante y el feto (Dhallan R, et al. 2007 *Lancet,* 369, 474-481). Un ejemplo de este método sería dirigirse a sitios polimórficos en los que la gestante es homocigota y el feto es heterocigoto. La cantidad de alelo específico de feto puede compararse con la cantidad de alelo común para determinar la concentración del ADN fetal fraccional.

A diferencia de las técnicas existentes para detectar aberraciones cromosómicas, incluyendo hibridación genómica comparativa, hibridación genómica comparativa por micromatriz, reacción en cadena de polimerasa cuantitativa en tiempo real, que detectan y cuantifican una o más secuencias específicas, la secuenciación masiva en paralelo no depende de la detección o análisis de secuencias predeterminadas o de un conjunto predefinido de secuencias de ADN. Se secuencia una fracción representativa aleatoria de moléculas de ADN del conjunto del espécimen. El número de etiquetas secuenciadas diferentes alineadas con diversas regiones cromosómicas se compara entre especímenes que contienen o no la especie de ADN de interés. Se manifestarían aberraciones cromosómicas por diferencias en el número (o porcentaje) de secuencias alineadas con cualquier región cromosómica determinada en los especímenes.

En otro ejemplo, la técnica de secuenciación en ADN libre en plasma puede usarse para detectar las aberraciones cromosómicas en el ADN plasmático para la detección de un cáncer específico. Diferentes cánceres tienen un grupo de aberraciones cromosómicas típicas. Pueden usarse cambios (amplificaciones y deleciones) en regiones cromosómicas múltiples. Por tanto, habría una proporción aumentada de secuencias alineadas con las regiones amplificadas y una proporción disminuida de secuencias alineadas con regiones disminuidas. El porcentaje de representación por cromosoma podría compararse con el tamaño de cada cromosoma correspondiente en un genoma de referencia expresado como porcentaje de representación genómica de cualquier cromosoma determinado en relación con el genoma completo. También pueden usarse comparaciones directas o comparaciones con un cromosoma de referencia.

II. SECUENCIACIÓN EXCLUSIVA DE UNA FRACCIÓN DEL GENOMA HUMANO

En el experimento descrito en el Ejemplo I anterior, el ADN plasmático materno de cada espécimen individual se secuenció usando solo una célula de flujo. El número de etiquetas secuenciales generadas de cada uno de los especímenes ensayados por el proceso de secuenciación se muestra en la FIG.6. T21 indica una muestra obtenida de una gestante portadora de un feto con trisomía 21.

Como se secuenciaron 36 pb de cada una de los fragmentos de ADN plasmático materno, el número de nucleótidos/pares de bases secuenciado de cada espécimen podría determinarse por 36 pb multiplicado por el recuento de etiquetas secuenciadas y también se muestra en la FIG. 6. Dado que en el genoma humano hay aproximadamente 3 billones de pares de bases, la cantidad de datos de secuenciación generada de cada espécimen plasmático materno representa solo una fracción, que varía del 10 % al 13 %.

Adicionalmente, en este estudio, solo se usaron las etiquetas únicamente secuenciadas mapeables, denominadas U0 en la nomenclatura del programa informático Eland, para demostrar la presencia de la sobre-representación en la cantidad de las secuencias de cromosoma 21 en los especímenes de plasma materno de gestantes que portaban cada una un feto con trisomía 21, como se describe en el ejemplo I anterior. Como se muestra en la FIG. 6, las secuencias U0 solo representan un subconjunto de todas las etiquetas secuenciadas generadas de cada espécimen y adicionalmente representan una proporción homogénea más pequeña, sobre el 2 %, del genoma humano. Estos datos indican que la secuenciación de solo una parte de las secuencias del genoma humano presente en el espécimen ensayado es suficiente para conseguir el diagnóstico de aneuploidía fetal.

III. DETERMINACIÓN DEL NÚMERO DE SECUENCIAS NECESARIO

En este análisis se utilizó el resultado de la secuenciación del ADN plasmático de una gestante portadora de un feto de sexo masculino euploide. El número de etiquetas secuenciadas que puede mapearse sin desapareamientos con respecto a la secuencia genómica humana de referencia fue de 1.990.000. Los subconjuntos de secuencias se seleccionaron aleatoriamente a partir de estas 1.990.000 etiquetas y el porcentaje de secuencias alineadas con el cromosoma 21 se calculó dentro de cada subconjunto. El número de secuencias en los subconjuntos varió de 60.000 a 540.000 secuencias. Para cada tamaño del subconjunto, se reunieron múltiples subconjuntos del mismo número de etiquetas secuenciadas por selección aleatoria de las etiquetas secuenciadas del conjunto total hasta ninguna otra combinación posible. El porcentaje medio de secuencias alineadas con el cromosoma 21 y su desviación típica (DT) se calculó después a partir de los subconjuntos múltiples dentro de cada tamaño del subconjunto. Estos datos se compararon a través de diferentes tamaños de subconjuntos para determinar el efecto del tamaño de subconjunto sobre la distribución del porcentaje de secuencias alineadas con el cromosoma 21. Los percentiles 5 y 95 de los porcentajes se calcularon después de acuerdo con media y la DT.

Cuando una gestante es portadora de un feto con trisomía 21, las etiquetas secuenciadas alineadas con el cromosoma 21 deben sobre-representarse en el plasma materno debido a una dosis extra de cromosoma 21 procedente del feto. El grado de sobre-representación depende del porcentaje de ADN fetal en la muestra de ADN plasmático materno según la siguiente ecuación:

$$Per_{T21} = Per_{Eu} \times (1 + f/2)$$

donde

Per_{T21} representa el porcentaje de secuencias alineadas con el cromosoma 21 en una gestante con un feto con trisomía 21; y

Per_{Eu} representa el porcentaje de secuencias alineadas con el cromosoma 21 en una gestante con un feto euploide; y

f representa el porcentaje de ADN fetal en el ADN plasmático materno.

Como se muestra en la FIG. 7, la DT para los porcentajes de secuencias alineadas con el cromosoma 21 disminuye cuando aumenta el número de secuencias en cada subconjunto. Por lo tanto, cuando el número de secuencias en cada subconjunto aumenta, el intervalo entre el percentil 5 y 95 disminuye. Cuando el intervalo del 5 % - 95 % para los casos euploides y de trisomía 21 no se solapa, entonces la diferenciación entre los dos grupos de casos sería posible con una precisión > 95 %.

