(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 886 508**

(51) Int. CI.:

G16B 20/00 (2009.01)
C12Q 1/6827 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **05.10.2012 PCT/US2012/059123**

(87) Fecha y número de publicación internacional: **11.04.2013 WO13052913**

(96) Fecha de presentación y número de la solicitud europea: **05.10.2012 E 12777999 (9)**

(97) Fecha y número de publicación de la concesión europea: **30.06.2021 EP 2764459**

(54) Título: **Métodos y procedimientos para la evaluación no invasiva de variaciones genéticas**

(30) Prioridad:

06.10.2011 US 201161544251 P
22.06.2012 US 201261663477 P
04.10.2012 US 201261709899 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
20.12.2021

(73) Titular/es:

SEQUENOM, INC. (100.0%)
3595 John Hopkins Court
San Diego, CA 92121, US

(72) Inventor/es:

DECIU, COSMIN;
DZAKULA, ZELJKO;
EHRICH, MATHIAS y
KIM, SUNG, KYUN

(74) Agente/Representante:

DEL VALLE VALIENTE, Sonia

DESCRIPCIÓN

Métodos y procedimientos para la evaluación no invasiva de variaciones genéticas

Campo

La tecnología proporcionada en el presente documento se refiere en parte a métodos, procesos y aparatos para la evaluación no invasiva de variaciones genéticas.

Antecedentes

La información genética de organismos vivos (por ejemplo, animales, plantas y microorganismos) y otras formas de información genética de replicación (por ejemplo, virus) está codificada en ácido desoxirribonucleico (ADN) o ácido ribonucleico (ARN). La información genética es una sucesión de nucleótidos o nucleótidos modificados que representan la estructura principal de los ácidos nucleicos químicos o hipotéticos. En seres humanos, el genoma completo contiene aproximadamente 30.000 genes localizados en veinticuatro (24) cromosomas (véase The Human Genome, T. Strachan, BIOS Scientific Publishers, 1992). Cada gen codifica una proteína específica que, después de la expresión a través de transcripción y traducción, cumple una función bioquímica específica dentro de una célula viva.

Muchas afecciones médicas están provocadas por una o más variaciones genéticas. Determinadas variaciones genéticas provocan afecciones médicas que incluyen, por ejemplo, hemofilia, talasemia, distrofia muscular de Duchenne (DMD), enfermedad de Huntington (EH), enfermedad de Alzheimer y fibrosis quística (FQ) (Human Genome Mutations, D. N. Cooper y M. Krawczak, BIOS Publishers, 1993). Tales enfermedades genéticas pueden resultar de una adición, sustitución o deleción de un solo nucleótido en el ADN de un gen particular. Determinados defectos congénitos están provocados por una anomalía cromosómica, denominada además aneuploidía, tal como trisomía 21 (síndrome de Down), trisomía 13 (síndrome de Patau), trisomía 18 (síndrome de Edward), monosomía X (síndrome de Turner) y determinadas aneuploidías en cromosomas sexuales, tales como síndrome de Klinefelter (XXY), por ejemplo. Otra variación genética es el sexo del feto, que puede determinarse a menudo basándose en los cromosomas sexuales X e Y. Algunas variaciones genéticas pueden predisponer a un individuo a, o provocar, cualquiera de varias enfermedades tales como, por ejemplo, diabetes, arteriosclerosis, obesidad, diversas enfermedades autoinmunitarias y cáncer (por ejemplo, colorrectal, de mama, de ovario, de pulmón).

La identificación de una o más varianzas o variaciones genéticas puede conducir al diagnóstico de, o determinar la predisposición a, una afección médica particular. La identificación de una varianza genética puede dar como resultado que se facilite una decisión médica y/o se emplee un procedimiento médico útil. En algunos casos, la identificación de una o más varianzas o variaciones genéticas implica el análisis del ADN libre de células.

El ADN libre de células (ADNlc) se compone de fragmentos de ADN que se originan a partir de la muerte celular y circulan en la sangre periférica. Las altas concentraciones de ADNlc pueden ser indicativas de determinadas afecciones clínicas, tales como cáncer, traumatismo, quemaduras, infarto de miocardio, accidente cerebrovascular, septicemia, infección y otras enfermedades. Adicionalmente, el ADN fetal libre de células (ADNflc) puede detectarse en el torrente sanguíneo materno y usarse para diversos diagnósticos prenatales no invasivos.

La presencia de ácido nucleico fetal en el plasma materno permite el diagnóstico prenatal no invasivo a través del análisis de una muestra de sangre materna. Por ejemplo, las anomalías cuantitativas del ADN fetal en el plasma materno pueden asociarse con varios trastornos asociados con el embarazo, incluyendo preeclampsia, parto prematuro, hemorragia prenatal, placentación invasiva, síndrome de Down fetal y otros aneuploidías cromosómicas fetales. Por tanto, el análisis de ácidos nucleicos fetales en plasma materno puede ser un mecanismo útil para la monitorización del bienestar fetomaterno.

La detección precoz de afecciones relacionadas con el embarazo, incluyendo complicaciones durante el embarazo y defectos genéticos del feto, es importante, ya que permite la intervención médica temprana necesaria para la seguridad tanto de la madre como del feto. Tradicionalmente, se ha realizado diagnóstico prenatal usando células aisladas del feto a través de procedimientos tales como biopsia de vellosidades coriónicas (BVC) o amniocentesis. Sin embargo, estos métodos convencionales son invasivos y presentan un riesgo apreciable tanto para la madre como para el feto. Actualmente, el Servicio Nacional de Salud cita una tasa de abortos espontáneos de entre el 1 y el 2 por ciento después de las pruebas invasivas de amniocentesis y biopsia de vellosidades coriónicas (BVC). El uso de técnicas de examen no invasivas que utilizan ADNflc circulante puede ser una alternativa a estos enfoques invasivos.

Chu *et al.* (2009), Bioinformatics, 25(10):1244-1250 describe un modelo estadístico para datos de secuenciación de genoma completo, y basándose en este modelo, un método altamente sensible de cariotipado mínimamente invasivo (MINK) para el diagnóstico de una enfermedad genética fetal.

El documento US-2011/177517 proporciona métodos para resolver problemas de medición en la medición del número de copias cromosómicas. Los métodos pueden implicar seleccionar en primer lugar un elemento de ensayo primario característico para la división, en el que la característica puede ser una fuente de variabilidad experimental, tal como el contenido de GC de secuencias de ADN medidas.

Benjamini *et al*. (2012), Nucleic Acid Research, 40(10):1-14 se refiere a medios y métodos para corregir el sesgo del contenido de GC en la secuenciación de alto rendimiento.

Sumario

La presente invención se define mediante las reivindicaciones. Por consiguiente, la presente invención se refiere a un método implementado por ordenador para calcular con sesgo reducido niveles de sección genómica para una muestra de prueba, que comprende: (a) obtener recuentos de lecturas de secuencia mapeadas en porciones de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante, libre de células a partir de una muestra de prueba; (b) determinar un sesgo de guanina y citosina (GC) para cada una de las porciones del genoma de referencia para múltiples muestras a partir de una relación lineal ajustada para cada muestra entre (i) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, y (ii) el contenido de GC para cada una de las porciones, en el que cada sesgo de GC es un coeficiente de sesgo de GC, coeficiente de sesgo de GC que es la pendiente de la relación lineal entre (i) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, y (ii) el contenido de GC para cada una de las porciones; y (c) calcular un nivel de sección genómica para cada una de las porciones del genoma de referencia a partir de una relación lineal ajustada entre (i) el sesgo de GC y (ii) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, proporcionándose de ese modo niveles de sección genómica calculados, en el que el nivel de sección genómica Li se determina para cada una de las porciones del genoma de referencia según la ecuación α: Li = (mi - GiS) 1-1 (ecuación α), en la que Gi es el sesgo de GC, I es la ordenada en el origen de la relación ajustada en (c), S es la pendiente de la relación en (c), mi es los recuentos medidos mapeados en cada porción del genoma de referencia e i es una muestra, mediante lo cual el sesgo en los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia se reduce en los niveles de sección genómica calculados.

La presente invención también se refiere a un sistema que comprende uno o más procesadores y memoria, memoria que comprende instrucciones ejecutables por el uno o más microprocesadores y memoria que comprende recuentos de lecturas de secuencia mapeadas en porciones de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante, libre de células de una muestra de prueba; e instrucciones ejecutables por el uno o más procesadores que están configuradas para: (b) determinar un sesgo de guanina y citosina (GC) para cada una de las porciones del genoma de referencia para múltiples muestras a partir de una relación lineal ajustada para cada muestra entre (i) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, y (ii) el contenido de GC para cada una de las porciones, en el que cada sesgo de GC es un coeficiente de sesgo de GC, coeficiente de sesgo de GC que es la pendiente de la relación lineal entre (i) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, y (ii) el contenido de GC para cada una de las porciones; y (c) calcular un nivel de sección genómica para cada una de las porciones del genoma de referencia a partir de una relación lineal ajustada entre (i) el sesgo de GC y (ii) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, proporcionándose de ese modo niveles de sección genómica calculados, en el que el nivel de sección genómica Li se determina para cada una de las porciones del genoma de referencia según la ecuación α: Li = (mi - GiS) 1-1 (ecuación α), en el que Gi es el sesgo de GC, I es la ordenada en el origen de la relación ajustada en (c), S es la pendiente de la relación en (c), mi es los recuentos medidos mapeados en cada porción del genoma de referencia e i es una muestra, mediante lo cual el sesgo en los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia se reduce en los niveles de sección genómica calculados.

Según una realización preferida del método y el sistema de la invención, cada una de la relación ajustada de (b) y la relación ajustada de (c) se ajustan independientemente mediante una regresión lineal.

Según una realización preferida del sistema de la invención, las instrucciones ejecutables por el uno o más procesadores están configuradas para determinar la presencia o ausencia de una aneuploidía cromosómica fetal para la muestra de prueba según los niveles de sección genómica calculados. De manera similar, según una realización preferida del método de la invención, el método comprende determinar la presencia o ausencia de una aneuploidía cromosómica fetal para la muestra de prueba según los niveles de sección genómica calculados.

Según una realización preferida del método y el sistema de la invención, la aneuploidía cromosómica fetal es una trisomía. Esta trisomía se elige preferiblemente de una trisomía del cromosoma 21, el cromosoma 18, el cromosoma 13 o combinación de los mismos.

Según una realización preferida del método de la invención, el método comprende, antes de (b), calcular una medida de error para los recuentos de lecturas de secuencia mapeadas en algunas o todas las porciones del genoma de referencia y eliminar o ponderar los recuentos de lecturas de secuencia para determinadas porciones del genoma de referencia según un umbral de la medida de error. Según una realización preferida del método de la invención. Según una realización más preferida del método de la invención. el umbral se selecciona según una brecha de desviación estándar entre un primer nivel de sección genómica y un segundo nivel de sección genómica de 3,5 o mayor. Según otra realización más preferida del método de la invención, la medida de error es un factor R. Según una realización

incluso más preferida del método de la invención, los recuentos de lecturas de secuencia para una porción del genoma de referencia que tiene un factor R de aproximadamente el 7 % a aproximadamente el 10 % se eliminan antes de (b).

También se proporciona en el presente documento un método para detectar la presencia o ausencia de una aberración cromosómica segmentaria o una aneuploidía fetal o ambas que incluye: (a) obtener de un sujeto de prueba una muestra que incluye ácido nucleico circulante, libre de células; (b) aislar el ácido nucleico de muestra libre de células de la muestra; (c) obtener lecturas de secuencia del ácido nucleico de muestra libre de células; (d) mapear las lecturas de secuencia obtenidas en (c) en un genoma conocido, genoma conocido que se ha dividido en secciones genómicas; (e) contar las lecturas de secuencia mapeadas dentro de las secciones genómicas; (f) generar un perfil de recuento normalizado de muestra normalizando los recuentos para las secciones genómicas obtenidas en (e); y (g) determinar la presencia o ausencia de una aberración cromosómica segmentaria o una aneuploidía fetal o ambas a partir del perfil de recuento normalizado de muestra en (f). La expresión "genoma conocido", tal como se usa en el presente documento con respecto al mapeo de lecturas de secuencia, se refiere a un genoma de referencia o mapeo o segmentos del mismo (por ejemplo, genoma intacto, uno o más cromosomas, porciones de cromosomas, segmentos o secciones genómicos seleccionados, similares o combinaciones de los anteriores).

También se proporciona en el presente documento un método para detectar la presencia o ausencia de una variación genética, que incluye: (a) obtener de un sujeto de prueba una muestra que incluye ácido nucleico; (b) aislar ácido nucleico de muestra de la muestra; (c) obtener lecturas de secuencia del ácido nucleico de muestra; (d) mapear las lecturas de secuencia obtenidas en (c) en un genoma conocido, genoma conocido que se ha dividido en secciones genómicas; (e) contar las lecturas de secuencia mapeadas dentro de las secciones genómicas; (f) generar un perfil de recuento normalizado de muestra normalizando los recuentos para las secciones genómicas obtenidas en (e); y (g) determinar la presencia o ausencia de una variación genética del perfil de recuento normalizado de muestra en (f).