Como se muestra en la FIG. 7, el tamaño del subconjunto mínimo para la diferenciación de casos de trisomía 21 de euploides depende del porcentaje de ADN fetal. Los tamaños de subconjunto mínimo para la diferenciación de casos de trisomía 21 de euploides fueron 120.000, 180.000 y 540.000 secuencias para porcentajes de ADN fetal del 20 %, 10 % y 5 %, respectivamente. En otras palabras, el número de secuencias necesario a analizar sería de 120.000 para determinar si un feto tiene trisomía 21 cuando una muestra de ADN plasmático materno contiene ADN fetal al 20 %. El número de secuencias necesario a analizar debería aumentar a 540.000 cuando el porcentaje de ADN fetal disminuye al 5 %.

Dado que los datos se generaron usando una secuenciación de 36 pares de bases, 120.000, 180.000 y 540.000 secuencias corresponden al 0,14 %, 0,22 % y 0,65 % del genoma humano, respectivamente. Dado que se indicó que el intervalo más bajo de concentraciones de ADN fetal en el plasma materno obtenido de embarazos tempranos era del 5 % (Lo, YMD et al. 1998 *Am J Hum Genet* 62, 768-775), la secuenciación de aproximadamente 0,6 % del genoma humano puede representar la cantidad mínima de secuenciación necesaria para el diagnóstico con una precisión de al menos el 95 % en la detección de aneuploidía cromosómica fetal para cualquier embarazo.

IV. SECUENCIACIÓN ALEATORIA

Para ilustrar que los fragmentos de ADN secuenciados se seleccionaban aleatoriamente durante el proceso de secuenciación, se obtuvieron las etiquetas secuenciadas generadas a partir de las ocho muestras de plasma materno analizadas en el ejemplo I. Para cada espécimen de plasma materno, se determinaron las posiciones de partida en relación con la secuencia genómica humana de referencia, ensamblaje NCBI, 36 de cada uno de los 36 pb secuenciados que se alineaban únicamente con el cromosoma 21 sin emparejamientos erróneos. Después se ordenó, en orden ascendiente, el número de posición de partida de los grupos de etiquetas secuenciadas alineadas de cada espécimen. Para el cromosoma 22 se realizó un análisis similar. Con fines ilustrativos, en las FIG. 8A y 8B, respectivamente, se muestran las diez primeras posiciones de partida del cromosoma 21 y 22 de cada uno de los especímenes de plasma materno. Como puede apreciarse a partir de estas Tablas, los conjuntos de fragmentos de ADN secuenciado no eran idénticos entre las muestras.

Cualquiera de los componentes informáticos o funciones descritas en esta solicitud pueden implementarse como código de programación para ejecutar por un procesador usando cualquier lenguaje informático adecuado tal como, por ejemplo, Java, C+ o Perl usando, por ejemplo, técnicas convencionales orientadas a objeto. El código de programación puede almacenarse como una serie de instrucciones o comandos en un medio legible por ordenador para almacenamiento y/o transmisión, los medios adecuados incluyen memoria de acceso aleatorio (RAM), o memoria solo de lectura (ROM), un medio magnético, tal como, un disco duro o un disquete, o un medio óptico tal como un disco compacto (CD) o un DVD (disco versátil digital), memoria flash, y similar. El medio legible por ordenador puede ser cualquier combinación de dichos dispositivos de almacenamiento o transmisión.

Dichos programas también pueden codificarse y transmitirse usando señales de transmisión adaptadas para la transmisión mediante redes alámbricas, ópticas y/o inalámbricas que conforman una variedad de protocolos incluyendo el de Internet. Como tal, un medio legible por ordenador de acuerdo con una realización de la presente invención puede crearse usando una señal de datos codificada con dichos programas. Los medios legibles por ordenador codificados con el código de programación pueden empaquetarse con un dispositivo compatible o proporcionarse por separado de otros dispositivos (por ejemplo, descarga vía Internet). Cualquier medio legible por ordenador puede residir sobre o dentro de un solo producto de programa informático (por ejemplo, un disco duro o todo un sistema informatizado) y puede estar presente sobre o dentro de productos de programas informáticos diferentes dentro de un sistema o red. Un sistema informatizado puede incluir un monitor, una impresora u otra pantalla adecuada para proporcionar a un usuario cualquiera de los resultados mencionados en el presente documento.

En la FIG. 9. se muestra un ejemplo de un sistema informatizado. Los subsistemas mostrados en la FIGURA 9 están interconectados mediante un bus de sistema 975. Se muestran subsistemas adicionales, tales como una impresora 974, un teclado 978, un disco duro 979, un monitor 976, que está acoplado a un adaptador de pantalla 982 y otros. Los dispositivos periféricos y de entrada/salida (E/S), acoplados al controlador E/O 971 pueden conectarse al sistema informatizado mediante cualquiera de los diversos medios conocidos en la técnica, tal como un puerto en serie 977. Por ejemplo, el puerto en serie 977 o interfaz externa 981 pueden usarse para conectar el aparato de ordenador a una red de área amplia tal como Internet, un dispositivo de entrada de ratón o un escáner. La interconexión mediante el bus del sistema permite que al procesador central 973 comunicar con cada subsistema y controlar la ejecución de instrucciones desde la memoria del sistema 972 o del disco duro 979, así como el intercambio de información entre subsistemas. La memoria del sistema 972 y/o el disco duro 979 pueden representar un medio legible por ordenador.

La descripción anterior de realizaciones ejemplares de la invención se ha presentado con fines ilustrativos y descriptivos. No pretende ser exhaustiva o limitar la invención a la forma precisa descrita, y son posibles muchas modificaciones y variaciones a la luz de las enseñanzas anteriores. Las realizaciones se eligieron y describieron para explicar mejor los principios de la invención y sus aplicaciones prácticas para permitir a otros expertos en la técnica utilizar mejor la invención en diversas realizaciones y con diversas modificaciones que sean adecuadas para el uso particular contemplado.

REIVINDICACIONES

1. Un método implementado por ordenador de análisis de una muestra biológica de un organismo para deleciones o amplificaciones en una o más regiones cromosómicas asociadas con cáncer, incluyendo la muestra biológica moléculas de ADN sin células que se originan a partir de células no malignas y potencialmente de células tumorales asociadas con cáncer, comprendiendo el método:

recibir etiquetas secuenciadas obtenidas de una secuenciación aleatorizada de moléculas de ADN contenidas en la muestra biológica, siendo la muestra biológica plasma, suero, saliva u orina;
alinear una pluralidad de las etiquetas secuenciadas con un genoma humano de referencia;
determinar una primera cantidad de las etiquetas secuenciadas identificadas como alineación con una primera región cromosómica que es parte de un primer cromosoma;
determinar una segunda cantidad de etiquetas secuenciadas como alineaciones con una o más segundas regiones cromosómicas;
determinar un primer parámetros a partir de la primera cantidad y de la segunda cantidad, en donde el parámetro incluye una proporción entre la primera cantidad y la segunda cantidad, en donde una anomalía en el primer parámetro para la primera región cromosómica se asocia con cáncer; y
comparar el primer parámetro con uno o más valores umbral para determinar una clasificación de si la primera región cromosómica presenta una deleción y/o una amplificación asociada con cáncer.

2. El método de la reivindicación 1, en donde la clasificación incluye si la primera región cromosómica se amplifica, y en donde se hace una clasificación de la amplificación cuando el primer parámetro es más de un primer valor umbral de los uno o más valores umbrales.

3. El método de la reivindicación 1, en donde la clasificación incluye si la primera región cromosómica tiene una deleción, y en donde se hace una clasificación de la deleción cuando el primer parámetro es menos de un primer valor umbral de los uno o más valores umbrales.

4. El método de la reivindicación 1, que además comprende:

determinar una tercera cantidad de etiquetas secuenciadas identificadas como alineación con una tercera región cromosómica, en donde una anomalía en la tercera región cromosómica se asocia con cáncer;
determinar un segundo parámetro para una tercera región cromosómica, en donde el segundo parámetro incluye una proporción entre la tercera cantidad y la segunda cantidad;
comparar el segundo parámetro con uno o más valores umbral para determinar una clasificación de si la tercera región cromosómica presenta una deleción y/o una amplificación asociada con cáncer; y usar las clasificaciones de la primera y tercera región cromosómica para detectar si el organismo tiene cáncer.

5. El método de la reivindicación 1, en donde el primer parámetro se calcula a partir de una proporción que incluye la primera cantidad.

6. El método de la reivindicación 5, en donde la proporción proporciona un porcentaje de las etiquetas secuenciadas identificadas como alineación con la primera región cromosómica entre las etiquetas secuenciadas.

7. El método de la reivindicación 5, en donde la proporción se obtiene a partir de uno cualquiera o más de un recuento fraccionado del número de etiquetas secuenciadas, un número fraccionado de nucleótidos secuenciados, y una longitud fraccionada de las secuencias acumuladas.

8. El método de la reivindicación 5, en donde las secuencias que se usan para determinar la primera cantidad se seleccionan para que sean menos de un número específico de bases.

9. El método de la reivindicación 1, en donde las moléculas de ADN de la muestra se han enriquecido para secuencias que se originan a partir de la primera región cromosómica.

10. El método de la reivindicación 1, en donde la proporción alineada representa al menos una fracción predeterminada del genoma humano de referencia, en donde la fracción representa al menos el 0,1% del genoma humano de referencia.

11. El método de la reivindicación 1, en donde al menos uno de los valores umbral es un valor de referencia establecido en una muestra biológica normal, y en donde el al menos un valor umbral se determina a partir de una media del valor del primer parámetro en una pluralidad de muestras biológicas normales.

12. El método de la reivindicación 1, que comprende además:

identificar una cantidad de ADN del tumor en la muestra biológica; y
calcular un número N de secuencias que se van a analizar basándose en una precisión deseada, en donde al

menos uno de los valores umbrales proviene de un porcentaje de células tumorales en la muestra biológica.

13. El método de la reivindicación 1, en donde las etiquetas secuenciadas identificadas como alineación con la primera región cromosómica únicamente se alinean con la primera región cromosómica, y en donde las etiquetas secuenciadas identificadas como alineación con las una o más segundas regiones cromosómicas únicamente se alinean con solo una de las segundas regiones cromosómicas.

14. El método de una cualquiera de las reivindicaciones 1, en donde la primera cantidad se determina basándose en un grupo de etiquetas secuenciadas que se alinean con una pluralidad de posiciones de la primera región cromosómica, en donde no se predetermina qué parte del genoma humano de referencia que se corresponde a las etiquetas secuenciadas se alinea con la primera región cromosómica.

15. El método de la reivindicación 1, en donde el primer parámetro es una representación fraccionada de las etiquetas secuenciadas identificadas como alineación con la primera región cromosómica, siendo la representación fraccionada una medida de una proporción de moléculas de ADN en la muestra biológica que son de la primera región cromosómica y en donde los uno o más valores umbrales tienen en cuenta un tamaño de la primera región cromosómica en relación con un tamaño de las una o más segundas regiones cromosómicas.

16. Un programa de ordenador que comprende un código de programa informático capaz de ejecutarse mediante un procesador de un sistema informático, estando el programa de ordenador configurado, cuando se ejecuta mediante el procesador, para provocar que el procesador realice un método de acuerdo con cualquiera de las reivindicaciones anteriores.