En algunas implementaciones, el sujeto de prueba se elige de un ser humano, un animal y una planta. En determinadas realizaciones, un sujeto de prueba humano incluye una mujer, una mujer embarazada, un hombre, un feto o un recién nacido. En algunas realizaciones, (f) incluye ponderar los recuentos para las secciones genómicas obtenidas en (e) usando la inversa de la desviación estándar al cuadrado.

También se proporciona en el presente documento un método para detectar la presencia o ausencia de una aberración cromosómica segmentaria o una aneuploidía fetal o ambas, que incluye: (a) obtener lecturas de secuencia de ácido nucleico de muestra circulante, libre de células de un sujeto de prueba; (b) mapear las lecturas de secuencia obtenidas en (a) en un genoma conocido, genoma conocido que se ha dividido en secciones genómicas; (c) contar las lecturas de secuencia mapeadas dentro de las secciones genómicas; (d) generar un perfil de recuento normalizado de muestra normalizando los recuentos para las secciones genómicas obtenidas en (c); y (e) determinar la presencia o ausencia de una aberración cromosómica segmentaria o una aneuploidía fetal o ambas a partir del perfil de recuento normalizado de muestra en (d).

También se proporciona en el presente documento un método para detectar la presencia o ausencia de una variación genética, que incluye: (a) obtener lecturas de secuencia de ácido nucleico de muestra de un sujeto de prueba; (b) mapear las lecturas de secuencia obtenidas en (a) en un genoma conocido, genoma conocido que se ha dividido en secciones genómicas; (c) contar las lecturas de secuencia mapeadas dentro de las secciones genómicas; (d) generar un perfil de recuento normalizado de muestra normalizando los recuentos para las secciones genómicas obtenidas en (c); y (e) determinar la presencia o ausencia de una variación genética del perfil de recuento normalizado de muestra en (d).

En algunas implementaciones, el ácido nucleico de muestra libre de células se aísla de sangre obtenida del sujeto de prueba. En determinadas implementaciones, el ácido nucleico de muestra libre de células se aísla de suero obtenido del sujeto de prueba y, en algunas implementaciones, el ácido nucleico de muestra libre de células se aísla de plasma obtenido del sujeto de prueba. En determinadas realizaciones, el sujeto de prueba se elige de un ser humano, un animal y una planta. En algunas implementaciones, un sujeto de prueba humano incluye una mujer, una mujer embarazada, un hombre, un feto o un recién nacido. En determinadas implementaciones, (d), incluye ponderar los recuentos para las secciones genómicas obtenidas en (c) usando la inversa de la desviación estándar al cuadrado.

En algunas implementaciones, las lecturas de secuencia del ácido nucleico de muestra libre de células están en forma de fragmentos de polinucleótidos. En determinadas implementaciones, los fragmentos de polinucleótidos tienen entre aproximadamente 20 y aproximadamente 50 nucleótidos de longitud. En algunas implementaciones, los polinucleótidos tienen entre aproximadamente 30 y aproximadamente 40 nucleótidos de longitud. En determinadas implementaciones, el genoma conocido se divide en secciones genómicas que comparten un tamaño común.

En algunas implementaciones, el recuento de las lecturas de secuencia mapeadas dentro de las secciones genómicas (c) se realiza después de eliminar las lecturas de secuencia redundantes mapeadas en las secciones genómicas en (b). En determinadas implementaciones, el perfil de recuento normalizado de muestra se genera normalizando un perfil de recuento sin procesar de muestra a una mediana de perfil de recuento de referencia. En algunas implementaciones, el perfil de recuento sin procesar de muestra se genera construyendo un perfil de recuento medido de muestra que representa la distribución de los recuentos medidos a través del genoma o segmento del mismo. En determinadas implementaciones, el método incluye además normalizar el perfil de

recuento medido de muestra con respecto al número total de recuentos mapeados no redundantes a través del genoma o segmento del mismo, generándose de ese modo el perfil de recuento sin procesar de muestra.

En algunas implementaciones, la mediana de perfil de recuento de referencia se genera mediante un procedimiento que incluye: (i) obtener lecturas de secuencia de ácido nucleico de muestra de referencia circulante, libre de células de múltiples sujetos de referencia; (ii) mapear las lecturas de secuencia obtenidas en (i) en un genoma conocido, genoma conocido que se ha dividido en secciones genómicas; (iii) contar las lecturas de secuencia mapeadas dentro de las secciones genómicas; (iv) generar un perfil de recuento sin procesar a partir del recuento en (iii); (v) eliminar segmentos genómicos con medianas de recuento de valor cero en muestras de referencia; y determinar la mediana de recuento y la incertidumbre para los segmentos genómicos; en el que realizar (i) a (vi) genera una mediana de perfil de recuento de referencia, un perfil de incertidumbre y/o identificadores de segmentos. En determinadas implementaciones, los sujetos de referencia se seleccionan de seres humanos, animales y plantas. En algunas implementaciones, los sujetos de referencia humanos incluyen mujeres, mujeres embarazadas, hombres, fetos o recién nacidos. En determinadas implementaciones, las mujeres embarazadas como sujeto de referencia portan fetos que no tienen aberraciones cromosómicas y/o fetos que se sabe que son euploides. En algunas implementaciones, generar una mediana de perfil de recuento de referencia incluye seleccionar un punto de corte de incertidumbre después de (iii).

En determinadas implementaciones, el punto de corte de incertidumbre se obtiene mediante un procedimiento que incluye: calcular la desviación estándar del perfil generado en (iv) y multiplicar la desviación estándar del perfil por una constante, en el que la constante es equivalente a un intervalo de confianza seleccionado (por ejemplo, 2 desviaciones estándar = 2, 3 desviaciones estándar = 3); generándose de ese modo un valor de para el punto de corte de incertidumbre. En algunas implementaciones, el punto de corte de incertidumbre se obtiene mediante un procedimiento que incluye: calcular la mediana de la desviación absoluta del perfil generado en (iv); y multiplicar la mediana de la desviación absoluta del perfil por una constante, en la que la constante es equivalente a un intervalo de confianza seleccionado; generándose de ese modo un valor de para el punto de corte de incertidumbre. En determinadas implementaciones, se elimina cualquier sección genómica con un valor que supere el punto de corte de incertidumbre. En algunas implementaciones, el método incluye además eliminar segmentos con incertidumbres de recuento que superen un punto de corte de incertidumbre después de (vi). En determinadas implementaciones, se genera una mediana de perfil de recuento de referencia construyendo un perfil de recuento medido de referencia que representa la distribución de recuentos medidos de referencia a través del genoma o segmento del mismo.

En algunas implementaciones, se genera un perfil de recuento normalizado de muestra para cada segmento genómico eliminando los segmentos genómicos del perfil de recuento sin procesar de muestra que se eliminaron del perfil de recuento de muestra de referencia en (v), asignando una incertidumbre generada en (vi) y normalizando los recuentos medidos de muestra para cada segmento restante con respecto a la suma de recuentos de segmentos restantes en la mediana de perfil de recuento de referencia.

En determinadas implementaciones, los picos de perfil de muestra con valor predictivo para detectar aberración segmentaria cromosómica fetal o aneuploidía fetal o ambas se identifican en una ubicación en el genoma mediante un procedimiento que incluye: seleccionar un nivel de confianza en el cual evaluar el perfil de recuento normalizado generado en (iv), perfil de recuento normalizado que incluye picos; seleccionar una longitud máxima de segmento genómico a lo largo de la cual evaluar los picos; y evaluar elevaciones de pico y/o anchura de pico para segmentos genómicos de diversas longitudes en una ubicación en el genoma, en los que los picos con valor predictivo para detectar aberración segmentaria cromosómica fetal o aneuploidía fetal o ambas se detectan con el nivel de confianza en la ubicación en el genoma. En algunas implementaciones, el nivel de confianza seleccionado es del 95 %. En determinadas realizaciones, el nivel de confianza seleccionado es del 99 %. En algunas realizaciones, el nivel de confianza se selecciona basándose en la calidad de los recuentos medidos. En determinadas realizaciones, la longitud máxima del segmento genómico a lo largo de la cual evaluar los picos incluye uno o más segmentos genómicos o porciones de los mismos.

En algunas implementaciones, el método incluye además: seleccionar una ubicación en el genoma; generar un perfil de valor de p que incluye picos; eliminar segmentos genómicos con valores de p por debajo del nivel de confianza seleccionado; eliminar segmentos redundantes y/o solapantes de longitudes diferentes; determinar las ubicaciones de borde de pico y sus incertidumbres asociadas; e identificar y opcionalmente eliminar los picos que se encuentran habitualmente entre muestras seleccionadas aleatoriamente, en el que los picos con valor predictivo para detectar aberración segmentaria cromosómica fetal o aneuploidía fetal o ambas se detectan dentro de una ubicación en el genoma. En algunas implementaciones, se eliminan algunos de los segmentos redundantes y/o solapantes de diferentes longitudes. En determinadas implementaciones, se eliminan todos los segmentos redundantes y/o solapantes de diferentes longitudes.

En algunas implementaciones, se genera un perfil de valor de p mediante un procedimiento que incluye: seleccionar una ubicación deseada en el genoma para su evaluación; seleccionar una longitud deseada del segmento genómico; evaluar la elevación de perfil promedio para determinar la ubicación en el genoma y el error asociado de la media en el perfil de recuento normalizado de muestra; y asignar un valor de p a los segmentos genómicos seleccionados, en el que se genera un perfil de valor de p. En determinadas implementaciones, los valores de p asignados a los segmentos genómicos seleccionados

$$t = \frac{(x_1) - (x_2)}{\sqrt{\frac{\sigma_1^2}{n_1} - \frac{\sigma_2^2}{n_2}}}$$

se calculan según la fórmula, en la que x_1 y x_2 representan valores promedio, n_1 y n_2 representan tamaños de muestra, y σ_1 y σ_2 representan la desviación estándar.

En algunas implementaciones, asignar un valor de p a los segmentos genómicos seleccionados incluye además: (1) seleccionar un segmento inicial; (2) determinar la elevación promedio y el error estándar de la media para la ubicación seleccionada en el genoma; (3) evaluar la elevación de segmento promedio y el error estándar correspondiente de la media; (4) evaluar el valor de Z en relación con la elevación promedio para la ubicación seleccionada en el genoma y/o en relación con un valor de elevación predeterminado; (5) repetir las etapas 1-4 para uno o más segmentos iniciales y/o longitudes de segmento; y (6) realizar una prueba de la t a lo largo de toda la longitud de segmento de cada uno de los segmentos iniciales y/o longitudes de segmento seleccionados, en el que se asigna un valor de p al segmento genómico seleccionado. En determinadas

$$Z = \frac{\Delta_1 - \Delta_2}{\sqrt{\sigma_1^2\left(\frac{1}{N_1} + \frac{1}{n_1}\right) + \sigma_2^2\left(\frac{1}{N_2} + \frac{1}{n_2}\right)}}$$

implementaciones, los valores de Z se calculan por medio del uso de la fórmula anterior, en la que N y n se refieren a los números de bins en todo el cromosoma y dentro de la aberración, σ_1 y σ_2 representan la desviación estándar, y Δ_1 representa la diferencia entre la elevación promedio de una región de variación genética para el sujeto 1 y la elevación promedio del cromosoma en el que está la región para el sujeto 1 y Δ_2 representa la diferencia entre la elevación promedio de una región de variación genética para el sujeto 2 y la elevación promedio del cromosoma en el que está la región para el sujeto 2. El término "diferencia", tal como se usa en el presente documento, con respecto a funciones matemáticas y/o estadísticas, se refiere a una resta matemática entre dos o más valores. En determinadas implementaciones, el valor de elevación predeterminado es igual a 1. En algunas implementaciones, el valor de elevación predeterminado es menor de 1. En determinadas implementaciones, el valor de elevación predeterminado es mayor de 1. En algunas implementaciones, el método incluye una corrección opcional para la autocorrelación.

En determinadas implementaciones, los picos hallados habitualmente se identifican mediante un procedimiento que incluye: obtener lecturas de ácido nucleico de muestra libre de células a partir de múltiples muestras medidas en las mismas condiciones o en condiciones similares; seleccionar un conjunto de muestras de prueba; generar una mediana de perfil de recuento de referencia que incluye picos; e identificar picos hallados en común entre las muestras en el conjunto de muestras de prueba. En algunas implementaciones, las múltiples muestras se seleccionan aleatoriamente. En determinadas implementaciones, la identificación de los picos hallados en común entre las muestras de prueba incluye: comparar la mediana de perfiles de recuento de referencia que incluyen picos, perfiles de valores de Z que incluyen picos, perfiles de valor de p que incluyen picos o combinaciones de los mismos, e identificar los picos identificados habitualmente en cada muestra. En determinadas implementaciones, el método incluye determinar ubicaciones de borde de pico, tolerancias laterales de pico e incertidumbres asociadas mediante un procedimiento que incluye: seleccionar una o más regiones en un perfil de recuento normalizado de muestra que incluye picos y/o mediana de perfil de recuento de referencia que incluye picos; determinar la primera derivada del perfil normalizado y/o sus potencias; y caracterizar picos de derivada, en el que el procedimiento genera máximos de pico de derivada y anchuras de pico de derivada con valor predictivo para detectar aberración segmentaria cromosómica fetal o aneuploidía fetal o ambas.