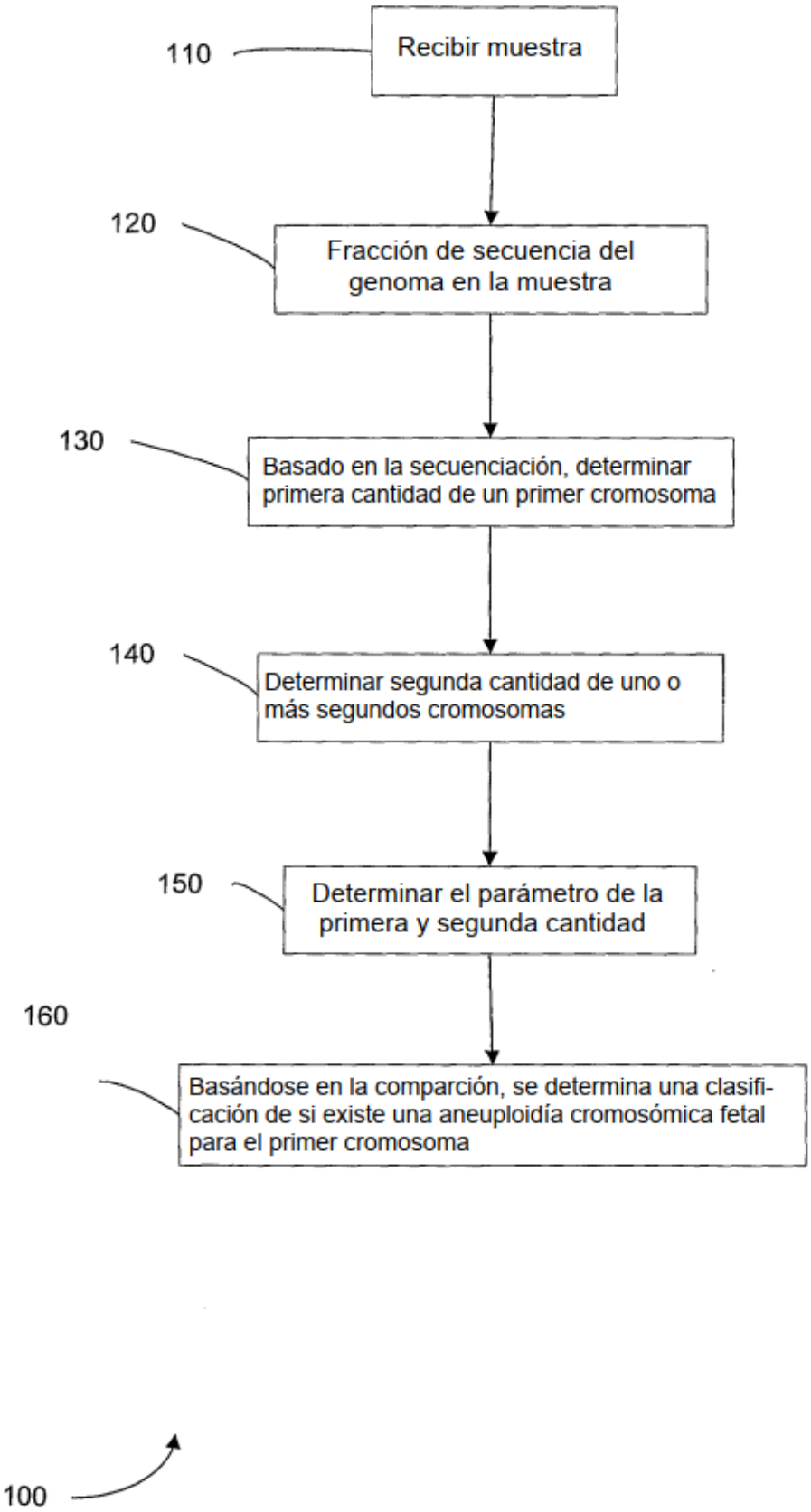
110
Recibir muestra
120
Fracción de secuencia del genoma en la muestra
130
Basado en la secuenciación, determinar primera cantidad de un primer cromosoma
140
Determinar segunda cantidad de uno o más segundos cromosomas
150
Determinar el parámetro de la primera y segunda cantidad
160
Basándose en la comparción, se determina una clasificación de si existe una aneuploidía cromosómica fetal para el primer cromosoma
100

FIG. 1

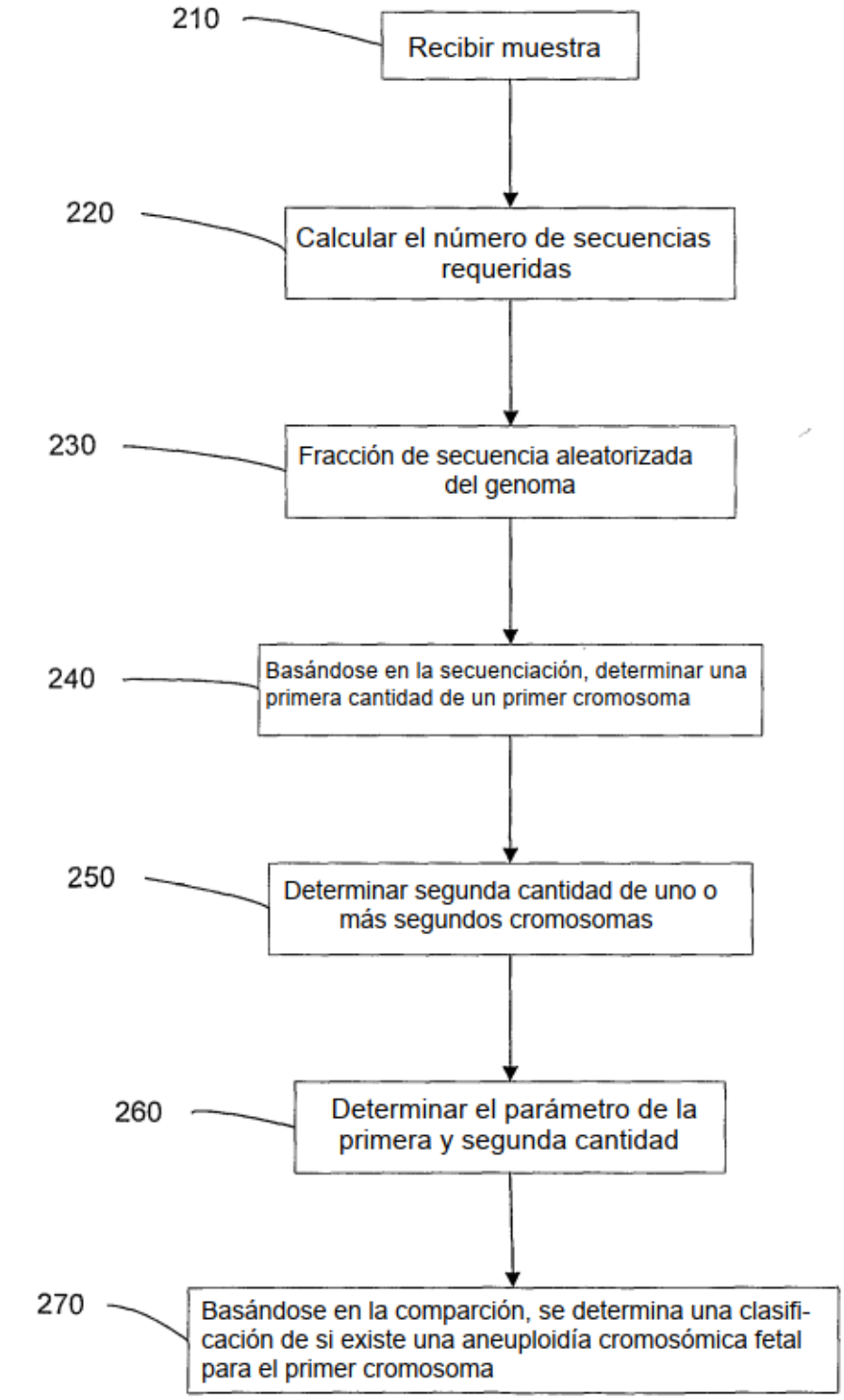
210
Recibir muestra
220
Calcular el número de secuencias requeridas
230
Fracción de secuencia aleatorizada del genoma
240
Basándose en la secuenciación, determinar una primera cantidad de un primer cromosoma
250
Determinar segunda cantidad de uno o más segundos cromosomas
260
Determinar el parámetro de la primera y segunda cantidad
270
Basándose en la comparción, se determina una clasificación de si existe una aneuploidía cromosómica fetal para el primer cromosoma
200

FIG. 2

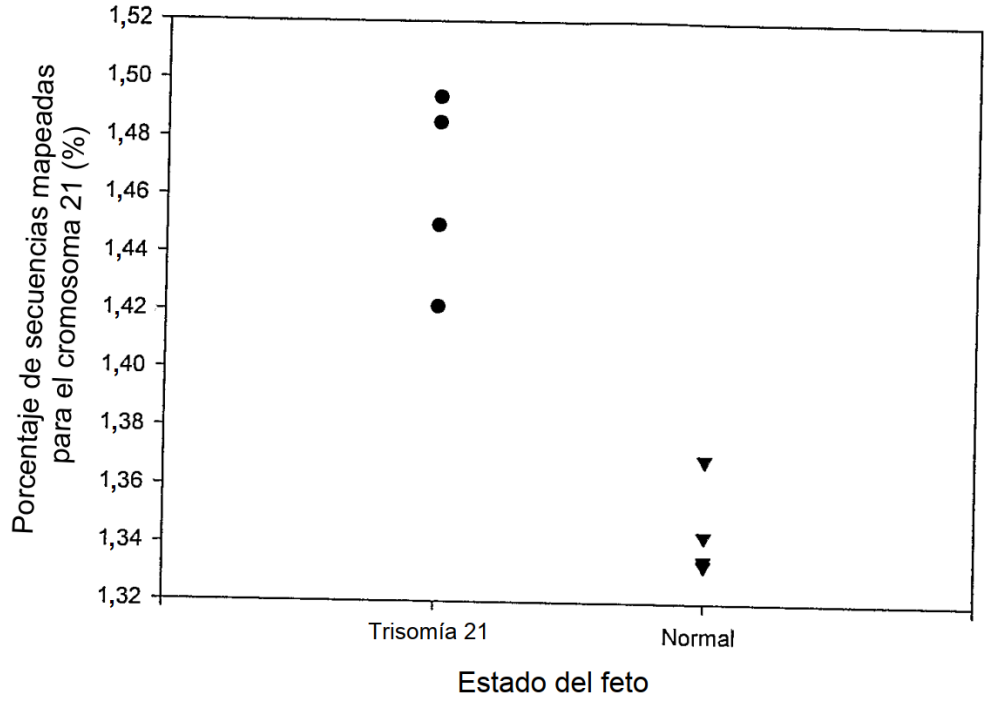
Porcentaje de secuencias mapeadas para el cromosoma 21 (%)
1,52
1,50
1,48
1,46
1,44
1,42
1,40
1,38
1,36
1,34
1,32
Trisomía 21
Normal
Estado del feto

FIG. 3A

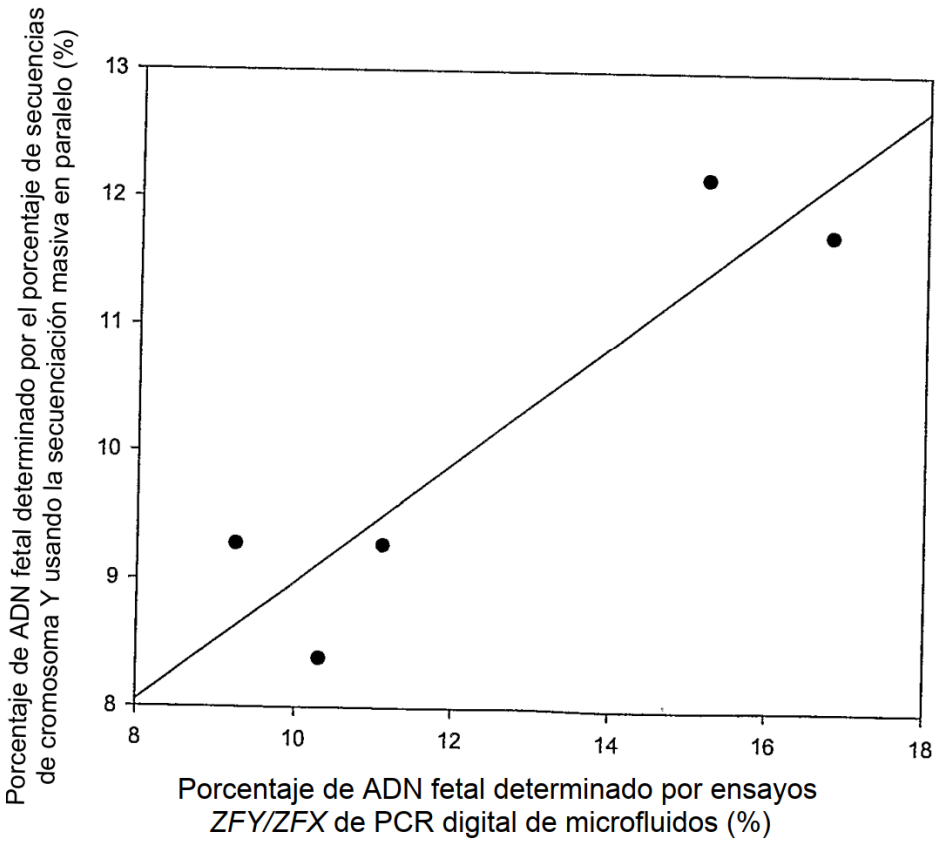
Porcentaje de ADN fetal determinado por el porcentaje de secuencias de cromosoma Y usando la secuenciación masiva en paralelo (%)
13
12
11
10
9
8
8
10
12
14
16
18
Porcentaje de ADN fetal determinado por ensayos *ZFY/ZFX* de PCR digital de microfluidos (%)