También se proporciona en el presente documento un método para determinar si dos muestras son del mismo donante, incluyendo el método: obtener lecturas de secuencia de ácido nucleico de muestra circulante, libre de células de las muestras de uno o más donantes; mapear las lecturas de secuencia obtenidas en un genoma conocido, genoma conocido que se ha dividido en secciones genómicas; contar las lecturas de secuencia mapeadas dentro de secciones genómicas; generar perfiles de recuento normalizados que incluyen picos; identificar picos de perfil de recuento normalizado con valor predictivo en cada muestra; comparar picos en una muestra con los picos de otra muestra; evaluar la probabilidad conjunta basándose en pares de picos coincidentes; determinar la probabilidad de que las muestras procedan del mismo donante, en el que se realiza una determinación con respecto a la probabilidad de que las muestras procedan del mismo donante. En algunas implementaciones, el método incluye además comparar los picos en una muestra con los picos en otra muestra usando uno o más de los siguientes procedimientos: determinar si los bordes de los picos coinciden dentro de sus tolerancias laterales usando anchuras de pico de derivada; determinar si las elevaciones de pico coinciden dentro de sus errores estándar de la media usando máximos de pico de derivada; ajustar los valores de p para la prevalencia de población de un pico dado, en el que se realiza una determinación de si las muestras proceden del mismo donante mediante la realización de uno o más de los procedimientos. En determinadas implementaciones, determinar si las elevaciones de pico coinciden dentro de sus errores estándar de la media incluye además usar

$$t = \frac{(x_1) - (x_2)}{\sqrt{\frac{\sigma_1^2}{n_1} - \frac{\sigma_2^2}{n_2}}}$$

una prueba de la t. En algunas implementaciones, se calcula una prueba de la t según la fórmula, en la que x_1 y x_2 representan valores promedio, n_1 y n_2 representan tamaños de muestra y σ_1 y σ_2 representan la desviación estándar.

También se proporciona en el presente documento un método para clasificar una muestra como euploide o aneuploide usando la mediana de elevaciones del perfil de recuento que incluye: obtener una muestra de un sujeto de prueba que incluye ácido nucleico circulante, libre de células; aislar ácido nucleico de muestra libre de células de la muestra; obtener lecturas de secuencia del ácido nucleico de muestra libre de células aislado; mapear las lecturas de secuencia obtenidas en un genoma conocido, genoma conocido que se ha dividido en secciones genómicas; contar las lecturas de secuencia mapeadas dentro de las secciones genómicas; obtener, a partir de las lecturas de secuencia mapeadas contadas, un perfil de recuento normalizado que incluye la mediana de elevaciones seleccionadas de secciones genómicas del perfil de recuento y una incertidumbre asociada; seleccionar una ubicación en el genoma para la evaluación; evaluar la mediana de elevación de perfil y la incertidumbre asociada para una ubicación en el genoma; y determinar si la mediana de elevación supera significativamente un valor predeterminado, en el que determinar si la mediana de elevación determina significativamente el valor predeterminado si la muestra es euploide o aneuploide. En algunas implementaciones, el valor predeterminado es igual a 1. En determinadas implementaciones, el valor predeterminado es menor de 1. En algunas implementaciones, el valor predeterminado es mayor de 1. En determinadas implementaciones, el método incluye identificar elevaciones de pico del perfil de recuento normalizado con valor predictivo dentro de una ubicación en el genoma y corregir deleciones y/o duplicaciones, si se identifican, antes de evaluar la mediana de elevación de perfil y la incertidumbre asociada para una ubicación en el genoma.

También se proporciona en el presente documento un método para clasificar una muestra como euploide o aneuploide usando razones de áreas de pico con valor predictivo que incluye: obtener una muestra de un sujeto de prueba que incluye ácido nucleico circulante, libre de células; aislar ácido nucleico de muestra libre de células de la muestra; obtener lecturas de secuencia del ácido nucleico de muestra libre de células aislado; mapear las lecturas de secuencia, en un genoma conocido, genoma conocido que se ha dividido en secciones genómicas; contar las lecturas de secuencia mapeadas dentro de las secciones genómicas; obtener un perfil de recuento normalizado que incluye una distribución de recuentos para una sección genómica seleccionada; seleccionar una ubicación en el genoma para evaluación; evaluar la ubicación seleccionada para detectar picos con valor predictivo y las razones de área asociadas para los picos; y determinar si la razón de área para un pico es significativamente diferente con respecto a un valor predeterminado, en el que determinar si las razones de área para un pico supera significativamente el valor predeterminado determina si la muestra es euploide o aneuploide. En algunas implementaciones, el valor predeterminado es igual a 1. En determinadas implementaciones, el valor predeterminado es menor de 1. En algunas implementaciones, el valor predeterminado es mayor de 1. En determinadas implementaciones, el método incluye identificar razones de área de pico dentro de una ubicación en el genoma y corregir deleciones y/o duplicaciones, si se identifican, antes de evaluar la razón de área de pico con valor predictivo para una ubicación en el genoma.

También se proporciona en el presente documento un método para clasificar una muestra como euploide o aneuploide combinando múltiples criterios de clasificación, incluyendo el método: obtener de un sujeto de prueba y múltiples sujetos de referencia euploides conocidos a partir de una muestra que incluye ácido nucleico circulante, libre de células; aislar ácido nucleico de muestra libre de células de la muestra; obtener lecturas de secuencia a partir del ácido nucleico de muestra libre de células aislado; mapear las lecturas de secuencia obtenidas en un genoma conocido, genoma conocido que se ha dividido en secciones genómicas; contar las lecturas de secuencia mapeadas dentro de las secciones genómicas; obtener un perfil de recuento normalizado a partir del recuento para los sujetos de prueba y referencia; seleccionar una ubicación en el genoma para evaluación; evaluar la ubicación seleccionada en el genoma de la referencia euploide usando múltiples criterios de clasificación; determinar el espacio N-dimensional mínimo poblado exclusivamente por euploides; evaluar una ubicación en el genoma del sujeto de prueba usando múltiples criterios de clasificación; y determinar si el punto N-dimensional para el sujeto de prueba se encuentra dentro del espacio exclusivamente poblado por euploides, en el que determinar si el punto N-dimensional para el sujeto de prueba se encuentra dentro del espacio exclusivamente poblado por euploides determina si el sujeto de prueba es euploide o aneuploide.

En algunas implementaciones, el espacio N-dimensional para euploides y el punto N-dimensional para el sujeto de prueba se evalúan usando uno o más criterios de clasificación seleccionados de mediana de elevación de perfil, razón de área, valores de Z, ploidía ajustada, fracción fetal ajustada, sumas de residuos al cuadrado y valores de p bayesianos. En determinadas implementaciones, la obtención de lecturas de secuencia incluye someter el ácido nucleico de muestra libre de células a un procedimiento de secuenciación de ácidos nucleicos. En algunas implementaciones, el procedimiento de secuenciación incluye un método seleccionado de secuenciación de alto rendimiento, secuenciación por nanoporos, secuenciación por síntesis, pirosecuenciación, secuenciación basada en ligamiento, secuenciación basada en celdas de flujo, secuenciación basada en semiconductores, secuenciación de una sola molécula basada en microscopia electrónica, secuenciación por PCR, secuenciación de didesoxi o

combinaciones de las mismas. En determinadas implementaciones, determinar la presencia o ausencia de una aberración cromosómica segmentaria o una aneuploidía fetal o ambas incluye, proporcionar un gráfico del resultado, un informe del resultado, un archivo electrónico que incluye el resultado, una representación bidimensional del resultado, una representación tridimensional del resultado, o combinaciones de los mismos, a un profesional sanitario. En algunas implementaciones, el profesional sanitario proporciona una recomendación basada en el resultado proporcionado. En algunas implementaciones, el ácido nucleico de muestra, el ácido nucleico de muestra de referencia o ambos son ácido nucleico libre de células. En determinadas implementaciones, el ácido nucleico libre de células es ácido nucleico circulante, libre de células. En algunas implementaciones, una variación genética es determinante de una afección médica.

También se proporciona en el presente documento un producto de programa informático, incluyendo un medio utilizable por ordenador que tiene un código de programa legible por ordenador incorporado en el mismo, incluyendo el código de programa legible por ordenador distintos módulos de software que incluyen un módulo de procesamiento lógico, un módulo de secuenciación y un módulo de organización de visualización de datos, estando el código de programa legible por ordenador adaptado para ejecutarse para implementar un método para identificar la presencia o ausencia de una aberración cromosómica segmentaria o una aneuploidía fetal o ambas, incluyendo el método: (a) obtener, por el módulo de secuenciación, lecturas de secuencia de ácido nucleico de muestra circulante, libre de células de un sujeto de prueba; (b) mapear, por el módulo de procesamiento lógico, las lecturas de secuencia obtenidas en (a) en un genoma conocido, genoma conocido que se ha dividido en secciones genómicas; (c) contar, por el módulo de procesamiento lógico, las lecturas de secuencia mapeadas dentro de las secciones genómicas; (d) generar, por el módulo de procesamiento lógico, un perfil de recuento normalizado de muestra normalizando los recuentos para las secciones genómicas obtenidas en (c); (e) proporcionar, por el módulo de procesamiento lógico, una determinación de la presencia o ausencia de una aberración cromosómica segmentaria o una aneuploidía fetal o ambas a partir del perfil de recuento normalizado de muestra en (d); y (f) organizar, por el módulo de organización de visualización de datos en respuesta a que se determina por el módulo de procesamiento lógico, una visualización de datos que indica la presencia o ausencia de una aberración cromosómica segmentaria o una aneuploidía fetal o ambas. También se proporciona en el presente documento un aparato, que incluye una memoria en la que se almacena un producto de programa informático descrito en el presente documento. En algunas implementaciones, el aparato incluye un procesador que implementa una o más funciones del producto de programa informático especificado en el presente documento.

También se proporciona en el presente documento un sistema que incluye un aparato de secuenciación de ácido nucleico y un aparato de procesamiento, en el que el aparato de secuenciación obtiene lecturas de secuencia de una muestra, y el aparato de procesamiento obtiene las lecturas de secuencia del dispositivo de secuenciación y lleva a cabo un método que incluye: (a) obtener lecturas de secuencia de ácido nucleico de muestra circulante, libre de células de un sujeto de prueba; (b) mapear las lecturas de secuencia obtenidas en (a) en un genoma conocido, genoma conocido que se ha dividido en secciones genómicas; (c) contar las lecturas de secuencia mapeadas dentro de las secciones genómicas; (d) generar un perfil de recuento normalizado de muestra normalizando los recuentos para las secciones genómicas obtenidas en (c); y (e) determinar la presencia o ausencia de una aberración cromosómica segmentaria o una aneuploidía fetal o ambas a partir del perfil de recuento normalizado de muestra en (d).

También se proporciona en el presente documento un método para determinar la ploidía fetal, que incluye: (a) generar un perfil de recuento sin procesar basándose en lecturas de secuencia de ácidos nucleicos circulantes, libres de células obtenidos a partir de una muestra de un sujeto de prueba; (b) generar una mediana de perfil de recuento de referencia basándose en lecturas de secuencia de ácidos nucleicos circulantes, libres de células obtenidos de muestras de uno o más sujetos de referencia; (c) generar un perfil de recuento normalizado a partir de (a) con respecto a los recuentos totales de las lecturas de secuencia de sujeto de prueba; (d) generar un perfil de recuento normalizado a partir de (b) con respecto a los recuentos totales de la una o más lecturas de secuencia de sujeto de referencia; (e) calcular la suma de los residuos al cuadrado basándose en parte en perfiles de recuento normalizados y una o más suposiciones elegidas de ploidía fija o ploidía optimizada, y fracción fetal fija o fracción fetal optimizada; y (f) determinar la ploidía fetal basándose en la suma de los residuos al cuadrado en (e). En algunas implementaciones, el sujeto de prueba y/o uno o más sujetos de referencia se eligen de un ser humano, un animal y una planta. En determinadas implementaciones, un sujeto de prueba humano y/o uno o más sujetos de referencia incluye una mujer, una mujer embarazada, un hombre, un feto o un recién nacido.

En algunas implementaciones, el ácido nucleico de muestra libre de células se aísla de sangre obtenida de los sujetos de prueba y/o referencia. En determinadas implementaciones, el ácido nucleico de muestra libre de células se aísla de suero obtenido de los sujetos de prueba y/o referencia. En algunas implementaciones, el ácido nucleico de muestra libre de células se aísla de plasma obtenido de los sujetos de prueba y/o referencia.

En determinadas implementaciones, el método incluye además calcular la suma de los residuos al cuadrado en (e) usando un valor para la fracción fetal medida, en el que el valor de ploidía fija no es igual a 1. En algunas implementaciones, determinar la ploidía fetal basada en el valor numérico de la suma de residuos al cuadrado permite la clasificación de un feto como euploide o triploide. En determinadas implementaciones, la fracción fetal fija es una fracción fetal medida. En algunas implementaciones, (c), (d) o (c) y (d) incluyen ponderar los recuentos para las secciones genómicas generadas en (a), (b) o (a) y (b) usando la inversa de la desviación estándar al cuadrado.

En determinadas implementaciones, (a) incluye: (i) obtener lecturas de secuencia de ácido nucleico de muestra circulante, libre de células de un sujeto de prueba; (ii) mapear las lecturas de secuencia obtenidas en (i) en un genoma conocido, genoma conocido que se ha dividido en secciones genómicas; (iii) contar las lecturas de secuencia mapeadas dentro de las secciones genómicas; (iv) construir un perfil de recuento medido de muestra que representa la distribución de recuentos medidos a través del genoma o segmento del mismo; y (v) normalizar el perfil de recuento medido de muestra a partir de la muestra del sujeto de prueba con respecto al número total de recuentos mapeados no redundantes a través del genoma o segmento del mismo, generándose de ese modo el perfil de recuento sin procesar de muestra. En algunas implementaciones, (iii) se realiza después de eliminar las lecturas de secuencia redundantes mapeadas en las secciones genómicas en (ii).

En algunas implementaciones, (b) incluye: (1) obtener lecturas de secuencia de ácido nucleico de muestra de referencia circulante, libre de células de uno o más sujetos de referencia que se sabe que son euploides; (2) mapear las lecturas de secuencia obtenidas en (1) en un genoma conocido, genoma conocido que se ha dividido en secciones genómicas; (3) contar las lecturas de secuencia mapeadas dentro de las secciones genómicas; (4) generar un perfil de recuento sin procesar a partir del recuento en (2); (5) eliminar segmentos genómicos con mediana de recuentos de valor cero en las muestras de referencia; (6) determinar la mediana de recuento y la incertidumbre para las secciones genómicas; y (7) normalizar la mediana de recuento con respecto a la suma de recuentos en las secciones restantes, en el que realizar (1) a (7) genera una mediana de perfil de recuento de referencia, un perfil de incertidumbre y/o identificadores de segmentos. En algunas implementaciones, las lecturas de secuencia del ácido nucleico libre de células están en forma de fragmentos de polinucleótidos. En determinadas implementaciones, los fragmentos de polinucleótidos tienen entre aproximadamente 20 a aproximadamente 50 nucleótidos de longitud. En algunas implementaciones, los fragmentos de polinucleótidos tienen entre aproximadamente 30 y aproximadamente 40 nucleótidos de longitud. En determinadas implementaciones, el genoma conocido se divide en segmentos genómicos que comparten un tamaño común.

En algunas implementaciones, el método incluye seleccionar un punto de corte de incertidumbre después de (4). En determinadas implementaciones, el punto de corte de incertidumbre se obtiene mediante un procedimiento que incluye: calcular la desviación estándar del perfil generado en (4); y multiplicar la desviación estándar del perfil por 3, generándose de ese modo un valor para el punto de corte de incertidumbre. En algunas implementaciones, el punto de corte de incertidumbre se obtiene mediante un procedimiento que incluye: calcular la mediana de la desviación absoluta del perfil generado en (4) y multiplicar la mediana de la desviación absoluta del perfil por 3, generándose de ese modo un valor para el punto de corte de incertidumbre. En determinadas implementaciones, el método incluye eliminar segmentos con incertidumbres de recuento que superan un punto de corte de incertidumbre después de (7).

En algunas implementaciones, la mediana de perfil de recuento de referencia se genera construyendo un perfil de recuento medido de referencia que representa la distribución de recuentos medidos de referencia a través del genoma o segmento del mismo. En determinadas implementaciones, se genera un perfil de recuento normalizado para cada segmento genómico eliminando los segmentos genómicos del perfil de recuento sin procesar de muestra que se eliminaron del perfil de recuento de muestra de referencia en (5), asignando una incertidumbre generada en (6) y normalizando los recuentos medidos de muestra para cada segmento restante con respecto a la suma de los recuentos de segmentos restantes en la mediana de perfil de recuento de referencia. En determinadas implementaciones, obtener lecturas de secuencia de ácido nucleico de muestra circulante, libre de células incluye: obtener de un sujeto una muestra que incluye ácido nucleico circulante, libre de células; y aislar ácido nucleico de muestra libre de células de la muestra; en el que la muestra obtenida del sujeto incluye sangre, suero, plasma o una combinación de los mismos.

En determinadas implementaciones, evaluar la suma de los residuos al cuadrado incluye: calcular el resultado numérico de la fórmula $\Xi_{fy} = \sum_{i=1}^{N} \frac{y_i f_i}{\sigma_i^2}$; calcular el resultado numérico de la fórmula $\Xi_{ff} = \sum_{i=1}^{N} \frac{f_i^2}{\sigma_i^2}$; calcular el resultado numérico para phi usando la fórmula $\varphi = \varphi_E - \varphi_T = F\left(\Xi_{fy} - \Xi_{ff}\right) - \frac{1}{4}F^2\,\Xi_{ff}$ usar los valores numéricos de $\Xi_{fy} = \sum_{i=1}^{N} \frac{y_i f_i}{\sigma_i^2}$ $\Xi_{ff} = \sum_{i=1}^{N} \frac{f_i^2}{\sigma_i^2}$ e ; y determinar si phi es menor o mayor que el valor predeterminado, en el que phi representa la diferencia entre sumas de residuos al cuadrado evaluada suponiendo un resultado euploide o de trisomía, respectivamente, *f* representa la mediana de perfil de recuento de referencia, épsilon representa el perfil de recuento medido normalizado con respecto a los recuentos totales, *F* representa la fracción fetal, *N* representa el número total de secciones genómicas, i representa una sección genómica seleccionada, sigma (σ) representa la incertidumbre asociada con *f* para una sección genómica seleccionada, y en el que una determinación euploide o no euploide basada en el valor numérico de phi. En algunas implementaciones, la fracción fetal es una fracción fetal medida. En determinadas implementaciones, el valor predeterminado es igual a 0. En algunas implementaciones, el valor predeterminado es mayor de 0. En determinadas implementaciones, el valor predeterminado es menor de 0.

En algunas implementaciones, la ploidía fetal optimizada incluye: calcular el resultado numérico de la fórmula $\Xi_{fy} = \sum_{i=1}^{N} \frac{y_i f_i}{\sigma_i^2}$; calcular el resultado numérico de la fórmula $\Xi_{ff} = \sum_{i=1}^{N} \frac{f_i^2}{\sigma_i^2}$; calcular el resultado numérico para ploidía (por ejemplo, *X*) usando la fórmula $X = \frac{\Xi_{fy} - (1-F)\Xi_{ff}}{F\,\Xi_{ff}} = \frac{\Xi_{fy}}{F\,\Xi_{ff}} - \frac{1-F}{F} = 1 + \frac{1}{F}\left(\frac{\Xi_{fy}}{\Xi_{ff}} - 1\right)$; usando los valores numéricos de $\Xi_{fy} = \sum_{i=1}^{N} \frac{y_i f_i}{\sigma_i^2}$ y $\Xi_{ff} = \sum_{i=1}^{N} \frac{f_i^2}{\sigma_i^2}$; y determinar si X es menor o mayor que un valor predeterminado; en el que *f* representa la mediana de perfil de recuento de referencia, *y* representa el perfil de recuento medido normalizado con respecto a los recuentos totales, *F* representa la fracción fetal, *N* representa el número total de secciones genómicas, i representa una sección genómica seleccionada, sigma (σ) representa la incertidumbre asociada con f para una sección genómica seleccionada, épsilon es un número positivo usado como punto de corte para distinguir muestras triploides de las euploides, y en el que se realiza una determinación de euploide o no euploide basándose en el valor numérico de *X*. En determinadas implementaciones, el valor predeterminado es (1+épsilon). En algunas implementaciones, *X* es mayor de (1+épsilon). En determinadas implementaciones, *X* es menor de (1+épsilon). En algunas implementaciones, *X* es igual a (1+épsilon).

En determinadas implementaciones, la fracción fetal optimizada incluye: calcular el resultado numérico de la fórmula $S_{ff} = \frac{(\Delta F)^2}{4} \sum_{i=1}^{N} \frac{f_i^2}{\sigma_i^2}$; calcular el resultado numérico de la fórmula $S_{fy} = \frac{(\Delta F)^2}{4} \sum_{i=1}^{N} \frac{y_i f_i}{\sigma_i^2}$; calcular el resultado numérico para ploidía (por ejemplo, *X)* usando la fórmula $F = \frac{F_0 + 2S_{fy} - 2S_{ff}}{1 + S_{ff}}$; usando los valores numéricos de $S_{ff} = \frac{(\Delta F)^2}{4} \sum_{i=1}^{N} \frac{f_i^2}{\sigma_i^2}$ e $S_{fy} = \frac{(\Delta F)^2}{4} \sum_{i=1}^{N} \frac{y_i f_i}{\sigma_i^2}$; y determinar si el valor absoluto de la diferencia entre la fracción fetal ajustada y la fracción fetal medida es mayor que un valor predeterminado para el error en la fracción fetal medida, en el que *F* representa la fracción fetal ajustada, F_0 representa la fracción fetal medida, delta *F* (por ejemplo, *ΔF*) representa el error en la fracción fetal medida, S representa una variable auxiliar introducida para simplificar los cálculos, *f* representa la mediana de perfil de recuento de referencia, épsilon representa el perfil de recuento medido normalizado con respecto a los recuentos totales, *N* representa el número total de secciones genómicas, i representa una sección genómica seleccionada, sigma (σ) representa la incertidumbre asociada con f para una sección genómica seleccionada, y en el que se realiza una determinación euploide o no euploide basada en el valor numérico de *X*. En algunas implementaciones, el valor predeterminado se calcula usando la fórmula $|F-F_0|<\Delta F$. En determinadas implementaciones, *X* es mayor que $|F-F_0|<\Delta F$. En algunas implementaciones, *X* es menor que $|F-F_0|<\Delta F$. En determinadas implementaciones, *X* es igual a $|F-F_0|<\Delta F$

En determinadas implementaciones, evaluar la suma de los residuos al cuadrado suponiendo ploidía fija y fracción fetal optimizada incluye: medir la fracción fetal; obtener la fracción fetal optimizada; calcular el resultado numérico de la fórmula $\varphi_E - \varphi_T = \frac{-1}{(\Delta F)^2(1+S_{ff})}\left[F_0^2 S_{ff} + 4F_0(S_{ff} - S_{fy}) - 4(S_{ff} - S_{fy})^2\right]$ usando valores obtenidos de la implementación C12; y determinar si phi es menor o mayor que un valor predeterminado, en el que phi representa la diferencia entre sumas de residuos al cuadrado evaluados suponiendo un resultado euploide o de trisomía, respectivamente, F_0 representa la fracción fetal medida, delta F (por ejemplo, *ΔF*) representa el error en la fracción fetal medida, S representa una variable auxiliar introducida para simplificar los cálculos, *f* representa la mediana de perfil de recuento de referencia, *y* representa el perfil de recuento medido normalizado con respecto a los recuentos totales, y en el que se realiza una determinación de euploide o no euploide basándose en el valor numérico de phi. En algunas implementaciones, el valor predeterminado es 0. En determinadas implementaciones, phi es igual al valor predeterminado. En algunas implementaciones, phi es menor que el valor predeterminado. En determinadas implementaciones, phi es mayor que el valor predeterminado.

En algunas implementaciones, una determinación no euploide es una determinación de trisomía. En determinadas implementaciones, una determinación no euploide es una determinación de monoploidía. En algunas implementaciones, determinar la presencia o ausencia de una aberración cromosómica segmentaria o una aneuploidía fetal o ambas incluye, proporcionar un gráfico del resultado, un informe del resultado, un archivo electrónico que incluye el resultado, una representación bidimensional del resultado, una representación tridimensional del resultado, o combinaciones de los mismos, a un profesional sanitario. En determinadas implementaciones, el profesional sanitario proporciona una recomendación basada en lo proporcionado.