FIG. 3B

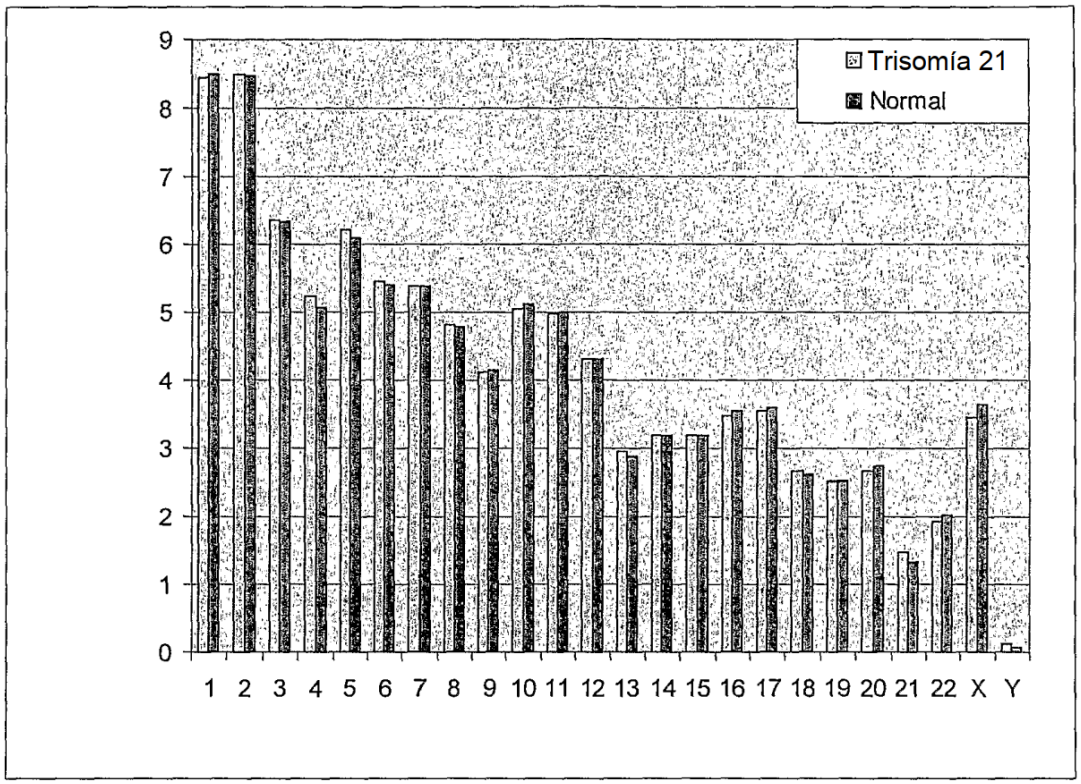
Trisomía 21
Normal
9
8
7
6
5
4
3
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

FIG. 4A

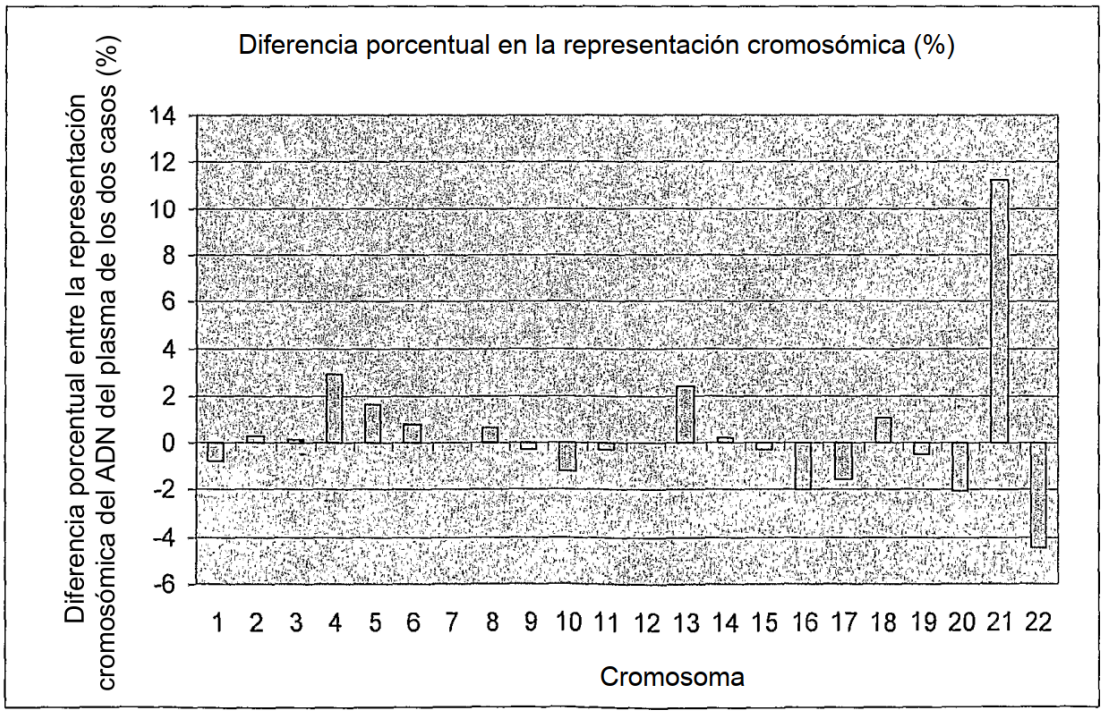
Diferencia porcentual en la representación cromosómica (%)
Diferencia porcentual entre la representación cromosómica del ADN del plasma de los dos casos (%)
14
12
10
8
6
4
2
0
-2
-4
-6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Cromosoma