También se proporciona en el presente documento un producto de programa informático, que incluye un medio utilizable por ordenador que tiene un código de programa legible por ordenador incorporado en el mismo, incluyendo el código de programa legible por ordenador distintos módulos de software que incluyen un módulo de secuenciación, un módulo de procesamiento lógico y un módulo de organización de visualización de datos, estando el código de programa legible por ordenador adaptado para ejecutarse para implementar un método para determinar la ploidía fetal, incluyendo el método: (a) generar, por el módulo de procesamiento lógico, un perfil de recuento sin procesar basándose en lecturas de secuencia de ácidos nucleicos circulantes, libres de células obtenidos, por el módulo de secuenciación, a partir de una muestra de un sujeto de prueba; (b) generar, por el módulo de procesamiento lógico, una mediana de perfil de recuento de referencia basándose en lecturas de secuencia de ácidos nucleicos circulantes, libres de células obtenidos, por el módulo de secuenciación, a partir de muestras de uno o más sujetos de referencia; (c) generar, por el módulo de procesamiento lógico, un perfil de recuento normalizado de (a) con respecto a los recuentos totales de las lecturas de secuencia de sujeto de prueba; (d) generar, por el módulo de procesamiento lógico, un perfil de recuento normalizado a partir de (b) con respecto a los recuentos totales de la una o más lecturas de secuencia de sujeto de referencia; (e) calcular, por el módulo de procesamiento lógico, la suma de residuos al cuadrado basándose en parte en los perfiles de recuento normalizados y una o más suposiciones elegidas de ploidía fija o ploidía optimizada, y fracción fetal fija o fracción fetal optimizada; (f) proporcionar, por el módulo de procesamiento lógico, una determinación de ploidía fetal basada en la suma de residuos al cuadrado en (e); y (g) organizar, por el módulo de organización de visualización de datos en respuesta a que se determina por el módulo de procesamiento lógico, una visualización de datos que indica la presencia o ausencia de una aberración cromosómica segmentaria o una aneuploidía fetal o ambas.

También se proporciona en el presente documento un aparato, que incluye una memoria en la que se almacena un producto de programa informático descrito en el presente documento. En algunas implementaciones, el aparato incluye un procesador que implementa una o más funciones del producto de programa informático descrito en el presente documento.

También se proporciona en el presente documento un sistema que incluye un aparato de secuenciación de ácido nucleico y un aparato de procesamiento, en el que el aparato de secuenciación obtiene lecturas de secuencia de una muestra, y el aparato de procesamiento obtiene las lecturas de secuencia del dispositivo de secuenciación y lleva a cabo un método que incluye: (a) generar un perfil de recuento sin procesar basándose en lecturas de secuencia de ácidos nucleicos circulantes, libres de células obtenidos a partir de una muestra de un sujeto de prueba; (b) generar una mediana de perfil de recuento de referencia basándose en lecturas de secuencia de ácidos nucleicos circulantes, libres de células obtenidos de muestras de uno o más sujetos de referencia; (c) generar un perfil de recuento normalizado a partir de (a) con respecto a los recuentos totales de las lecturas de secuencia de sujeto de prueba; (d) generar un perfil de recuento normalizado a partir de (b) con respecto a los recuentos totales de la una o más lecturas de secuencia de sujeto de referencia; (e) calcular la suma de los residuos al cuadrado basándose en parte en los perfiles de recuento normalizados y una o más suposiciones elegidas de ploidía fija o ploidía optimizada, y fracción fetal fija o fracción fetal optimizada; y (f) determinar la ploidía fetal basándose en la suma de los residuos al cuadrado en (e).

En algunas implementaciones, la profundidad de la secuenciación (por ejemplo, cobertura de la secuenciación o número de veces (por ejemplo, veces) que se secuencia todo el genoma) es equivalente a aproximadamente 0,1 veces o más, aproximadamente 0,2 veces o más, aproximadamente 0,3 veces o más, aproximadamente 0,4 veces o más, aproximadamente 0,5 veces o más, aproximadamente 0,6 veces o más, aproximadamente 0,7 veces o más, aproximadamente 0,8 veces o más, aproximadamente 0,9 veces o más, aproximadamente 1,0 veces o más, aproximadamente 1,1 veces o más, aproximadamente 1,2 veces o más, aproximadamente 1,3 veces o más, aproximadamente 1,4 veces o más, aproximadamente 1,5 veces o más, aproximadamente 1,6 veces o más, aproximadamente 1,7 veces o más, aproximadamente 1,8 veces o más, aproximadamente 1,9 veces o más, aproximadamente 2,0 veces o más, aproximadamente 2,5 veces o más, aproximadamente 3,0 veces o más, aproximadamente 3,5 veces o más, aproximadamente 4,0 veces o más, aproximadamente 4,5 veces o más, aproximadamente 5,0 veces o más, aproximadamente 5,5 veces o más, aproximadamente 6 veces o más, aproximadamente 6,5 veces o más, aproximadamente 7,0 veces o más, aproximadamente 7,5 veces o más, aproximadamente 8,0 veces o más, aproximadamente 8,5 veces o más, aproximadamente 9,0 veces o más, aproximadamente 9,5 veces o más, aproximadamente 10 veces o más, aproximadamente 20 veces o más, aproximadamente 30 veces o más, aproximadamente 40 veces o más, aproximadamente 50 veces o más, aproximadamente 60 veces o más, aproximadamente 70 veces o más, aproximadamente 80 veces o más, aproximadamente 90 veces o más, o 99 veces o más. En determinadas implementaciones, la fracción fetal de ácido nucleico circulante, libre de células es de aproximadamente el 50 por ciento o menos, aproximadamente el 45 por ciento o menos, aproximadamente el 40 por ciento o menos, aproximadamente el 35 por ciento o menos, aproximadamente el 30 por ciento o menos, aproximadamente el 25 por ciento o menos, aproximadamente el 20 por ciento o menos, aproximadamente el 15 por ciento o menos, aproximadamente el 10 por ciento o menos, aproximadamente el 5 por ciento o menos, o aproximadamente el 2 por ciento o menos, del ácido nucleico circulante, libre de células total.

En algunas implementaciones, la fracción fetal (por ejemplo, medida o estimada) se usa durante una o más etapas de procesamiento para modificar los valores obtenidos de una o más manipulaciones de procesamiento realizadas para generar una determinación de la presencia o ausencia de una variación genética. En determinadas implementaciones, la fracción fetal no se usa para alterar un valor de punto de corte umbral y, algunas veces, la fracción fetal se usa para alterar los recuentos de lectura mapeados o las derivaciones de los mismos.

También se proporciona en el presente documento un método para identificar una aberración cromosómica segmentaria o una aneuploidía fetal o ambas, que comprende: (a) obtener de un sujeto de prueba una muestra que comprende ácido nucleico circulante, libre de células; (b) aislar el ácido nucleico de muestra libre de células de la muestra; (c) obtener lecturas de secuencia del ácido nucleico de muestra libre de células; (d) mapear las lecturas de secuencia obtenidas en (c) en un genoma conocido, genoma conocido que se ha dividido en secciones genómicas; (e) contar las lecturas de secuencia mapeadas dentro de las secciones genómicas; (f) proporcionar una normalización de las lecturas de secuencia mapeadas contadas en (e) basándose en una normalización de ventana deslizante; y (g) proporcionar un resultado que identifica una aberración cromosómica segmentaria o una aneuploidía fetal o ambas a partir de la normalización en (f). En algunas implementaciones (f) comprende uno o más de: (i) generar un perfil de recuento normalizado de muestra; (ii) eliminar secciones genómicas de ruido; (iii) identificar secciones genómicas que se desvían significativamente de la elevación media; (iv) eliminar puntos de datos solitarios identificados en (iii); (v) agrupar puntos de datos vecinos que se desvían en la misma dirección; y (vi) caracterizar los bordes y las elevaciones de aberración. En determinadas implementaciones, (v) se realiza usando una tolerancia de brecha predefinida. En algunas implementaciones, puede usarse la caracterización de los bordes de aberración para determinar la anchura de una aberración.

También se proporciona en el presente documento un método para identificar una aberración cromosómica segmentaria, una aneuploidía fetal o ambas, que comprende: (a) obtener lecturas de secuencia de un ácido nucleico de muestra libre de células; (b) mapear las lecturas de secuencia obtenidas en (a) en un genoma conocido, genoma conocido que se ha dividido en secciones genómicas; (c) contar las lecturas de secuencia mapeadas dentro de las secciones genómicas; (d) proporcionar una normalización de las lecturas de secuencia mapeadas contadas en (c) basándose en una normalización de ventana deslizante; y (e) proporcionar un resultado que identifique una aberración cromosómica segmentaria, una aneuploidía fetal o ambas a partir de la normalización en (d). En algunas implementaciones (d) comprende uno o más de: (i) generar un perfil de recuento normalizado de muestra; (ii) eliminar secciones genómicas de ruido; (iii) identificar secciones genómicas que se desvían significativamente de la elevación media; (iv) eliminar puntos de datos solitarios identificados en (iii); (v) agrupar puntos de datos vecinos que se desvían en la misma dirección; y (vi) caracterizar los bordes y las elevaciones de aberración. En determinadas implementaciones, (v) se realiza usando una tolerancia de brecha predefinida. En algunas implementaciones, puede usarse la caracterización de los bordes de aberración para determinar la anchura de una aberración.

También se proporciona en el presente documento un método para identificar una variación genética, que comprende: (a) obtener de un sujeto de prueba una muestra que comprende ácido nucleico circulante, libre de células; (b) aislar el ácido nucleico de muestra libre de células de la muestra; (c) obtener lecturas de secuencia del ácido nucleico de muestra libre de células; (d) mapear las lecturas de secuencia obtenidas en (c) en un genoma conocido, genoma conocido que se ha dividido en secciones genómicas; (e) contar las lecturas de secuencia mapeadas dentro de las secciones genómicas; (f) proporcionar una normalización de las lecturas de secuencia mapeadas contadas en (e) basándose en una normalización de ventana deslizante; y (g) proporcionar un resultado que identifique una variación genética a partir de la normalización en (f). En algunas implementaciones (f) comprende uno o más de: (i) generar un perfil de recuento normalizado de muestra; (ii) eliminar secciones genómicas de ruido; (iii) identificar secciones genómicas que se desvían significativamente de la elevación media; (iv) eliminar puntos de datos solitarios identificados en (iii); (v) agrupar puntos de datos vecinos que se desvían en la misma dirección; y (vi) caracterizar los bordes y las elevaciones de aberración. En determinadas implementaciones, (v) se realiza usando una tolerancia de brecha predefinida. En algunas implementaciones, puede usarse la caracterización de los bordes de aberración para determinar la anchura de una aberración.

También se proporciona en el presente documento un método para identificar una variación genética, que comprende: (a) obtener lecturas de secuencia de un ácido nucleico de muestra libre de células; (b) mapear las lecturas de secuencia obtenidas en (a) en un genoma conocido, genoma conocido que se ha dividido en secciones genómicas; (c) contar las lecturas de secuencia mapeadas dentro de las secciones genómicas; (d) proporcionar una normalización de las lecturas de secuencia mapeadas contadas en (c) basándose en una normalización de ventana deslizante; y (e) proporcionar un resultado que identifique una variación genética a partir de la normalización en (d). En algunas implementaciones (d) comprende uno o más de: (i) generar un perfil de recuento normalizado de muestra; (ii) eliminar secciones genómicas de ruido; (iii) identificar secciones genómicas que se desvían significativamente de la elevación media; (iv) eliminar puntos de datos solitarios identificados en (iii); (v) agrupar puntos de datos vecinos que se desvían en la misma dirección; y (vi) caracterizar los bordes y las elevaciones de aberración. En algunas implementaciones, (v) se realiza usando una tolerancia de brecha predefinida. En algunas implementaciones, puede usarse la caracterización de los bordes de aberración para determinar la anchura de una aberración.

En determinadas implementaciones, caracterizar las elevaciones y los bordes de aberración comprende el uso de integrales sobre la aberración sospechada y su entorno inmediato. En algunas implementaciones, (vi) comprende: (1) realizar una regresión lineal en secciones genómicas seleccionadas en un lado de la aberración candidata; (2) realizar una regresión lineal en secciones genómicas seleccionadas en el otro lado de la aberración candidata; (3) determinar la elevación media dentro de la aberración candidata y/o la pendiente del segmento de línea que conecta dos líneas de regresión lineal; y (4) determinar la diferencia entre las ordenadas en el origen de dos líneas de regresión lineal, combinada con la elevación media dentro de la aberración, en el que realizar (1) a (4) produce la anchura de la

aberración. En algunas implementaciones, (1) a (4) se repiten en el rango de aproximadamente 1 a aproximadamente 100 veces, y en determinadas implementaciones, (1) a (4) se repiten en el rango de aproximadamente 1 a aproximadamente 10 veces. Las expresiones "anchura de una aberración" o "anchura de la aberración", tal como se usan en el presente documento, se refieren al número de bins, secciones genómicas y/o nucleótidos entre un lado de una aberración y el otro lado de una aberración (por ejemplo, los bordes de una microdeleción o microduplicación). En algunas implementaciones, las secciones genómicas seleccionadas en un lado o el otro lado de una aberración candidata son secciones genómicas adyacentes. En determinadas implementaciones, las secciones genómicas adyacentes comprenden secciones genómicas contiguas y/o ininterrumpidas y, en algunas implementaciones, las secciones genómicas adyacentes permiten brechas o interrupciones de un tamaño predeterminado.

También se proporciona en el presente documento un método para detectar y/o determinar la presencia o ausencia de una afección, un síndrome o una anomalía enumerados en la tabla 1B, que comprende: (a) obtener lecturas de secuencia de un ácido nucleico de muestra libre de células; (b) mapear las lecturas de secuencia obtenidas en (a) en un genoma conocido, genoma conocido que se ha dividido en secciones genómicas; (c) contar las lecturas de secuencia mapeadas dentro de las secciones genómicas; (d) determinar la presencia o ausencia de una afección, un síndrome o una anomalía enumerados en la tabla 1B, basándose en los recuentos obtenidos en (c) y/o derivaciones procesadas de los mismos. En algunas implementaciones, (d) comprende proporcionar un perfil de recuento normalizado de muestra (por ejemplo, normalización basada en bins). En algunas implementaciones, una determinación de la presencia o ausencia de una afección, un síndrome o una anomalía es, o incluye, la detección de una afección, un síndrome o una anomalía enumerados en la tabla 1B.

En algunas implementaciones, el ácido nucleico de muestra libre de células se aísla de la sangre obtenida de un sujeto de prueba. En determinadas implementaciones, el ácido nucleico de muestra libre de células se aísla de suero obtenido de un sujeto de prueba. En algunas implementaciones, el ácido nucleico de muestra libre de células se aísla de plasma obtenido de un sujeto de prueba. En determinadas implementaciones, el sujeto de prueba se elige de un ser humano, un animal y una planta. En algunas implementaciones, un sujeto de prueba humano se elige de una mujer, una mujer embarazada, un hombre, un feto o un recién nacido.

En determinadas implementaciones, las lecturas de secuencia del ácido nucleico de muestra libre de células están en forma de fragmentos de polinucleótidos. En algunas implementaciones, los fragmentos de polinucleótidos tienen entre aproximadamente 20 y aproximadamente 50 nucleótidos de longitud y, en determinadas implementaciones, los polinucleótidos tienen entre aproximadamente 30 y aproximadamente 40 nucleótidos de longitud.

En algunas implementaciones, también se proporcionan métodos para calcular con sesgo reducido niveles de sección genómica para una muestra de prueba, que comprenden: (a) obtener recuentos de lecturas de secuencia mapeadas en bins de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante, libre de células de una muestra de prueba; (b) determinar un sesgo de guanina y citosina (GC) para cada uno de los bins a través de múltiples muestras a partir de una relación ajustada para cada muestra entre (i) los recuentos de las lecturas de secuencia mapeadas en cada uno de los bins, y (ii) el contenido de GC para cada uno de los bins; y (c) calcular un nivel de sección genómica para cada uno de los bins a partir de una relación ajustada entre (i) el sesgo de GC y (ii) los recuentos de las lecturas de secuencia mapeadas en cada uno de los bins, proporcionándose de ese modo niveles de sección genómica calculados, mediante lo cual el sesgo en los recuentos de las lecturas de secuencia mapeadas en cada uno de los bins se reduce en los niveles de sección genómica calculados. Un bin comprende a veces uno o más segmentos de un genoma de referencia, tal como se describe con mayor detalle en el presente documento.

En determinadas implementaciones, se proporcionan métodos para identificar la presencia o ausencia de una aneuploidía en un feto, que comprenden: (a) obtener recuentos de lecturas de secuencia mapeadas en bins de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante, libre de células de una mujer embarazada que porta un feto; (b) determinar un sesgo de guanina y citosina (GC) para cada uno de los bins a través de múltiples muestras a partir de una relación ajustada para cada muestra entre (i) los recuentos de las lecturas de secuencia mapeadas en cada uno de los bins, y (ii) el contenido de GC para cada uno de los bins; (c) calcular un nivel de sección genómica para cada uno de los bins a partir de una relación ajustada entre el sesgo de GC y los recuentos de las lecturas de secuencia mapeadas en cada uno de los bins, proporcionándose de ese modo niveles de sección genómica calculados; y (d) identificar la presencia o ausencia de una aneuploidía para el feto según los niveles de sección genómica calculados con una sensibilidad del 95 % o más y una especificidad del 95 % o más.

En algunas implementaciones, también se proporcionan métodos para calcular con sesgo reducido niveles de sección genómica para una muestra de prueba, que comprenden: (a) obtener recuentos de lecturas de secuencia mapeadas en bins de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante, libre de células de una muestra de prueba; (b) determinar el sesgo experimental para cada uno de los bins a través de múltiples muestras a partir de una relación ajustada entre (i) los recuentos de las lecturas de secuencia mapeadas en cada uno de los bins, y (ii) una característica de mapeo para cada uno de los bins; y (c) calcular un nivel de sección genómica para cada uno de los bins a partir de una relación ajustada entre el sesgo experimental y los recuentos de las lecturas de secuencia mapeadas en cada uno de los bins, proporcionándose de ese modo niveles de sección genómica calculados, mediante lo cual el sesgo en los recuentos de las lecturas de secuencia mapeadas en cada uno de los bins se reduce en los niveles de sección genómica calculados.

Una variación del número de copias materno y/o fetal (por ejemplo, duplicación, deleción, inserción) puede dar lugar potencialmente a una identificación de falso positivo o falso negativo cuando se determina la presencia o ausencia de una aneuploidía cromosómica. En determinadas implementaciones proporcionadas en el presente documento, se trata de métodos que comprenden identificar una variación del número de copias materno, una variación del número de copias fetal, o una variación del número de copias materno y una variación del número de copias fetal en un segmento genómico (por ejemplo, un perfil) y ajustar las elevaciones de señal asociadas con tales variaciones del número de copias. Tales métodos se mencionan en el presente documento como "relleno". Realizar ajustes de tal manera puede reducir o eliminar las interferencias de las variaciones del número de copias materno y/o de las variaciones del número de copias fetal que pueden dar como resultado determinaciones de resultados de falso negativo o falso positivo. Un método de relleno puede convertir perfiles de falso positivo que indican una posible aneuploidía (por ejemplo, una trisomía 13) en un perfil indicativo de un resultado de verdadero negativo (por ejemplo, la ausencia de una trisomía), en algunos casos. Un método de relleno puede convertir perfiles de falso negativo que indican la ausencia de una aneuploidía en un perfil que indica un resultado de verdadero positivo (por ejemplo, la presencia de una trisomía), en algunos casos.

Por tanto, en determinados aspectos en el presente documento se proporcionan métodos para identificar la presencia o ausencia de una aneuploidía cromosómica en un feto con diagnósticos de falso negativo y falso positivo reducidos, que comprenden: (a) obtener recuentos de lecturas de secuencia de ácidos nucleicos mapeadas en secciones genómicas de un genoma de referencia, lecturas de secuencia que son lecturas de ácidos nucleicos circulantes, libres de células de una mujer embarazada, (b) normalizar los recuentos mapeados en las secciones genómicas del genoma de referencia, proporcionándose de ese modo un perfil de recuentos normalizados para las secciones genómicas, (c) identificar una primera elevación de los recuentos normalizados significativamente diferente de una segunda elevación de los recuentos normalizados en el perfil, primera elevación que es para un primer conjunto de secciones genómicas, y segunda elevación que es para un segundo conjunto de secciones genómicas, (d) determinar un rango de elevación esperado para una variación del número de copias homocigotas y heterocigotas según un valor de incertidumbre para un segmento del genoma, (e) ajustar la primera elevación en un valor predeterminado cuando la primera elevación está dentro de uno de los rangos de elevación esperados, proporcionándose de ese modo un ajuste de la primera elevación; y (f) determinar la presencia o ausencia de una aneuploidía cromosómica en el feto según las elevaciones de secciones genómicas que comprenden el ajuste de (e), mediante lo cual el resultado determinante de la presencia o ausencia de la aneuploidía cromosómica se genera a partir de las lecturas de secuencia de ácido nucleico.

Es algunos aspectos se proporcionan métodos para identificar una variación del número de copias materno y/o fetal dentro de un genoma de una mujer embarazada que porta un feto, que comprenden: (a) obtener recuentos de lecturas de secuencia de ácidos nucleicos mapeadas en secciones genómicas de un genoma de referencia, lecturas de secuencia que son lecturas de ácidos nucleicos circulantes, libres de células de una mujer embarazada, (b) normalizar los recuentos mapeados en las secciones genómicas del genoma de referencia, proporcionándose de ese modo un perfil de recuentos normalizados para las secciones genómicas, (c) identificar una primera elevación de los recuentos normalizados significativamente diferente de una segunda elevación de los recuentos normalizados en el perfil, primera elevación que es para un primer conjunto de secciones genómicas, y segunda elevación que es para un segundo conjunto de secciones genómicas, (d) determinar un rango de elevación esperado para una variación del número de copias homocigotas y heterocigotas según un valor de incertidumbre para un segmento del genoma, (e) identificar una variación del número de copias materno y/o fetal dentro de la sección genómica basándose en uno de los rangos de elevación esperados, mediante lo cual la variación del número de copias materno y/o fetal se identifica a partir de las lecturas de secuencia de ácido nucleico.

Tal como se usa en el presente documento, la expresión "secciones genómicas" de un genoma de referencia es igual a "porciones de un genoma de referencia".

Determinados aspectos de la tecnología se describen adicionalmente en la siguiente descripción, los ejemplos, las reivindicaciones y dibujos.

Breve descripción de los dibujos

Los dibujos ilustran implementaciones de la tecnología y no son limitativos. Para mayor claridad y facilidad de ilustración, los dibujos no se trazan a escala y, en algunos casos, diversos aspectos pueden mostrarse exagerados o ampliados para facilitar una comprensión de implementaciones particulares.

La Fig. 1 ilustra gráficamente cómo el aumento de la incertidumbre en los recuentos de bins dentro de una región genómica reduce a veces las brechas entre los valores de Z de euploide y trisomía. La Fig. 2 ilustra gráficamente cómo las diferencias disminuidas entre el número triploide y euploide de recuentos dentro de una región genómica reduce a veces la potencia predictiva de las puntuaciones Z. Véase el ejemplo 1 para detalles y resultados experimentales.

La Fig. 3 ilustra gráficamente la dependencia de los valores de p de la posición de los bins genómicos dentro del cromosoma 21. La Fig. 4 representa esquemáticamente un procedimiento de filtrado de bins. Se alinean un gran número de muestras euploides, se evalúan las incertidumbres del recuento de bins (valores de D.E. o D.M.A.) y a veces se filtran los bins con las mayores incertidumbres. La Fig. 5 ilustra gráficamente los perfiles de recuento para el

cromosoma 21 en dos pacientes. La Fig. 6 ilustra gráficamente los perfiles de recuento para pacientes usados para filtrar los bins no informativos del cromosoma 18. En la Fig. 6, las dos trazas inferiores muestran a un paciente con una deleción grande en el cromosoma 18. Véase el ejemplo 1 para detalles y resultados experimentales.

La Fig. 7 ilustra gráficamente la dependencia de los valores de p de la posición de los bins genómicos dentro del cromosoma 18. La Fig. 8 representa esquemáticamente la normalización del recuento de bins. El procedimiento alinea en primer lugar perfiles de recuento euploides conocidos, a partir de un conjunto de datos, y los normaliza con respecto a los recuentos totales. Para cada bin se evalúan la mediana de recuentos y desviaciones de las medianas. A veces se eliminan los bins con demasiada variabilidad (que supera 3 desviaciones medias absolutas (por ejemplo, D.M.A.)). Los bins restantes se normalizan de nuevo con respecto a los recuentos totales residuales, y las medianas se evalúan de nuevo después de la nueva normalización, en algunas implementaciones. Finalmente, se usa el perfil de referencia resultante (véase la traza inferior, panel izquierdo) para normalizar los recuentos de bins en las muestras de prueba (véase la traza superior, panel izquierdo), suavizar el contorno del recuento (véase la traza a la derecha) y dejar huecos en los que se excluyen los bins no informativos de la consideración. La Fig. 9 ilustra gráficamente el comportamiento esperado de los perfiles de recuento normalizados. La mayoría de los recuentos de bins normalizados a menudo se centrarán en 1, con ruido aleatorio superpuesto. Las deleciones y duplicaciones (por ejemplo, deleciones y duplicaciones maternas o fetales, o maternas y fetales) a veces cambian la elevación a un múltiplo entero de 0,5. Las elevaciones de perfil correspondientes a un cromosoma fetal triploide a menudo se desplazan hacia arriba en proporción a la fracción fetal. Véase el ejemplo 1 para detalles y resultados experimentales.