FIG. 4B

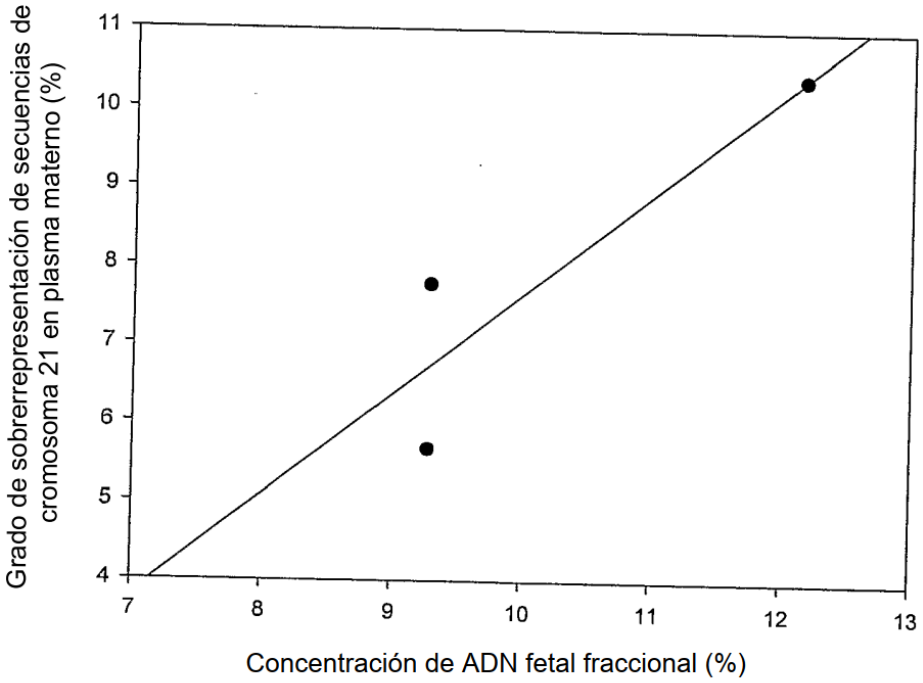
Grado de sobrerrepresentación de secuencias de cromosoma 21 en plasma materno (%)
11
10
9
8
7
6
5
4
7
8
9
10
11
12
13
Concentración de ADN fetal fraccional (%)

FIG. 5

Muestra	Número total de etiquetas secuenciadas	Número total de pares de bases secuenciados	Proporción del genoma humano (%)	Recuento U0	Recuento U0/ Recuento total (%)	Nº de pares de bases U0 secuenciados	Proporción U0 del genoma humano (%)
Muestra 1 T21	**1,12E+07**	**4,02E+08**	**13,4**	**1,93E+06**	**17,3**	**6,94E+07**	**2,31**
Muestra 2 T21	**1,04E+07**	**3,73E+08**	**12,4**	**1,80E+06**	**17,4**	**6,50E+07**	**2,17**
Muestra 3 T21	**8,90E+06**	**3,20E+08**	**10,7**	**2,09E+06**	**23,4**	**7,51 E+07**	**2,50**
Muestra 4 T21	**1,02E+07**	**3,69E+08**	**12,3**	**2,23E+06**	**21,7**	**8,02E+07**	**2,67**
Muestra 5 euploide	**1,06E+07**	**3,83+08**	**12,8**	**2,12E+06**	**19,9**	**7,62E+07**	**2,54**
Muestra 6 euploide	**9,58E+06**	**3,45E+08**	**11,5**	**1,91 E+06**	**20,0**	**6,88E+07**	**2,29**
Muestra 7 euploide	**9,55E+06**	**3,44E+08**	**11,5**	**2,01 E+06**	**21,0**	**7,22E+07**	**2,41**
Muestra 8 euploide	**9,09E+06**	**3,27E+08**	**10,9**	**2,09E+06**	**23,0**	**7,53E+07**	**2,51**

FIG. 6

ES 2 792 802 T3

N.º de secuencias en cada subconjunto	Euploide			T21 (5%)				T21 (10%)				T21 (20%)			
	Media	DT	percentil 5	percentil 95	Media	percentil 5	percentil 95	Diferenciación satisfactoria ¿(frente a euploide?)	Media	percentil 5	percentil 95	Diferenciación satisfactoria ¿(frente a euploide?	Media	percentil 5	percentil 95	Diferenciación satisfactoria ¿(frente a euploide?
58,000	1,322%	0,036%	1,250%	1,395%	1,355%	1,283%	1,428%	No	1,388%	1,316%	1,461%	No	1,454%	1,382%	1,527%	No
120,000	1,322%	0,022%	1,277%	1,367%	1,355%	1,311%	1,400%	No	1,388%	1,344%	1,433%	No	1,454%	1,410%	1,499%	Sí
180,000	1,322%	0,016%	1,291%	1,353%	1,355%	1,324%	1,386%	No	1,388%	1,357%	1,420%	Sí	1,454%	1,423%	1,486%	Sí
240,000	1,322%	0,013%	1,295%	1,349%	1,355%	1,328%	1,382%	No	1,388%	1,361%	1,415%	Sí	1,454%	1,427%	1,481%	Sí
300,000	1,322%	0,011%	1,301%	1,344%	1,355%	1,334%	1,377%	No	1,388%	1,367%	1,410%	Sí	1,454%	1,433%	1,476%	Sí
360,000	1,322%	0,010%	1,302%	1,343%	1,355%	1,335%	1,376%	No	1,388%	1,368%	1,409%	Sí	1,454%	1,434%	1,475%	Sí
420,000	1,322%	0,095%	1,303%	1,341%	1,355%	1,336%	1,374%	No	1,388%	1,369%	1,407%	Sí	1,454%	1,435%	1,473%	Sí
480,000	1,322%	0,009%	1,305%	1,339%	1,355%	1,338%	1,372%	No	1,388%	1,371%	1,405%	Sí	1,454%	1,437%	1,471%	Sí
540,000	1,322%	0,008%	1,306%	1,339%	1,355%	1,339%	1,372%	Sí	1,388%	1,372%	1,405%	Sí	1,454%	1,438%	1,471%	Sí