La Fig. 10 ilustra gráficamente un perfil de recuento de T18 normalizado con una deleción materna heterocigota en el cromosoma 18. El segmento de color gris claro del trazado del gráfico muestra una mayor elevación promedio que el segmento de color negro del trazado del gráfico. Véase el ejemplo 1 para detalles y resultados experimentales.

La Fig. 11 ilustra gráficamente perfiles de recuento en bins normalizados para dos muestras recogidas del mismo paciente con deleción materna heterocigota en el cromosoma 18. Los trazados sustancialmente idénticos pueden usarse para determinar si dos muestras son del mismo donante. La Fig. 12 ilustra gráficamente los perfiles de recuento de bins normalizados de una muestra de un estudio, en comparación con dos muestras de un estudio previo. La duplicación en el cromosoma 22 señala de manera inequívoca la identidad del paciente. La Fig. 13 ilustra gráficamente los perfiles de recuento de bins normalizados del cromosoma 4 en los mismos tres pacientes presentados en la Fig. 12. La duplicación en el cromosoma 4 confirma la identidad del paciente establecida en la Fig. 12. Véase el ejemplo 1 para detalles y resultados experimentales.

La Fig. 14 ilustra gráficamente la distribución de recuentos de bins normalizados en el cromosoma 5 de una muestra euploide. La Fig. 15 ilustra gráficamente dos muestras con diferentes niveles de ruido en sus perfiles de recuento normalizados. La Fig. 16 representa esquemáticamente factores que determinan la confianza en la elevación de pico: desviación estándar del ruido (por ejemplo, o) y desviación promedio con respecto a los valores iniciales de referencia (por ejemplo, △). Véase el ejemplo 1 para detalles y resultados experimentales.

La Fig. 17 ilustra gráficamente los resultados de aplicar una función de correlación a los recuentos de bins normalizados. La función de correlación que se muestra en la Fig. 17 se usó para normalizar los recuentos de bins en el cromosoma 5 de un paciente euploide seleccionado arbitrariamente. La Fig. 18 ilustra gráficamente la desviación estándar para la elevación de tramo promedio en el cromosoma 5, evaluada como una estimación de la muestra (puntos de datos cuadrados) y comparada con el error estándar de la media (puntos de datos triangulares) y con la estimación corregida para la autocorrelación p = 0,5 (puntos de datos circulares). La aberración representada en la Fig. 18 tiene una longitud de aproximadamente 18 bins. Véase el ejemplo 1 para detalles y resultados experimentales.

La Fig. 19 ilustra gráficamente los valores de Z calculados para la elevación de pico promedio en el cromosoma 4. El paciente tiene una duplicación materna heterocigota en el cromosoma 4 (véase la Fig. 13). La Fig. 20 ilustra gráficamente los valores de p para la elevación de pico promedio, basándose en una prueba de la t y los valores de Z de la Fig. 19. El orden de la distribución t se determina mediante la longitud de la aberración. Véase el ejemplo 1 para detalles y resultados experimentales.

La Fig. 21 representa esquemáticamente comparaciones de borde entre aberraciones coincidentes de diferentes muestras. Se ilustran en la Fig. 21 solapamientos, contención y desviaciones vecinas. La Fig. 22 ilustra gráficamente las duplicaciones heterocigotas coincidentes en el cromosoma 4 (traza superior de color gris oscuro y traza inferior de color negro), en contraste con una aberración que toca marginalmente en una muestra no relacionada (traza intermedia de color gris claro). Véase el ejemplo 1 para detalles y resultados experimentales.

La Fig. 23 representa esquemáticamente la detección de borde por medio de las primeras derivadas evaluadas numéricamente de los perfiles de recuento. La Fig. 24 ilustra gráficamente que la primera derivada de los perfiles de recuento, obtenida a partir de datos reales, son difíciles de distinguir del ruido. La Fig. 25 ilustra gráficamente la tercera potencia del perfil de recuento, desplazada en 1 para suprimir ruido y mejorar la señal (véase el trazo superior). También se ilustra en la Fig. 25 (véase la traza inferior) una primera derivada de la traza superior. Los bordes pueden detectarse de manera inconfundible. Véase el ejemplo 1 para detalles y resultados experimentales.

La Fig. 26 ilustra gráficamente los histogramas de la mediana de las elevaciones del cromosoma 21 para diversos pacientes. El histograma de color negro ilustra la mediana de las elevaciones del cromosoma 21 para 86 pacientes euploides. El histograma de color gris ilustra la mediana de las elevaciones del cromosoma 21 para 35 pacientes con trisomía 21. Se normalizaron los perfiles de recuento con respecto a un conjunto de referencia euploide antes de evaluar la mediana de las elevaciones. La Fig. 27 ilustra gráficamente una distribución de recuentos normalizados para el cromosoma 21 en una muestra de trisomía. La Fig. 28 representa gráficamente las razones de área para diversos pacientes. El histograma de color gris oscuro ilustra las razones de área del cromosoma 21 para 86 pacientes euploides. El histograma de color gris claro ilustra las razones de área del cromosoma 21 para 35 pacientes con trisomía 21. Se normalizaron los perfiles de recuento con respecto a un conjunto de referencia euploide antes de evaluar las razones de área. Véase el ejemplo 1 para detalles y resultados experimentales.

La Fig. 29 ilustra gráficamente la razón de área en el cromosoma 21 representada gráficamente frente a la mediana de elevaciones de recuento normalizado. Los puntos de datos de color gris claro representan aproximadamente 86 muestras euploides. Los puntos de datos de color gris oscuro representan aproximadamente 35 pacientes con trisomía. Véase el ejemplo 1 para detalles y resultados experimentales.

La Fig. 30 ilustra gráficamente relaciones entre 9 criterios de clasificación diferentes, según se evaluó para un conjunto de pacientes con trisomía. Los criterios implican puntuaciones Z, mediana de elevaciones de recuento normalizado, razones de área, fracciones fetales medidas, fracciones fetales ajustadas, la razón entre fracciones fetales ajustadas y medidas, suma de residuos al cuadrado para fracciones fetales ajustadas, suma de residuos al cuadrado con fracciones fetales ajustadas y ploidía fija, y valores de ploidía ajustados. Véase el ejemplo 1 para detalles y resultados experimentales.

La Fig. 31 ilustra gráficamente perfiles de phi funcionales simulados para casos de trisomía (de color gris claro) y de euploides (de color gris oscuro). La Fig. 32 ilustra gráficamente los valores de phi funcionales derivados de conjuntos de datos de trisomía (de color gris oscuro) y de euploides (de color gris claro) medidos. Véase el ejemplo 2 para detalles y resultados experimentales.

La Fig. 33 ilustra gráficamente la suma linealizada de diferencias al cuadrado en función de la fracción fetal medida. La Fig. 34 ilustra gráficamente las estimaciones de fracción fetal basadas en los recuentos de Y representadas gráficamente frente a los valores obtenidos de un ensayo cuantificador fetal (por ejemplo, FQA) de la fracción fetal. La Fig. 35 ilustra gráficamente los valores de Z para pacientes con T21 representados gráficamente frente a mediciones de la fracción fetal de FQA. Para las figuras 33-35, véase el ejemplo 2 para detalles y resultados experimentales.

La Fig. 36 ilustra gráficamente las estimaciones de fracción fetal basadas en el cromosoma Y representadas gráficamente frente a las fracciones fetales medidas. La Fig. 37 ilustra gráficamente las estimaciones de fracción fetal basadas en el cromosoma 21 (cr21) representadas gráficamente frente a las fracciones fetales medidas. La Fig. 38 ilustra gráficamente las estimaciones de fracción fetal derivadas de los recuentos del cromosoma X representados gráficamente frente a las fracciones fetales medidas. La Fig. 39 ilustra gráficamente las medianas de los recuentos de bins normalizados para casos de T21 representados gráficamente frente a fracciones fetales medidas. Para las figuras 36-39, véase el ejemplo 2 para detalles y resultados experimentales.

La Fig. 40 ilustra gráficamente los perfiles simulados de ploidía triploide ajustados (por ejemplo, X) en función de F_0 con errores fijos ΔF = +/- 0,2 %. La Fig. 41 ilustra gráficamente los valores de ploidía triploide ajustados en función de las fracciones fetales medidas. Para las Fig. 40 y 41, véase el ejemplo 2 para detalles y resultados experimentales.

La Fig. 42 ilustra gráficamente distribuciones de probabilidad para ploidía ajustada en diferentes niveles de errores en las fracciones fetales medidas. El panel superior en la Fig. 42 establece el error de la fracción fetal medida en el 0,2 %. El panel central en la Fig. 42 establece el error de la fracción fetal medida en el 0,4 %. El panel inferior en la Fig. 42 establece el error de la fracción fetal medida en el 0,6 %. Véase el ejemplo 2 para detalles y resultados experimentales.

La Fig. 43 ilustra gráficamente distribuciones euploides y de trisomía de valores de ploidía ajustados para un conjunto de datos derivado de muestras de pacientes. La Fig. 44 ilustra gráficamente las fracciones fetales ajustadas representadas gráficamente frente a las fracciones fetales medidas. Para las Fig. 43 y 44, véase el ejemplo 2 para detalles y resultados experimentales.

La Fig. 45 ilustra esquemáticamente la diferencia predicha entre sumas euploides y de trisomía de residuos al cuadrado para la fracción fetal ajustada en función de la fracción fetal medida. La Fig. 46 ilustra gráficamente la diferencia entre sumas euploides y de trisomía de residuos al cuadrado en función de la fracción fetal medida usando un conjunto de datos derivado de muestras de pacientes. Los puntos de datos se obtienen ajustando los valores de fracción fetal suponiendo incertidumbres fijas en las mediciones de fracción fetal. La Fig. 47 ilustra gráficamente la diferencia entre sumas euploides y de trisomía de residuos al cuadrado en función de la fracción fetal medida. Los puntos de datos se obtienen ajustando los valores de fracción fetal suponiendo que las incertidumbres en las mediciones de fracción fetal son proporcionales a las fracciones fetales: $\Delta F = 2/3 + F_0/Q$. Para las Fig. 45-47, véase el ejemplo 2 para detalles y resultados experimentales.

La Fig. 48 ilustra esquemáticamente la dependencia predicha de la fracción fetal ajustada representada gráficamente frente a los perfiles de fracción fetal medidos en desviaciones sistemáticas en recuentos de referencia. Las

ramificaciones inferior y superior representan casos de euploides y triploides, respectivamente. La Fig. 49 representa gráficamente los efectos de errores sistemáticos simulados A impuestos artificialmente sobre datos reales. La diagonal principal en el panel superior y la diagonal superior en el panel inferior derecho representan una concordancia ideal. La línea de color gris oscuro en todos los paneles representa las ecuaciones (51) y (53) para los casos euploide y triploide, respectivamente. Los puntos de datos representan mediciones reales que incorporan diversos niveles de desplazamientos sistemáticos artificiales. Los desplazamientos sistemáticos se dan como la desviación por encima de cada panel. Para las Fig. 48 y 49, véase el ejemplo 2 para detalles y resultados experimentales.

La Fig. 50 ilustra gráficamente la fracción fetal ajustada en función de la desviación sistemática, obtenida para un euploide y para un conjunto de datos triploide. La Fig. 51 ilustra gráficamente simulaciones basadas en la ecuación (61), junto con fracciones fetales ajustadas para datos reales. Las líneas de color negro representan dos desviaciones estándar (obtenidas como la raíz cuadrada de la ecuación (61)) por encima y por debajo de la ecuación (40). ΔF se establece en $2/3 + F_0/Q$. Para las Fig. 50 y 51, véase el ejemplo 2 para detalles y resultados experimentales.

El ejemplo 3 abordó las Fig. 52 a 61F. La Fig. 52 ilustra gráficamente un ejemplo de aplicación del algoritmo de suma acumulativa a una microdeleción materna heterocigota en el cromosoma 12, bin 1457. La diferencia entre las ordenadas en el origen asociadas con los modelos lineales izquierdo y derecho es de 2,92, lo que indica que la deleción heterocigota tiene una anchura de 6 bins.

La Fig. 53 ilustra gráficamente una deleción heterocigota hipotética, con una anchura de aproximadamente 2 secciones genómicas y su perfil de suma acumulativa asociado. La diferencia entre las ordenadas en el origen izquierda y derecha es de -1.

La Fig. 54 ilustra gráficamente una deleción homocigota hipotética, una anchura de aproximadamente 2 secciones genómicas y su perfil de suma acumulativa asociado. La diferencia entre las ordenadas en el origen izquierda y derecha es de -2.

La Fig. 55 ilustra gráficamente una deleción heterocigota hipotética, una anchura de aproximadamente 6 secciones genómicas y su perfil de suma acumulativa asociado. La diferencia entre las ordenadas en el origen izquierda y derecha es de -3.

La Fig. 56 ilustra gráficamente una deleción homocigota hipotética, una anchura de aproximadamente 6 secciones genómicas y su perfil de suma acumulativa asociado. La diferencia entre las ordenadas en el origen izquierda y derecha es de -6.