FIG. 7

Orden de etiqueta secuenciada	Muestra 1 T21	Muestra 2 T21	Muestra 3 T21	Muestra 4 T21	Muestra 5 Euploide	Muestra 6 Euploide	Muestra 7 Euploide	Muestra 8 Euploide
1	9796394	9798087	9798123	9795700	9797841	9795972	9796536	9795601
2	9797424	9798402	9798250	9797860	9798176	9796549	9796863	9797404
3	9797438	9798708	9798715	9797864	9798835	9797359	9798161	9798117
4	9798112	9799733	9799467	9798106	9800315	9797418	9 798401	9798175
5	9798394	9799852	9799730	9799209	9800385	9797446	9798722	9798387
6	9798729	9800362	9799788	9799440	9800554	9797860	9799752	9798816
7	9798768	9800914	9799834	9799440	9800820	9798062	9800317	9799732
8	9798816	9801528	9800332	9799741	9800829	9798135	9800800	9800144
9	9799421	9801779	9800869	9799812	9800832	9798707	9801380	9800421
10	9799464	9803801	9800881	9799833	9800866	9798715	9801699	9800827

FIG. 8A

Orden de etiqueta secuenciada	Muestra 1 T21	Muestra 2 T21	Muestra 3 T21	Muestra 4 T21	Muestra 5 Euploide	Muestra 6 Euploide	Muestra 7 Euploide	Muestra 8 Euploide
1	14469232	14510976	14510792	14510867	14434681	14510711	14512069	14510996
2	14503781	14511307	14510807	14511312	14506660	14510835	14512407	14522947
3	14510805	14511512	14510886	14511983	14511219	14511308	14522879	14530478
4	14510824	14511755	14511816	14522914	14511328	14511354	14609732	14565255
5	14510965	14511774	14512410	14522978	14511377	14511703	14628248	14609733
6	14511369	14512067	14522944	14523028	14511432	14511703	14675246	14628261
7	14511702	14512190	14523047	14565245	14511597	14512184	14675250	14668784
8	14511738	14522891	14564769	14628257	14511755	14512204	14680184	14680163
9	14523000	14539630	14565205	14635374	14512204	14524806	14692539	14689547
10	14524823	14541260	14565251	14665101	14512204	14565230	14694114	14689592

FIG. 8B

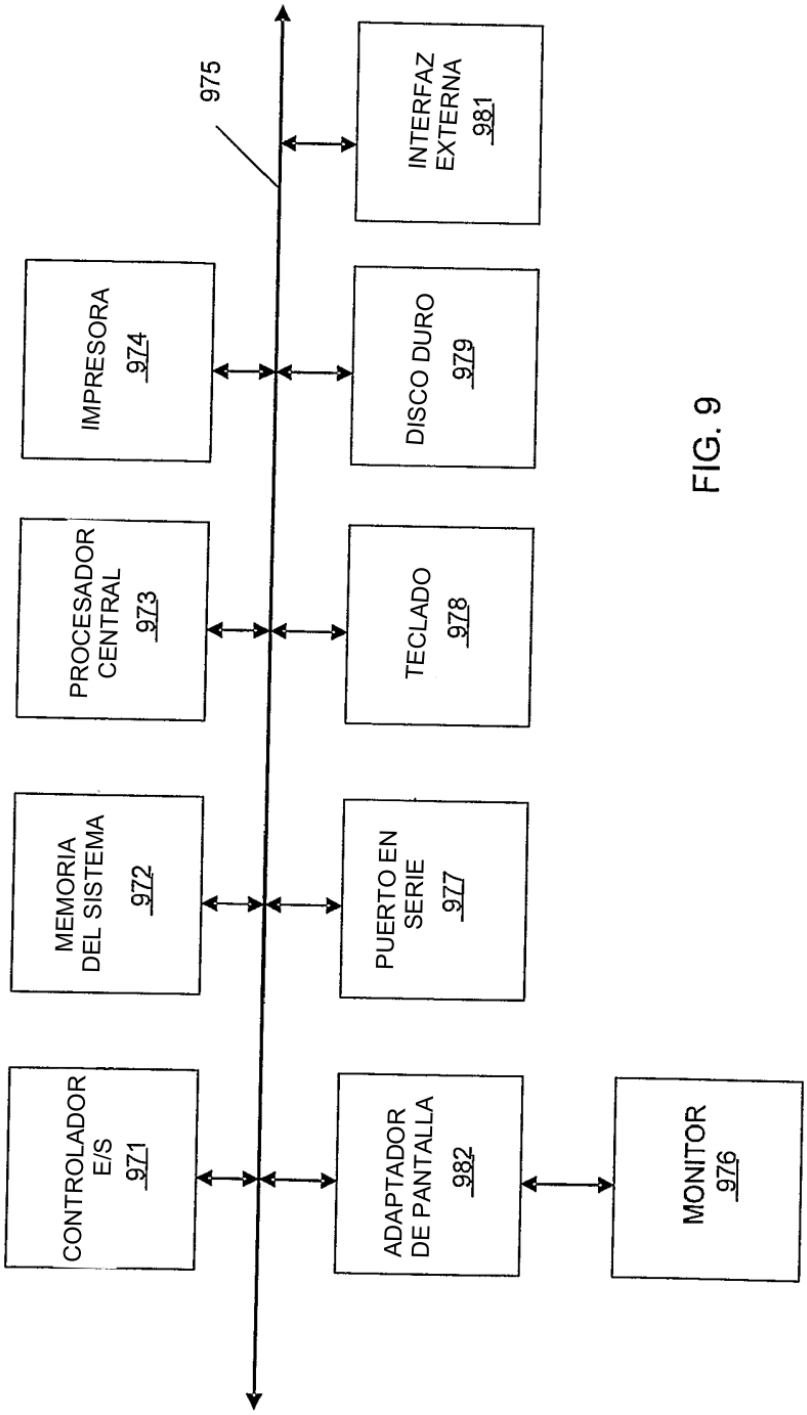
975
INTERFAZ EXTERNA 981
IMPRESORA 974
DISCO DURO 979
PROCESADOR CENTRAL 973
TECLADO 978
MEMORIA DEL SISTEMA 972
PUERTO EN SERIE 977
CONTROLADOR E/S 971
ADAPTADOR DE PANTALLA 982
MONITOR 976

FIG. 9