La Fig. 57 ilustra gráficamente una duplicación heterocigota hipotética, una anchura de aproximadamente 2 secciones genómicas y su perfil de suma acumulativa asociado. La diferencia entre las ordenadas en el origen izquierda y derecha es de 1.

La Fig. 58 ilustra gráficamente una duplicación homocigota hipotética, una anchura de aproximadamente 2 secciones genómicas y su perfil de suma acumulativa asociado. La diferencia entre las ordenadas en el origen izquierda y derecha es de 2.

La Fig. 59 ilustra gráficamente una duplicación heterocigota hipotética, una anchura de aproximadamente 6 secciones genómicas y su perfil de suma acumulativa asociado. La diferencia entre las ordenadas en el origen izquierda y derecha es de 3.

La Fig. 60 ilustra gráficamente una duplicación homocigota hipotética, una anchura de aproximadamente 6 secciones genómicas y su perfil de suma acumulativa asociado. La diferencia entre las ordenadas en el origen izquierda y derecha es de 6.

Las Fig. 61A-F ilustran gráficamente candidatos para duplicaciones heterocigotas fetales en datos obtenidos de estudios clínicos en mujeres y lactantes con altos valores de fracción fetal (el 40-50 %). Para descartar la posibilidad de que las aberraciones se originen de la madre y no del feto, se usaron perfiles maternos independientes. La elevación de perfil en las regiones afectadas es de aproximadamente 1,25, según las estimaciones de fracción fetal.

La Fig. 62 muestra un perfil de elevaciones para los cr20, cr21 (de ~55750 a ~56750) y cr22 obtenido de una mujer embarazada que porta un feto euploide.

La Fig. 63 muestra un perfil de elevaciones para los cr20, cr21 (de ~55750 a ~56750) y cr22 obtenido de una mujer embarazada que porta un feto con trisomía 21.

La Fig. 64 muestra un perfil de recuentos sin procesar para los cr20, cr21 (de ~55750 a ~56750) y cr22 obtenido de una mujer embarazada que porta un feto euploide.

La Fig. 65 muestra un perfil de recuentos sin procesar para los cr20, cr21 (de ~55750 a ~56750) y cr22 obtenido de una mujer embarazada que porta un feto con trisomía 21.

La Fig. 66 muestra un perfil de recuentos normalizados para los cr20, cr21 (de ~55750 a ~56750) y cr22 obtenido de una mujer embarazada que porta un feto euploide.

La Fig. 67 muestra un perfil de recuentos normalizados para los cr20, cr21 (de ~55750 a ~56750) y cr22 obtenido de una mujer embarazada que porta un feto con trisomía 21.

La Fig. 68 muestra un perfil de recuentos normalizados para los cr20, cr21 (de ~47750 a ~48375) y cr22 obtenido de una mujer embarazada que porta un feto euploide.

La Fig. 69 muestra un perfil de recuentos normalizados para los cr20, cr21 (de ~47750 a ~48375) y cr22 obtenido de una mujer embarazada que porta un feto con trisomía 21.

La Fig. 70 muestra un gráfico de recuentos (eje y) frente al contenido de GC (eje X) antes de corrección de GC de LOESS (panel superior) y después de GC de LOESS (panel inferior).

La Fig. 71 muestra un gráfico de recuentos normalizados mediante GC de LOESS (eje Y) frente a la fracción de GC para múltiples muestras del cromosoma 1.

La Fig. 72 muestra un gráfico de recuentos normalizados mediante GC de LOESS y corregidos para la inclinación (eje Y) frente a la fracción de GC (eje X) para múltiples muestras del cromosoma 1.

La Fig. 73 muestra un gráfico de varianza (eje Y) frente a la fracción de GC (eje X) para el cromosoma 1 antes de la inclinación (línea en forma de V de color negro) y después de la inclinación (línea inferior de color gris).

La Fig. 74 muestra un gráfico de frecuencia (eje Y) frente a la fracción de GC (eje X) para el cromosoma así como una mediana (línea de color gris vertical izquierda) y la media (línea en negrita vertical derecha).

Las Fig. 75A-F muestran un gráfico de recuentos normalizados mediante GC de LOESS y corregidos para la inclinación (eje Y) frente a la fracción de GC (eje X), paneles izquierdos y la frecuencia (eje Y) frente a la fracción de GC (eje X) (paneles derechos) para los cromosomas 4, 15 y X (Fig. 75A, enumerados de arriba abajo), los cromosomas 5, 6 y 3 (Fig. 75B, enumerados de arriba abajo), los cromosomas 8, 2, 7 y 18 (Fig. 75C, enumerados de arriba abajo), los cromosomas 12, 14, 11 y 9 (Fig. 75D, enumerados de arriba abajo), los cromosomas 21, 1, 10, 15 y 20 (Fig. 75E, enumerados de arriba abajo) y los cromosomas 16, 17, 22 y 19 (Fig. 75F, enumerados de arriba abajo). La mediana de los valores (línea de color gris vertical izquierda) y los valores medios (línea en negrita vertical derecha) se indican en los paneles derechos.

La Fig. 76 muestra un gráfico de recuentos normalizados mediante GC de LOESS y corregidos para la inclinación (eje Y) frente a la fracción de GC (eje X) para el cromosoma 19. El pivote del cromosoma se muestra en las regiones recuadradas derechas y el pivote del genoma se muestra en la región recuadrada izquierda.

La Fig. 77 muestra un gráfico del valor de p (eje Y) frente a bins (eje X) para los cromosomas 13 (parte superior derecha), 21 (parte superior intermedia) y 18 (parte superior derecha). La posición cromosómica de determinados bins se muestra en el panel inferior.

La Fig. 78 muestra la puntuación Z para el cromosoma 21, en el que los bins no informativos se excluyeron del cálculo de la puntuación Z (eje Y) y la puntuación Z para el cromosoma 21 para todos los bins (eje X). Los casos de trisomía 21 se indican con círculos rellenos. Los euploides se indican mediante círculos en blanco.

La Fig. 79 muestra la puntuación Z para el cromosoma 18, en el que los bins no informativos se excluyeron del cálculo de la puntuación Z (eje Y) y la puntuación Z para el cromosoma 18 para todos los bins (eje X).

La Fig. 80 muestra un gráfico de bins seleccionados (eje Y) frente a todos los bins (eje X) para el cromosoma 18.

La Fig. 81 muestra un gráfico de bins seleccionados (eje Y) frente a todos los bins (eje X) para el cromosoma 21.

La Fig. 82 muestra un gráfico de recuentos (eje Y) frente al contenido de GC (eje X) para 7 muestras.

La Fig. 83 muestra un gráfico de recuentos sin procesar (eje Y) frente a coeficientes de sesgo de GC (eje X).

La Fig. 84 muestra un gráfico de frecuencia (eje Y) frente a ordenadas en el origen (eje X).

La Fig. 85 muestra un gráfico de frecuencia (eje Y) frente a pendientes (eje X).

La Fig. 86 muestra un gráfico de log de mediana de recuento (eje Y) frente a log de ordenada en el origen (eje X).

La Fig. 87 muestra un gráfico de frecuencia (eje Y) frente a pendiente (eje X).

La Fig. 88 muestra un gráfico de frecuencia (eje Y) frente al contenido de GC (eje X).

La Fig. 89 muestra un gráfico de pendiente (eje Y) frente al contenido de GC (eje X).

La Fig. 90 muestra un gráfico de los errores de validación cruzada (eje Y) frente al trabajo R (eje X) para los bins cr2_2404.

La Fig. 91 muestra un gráfico de los errores de validación cruzada (eje Y) frente al trabajo R (eje X) (parte superior izquierda), recuentos sin procesar (eje Y) frente a coeficientes de sesgo de GC (eje X) (parte superior derecha), frecuencia (eje Y) frente a ordenadas en el origen (eje X) (parte inferior izquierda), y frecuencia (eje Y) frente a pendiente (eje X) (parte inferior derecha) para los bins cr2_2345.

La Fig. 92 muestra un gráfico de errores de validación cruzada (eje Y) frente a trabajo R (eje X) (parte superior izquierda), recuentos sin procesar (eje Y) frente a coeficientes de sesgo de GC (eje X) (parte superior derecha), frecuencia (eje Y) frente a ordenadas en el origen (eje X) (parte inferior izquierda), y frecuencia (eje Y) frente a pendiente (eje X) (parte inferior derecha) para los bins cr1_31.

La Fig. 93 muestra un gráfico de los errores de validación cruzada (eje Y) frente al trabajo R (eje X) (parte superior izquierda), recuentos sin procesar (eje Y) frente a coeficientes de sesgo de GC (eje X) (parte superior derecha), frecuencia (eje Y) frente a ordenadas en el origen (eje X) (parte inferior izquierda), y frecuencia (eje Y) frente a pendiente (eje X) (parte inferior derecha) para los bins cr1_10.

La Fig. 94 muestra un gráfico de los errores de validación cruzada (eje Y) frente al trabajo R (eje X) (parte superior izquierda), recuentos sin procesar (eje Y) frente a coeficientes de sesgo de GC (eje X) (parte superior derecha), frecuencia (eje Y) frente a ordenadas en el origen (eje X) (parte inferior izquierda), y frecuencia (eje Y) frente a pendiente (eje Y) (parte inferior derecha) para los bins cr1_9.

La Fig. 95 muestra un gráfico de errores de validación cruzada (eje Y) frente a trabajo R (eje X) (parte superior izquierda), recuentos sin procesar (eje Y) frente a coeficientes de sesgo de GC (eje X) (parte superior derecha), frecuencia (eje Y) frente a ordenadas en el origen (eje X) (parte inferior izquierda), y frecuencia (eje Y) frente a pendiente (eje X) (parte inferior derecha) para los bins cr1_8.

La Fig. 96 muestra un gráfico de frecuencia (eje Y) frente a (R_{cv}, $R_{trabajo}$) máx. (eje X).

La Fig. 97 muestra un gráfico de réplicas técnicas (eje X) frente a log10 de errores de validación cruzada (eje X).

La Fig. 98 muestra un gráfico de separación de brecha de puntuación Z (eje Y) frente al umbral de error de validación cruzada (eje X) para el cr21.

La Fig. 99A (todos los bins) y la Fig. 99B (bins validados de forma cruzada) demuestra que la selección de bins descrita en el ejemplo 4 elimina principalmente los bins con baja capacidad de mapeo.

La Fig. 100 muestra un gráfico de recuentos normalizados (eje Y) frente al sesgo de GC (eje X) para el cr18_6.

La Fig. 101 muestran un gráfico de recuentos normalizados (eje Y) frente al sesgo de GC (eje X) para el cr18_8.

La Fig. 102 muestra un histograma de frecuencia (eje Y) frente al error en la ordenada en el origen (eje X).

La Fig. 103 muestra un histograma de frecuencia (eje Y) frente al error en la pendiente (eje X).

La Fig. 104 muestra un gráfico del error en la pendiente (eje Y) frente a la ordenada en el origen (eje X).

La Fig. 105 muestra un perfil normalizado que incluye el cr4 (de aproximadamente 12400 a aproximadamente 15750) con elevación (eje Y) y número de bins (eje X).

La Fig. 106 muestra un perfil de recuentos sin procesar (panel superior) y recuentos normalizados (panel inferior) para los cr20, cr21 y cr22. También se muestra una distribución de desviaciones estándar (eje X) frente a la frecuencia (eje Y) para los perfiles antes (parte superior) y después (parte inferior) de la normalización PERUN.

La Fig. 107 muestra una distribución de representaciones cromosómicas para casos de euploides y de trisomía para recuentos sin procesar (parte superior), enmascaramiento repetido (parte central) y recuentos normalizados (parte inferior).

La Fig. 108 muestra un gráfico de los resultados obtenidos con un modelo aditivo lineal (eje Y) frente a un GCRM para el cr13.

La Fig. 109 muestra un gráfico de los resultados obtenidos con un modelo aditivo lineal (eje Y) frente a un GCRM para el cr18.

La Fig. 110 y la Fig. 111 muestran un gráfico de resultados obtenidos con un modelo aditivo lineal (eje Y) frente a un GCRM para el cr21.

Las Fig. 112A-C ilustran el relleno de un perfil autosómico normalizado para una muestra de WI euploide. La Fig. 112A es un ejemplo de un perfil sin relleno. La Fig. 112B es un ejemplo de un perfil con relleno. La Fig. 112C es un ejemplo de una corrección de relleno (por ejemplo, un perfil ajustado, una elevación ajustada).

Las Fig. 113A-C ilustran el relleno de un perfil autosómico normalizado para una muestra de WI euploide. La Fig. 113A es un ejemplo de un perfil sin relleno. La Fig. 113B es un ejemplo de un perfil con relleno. La Fig. 113C es un ejemplo de una corrección de relleno (por ejemplo, un perfil ajustado, una elevación ajustada).

Las Fig. 114A-C ilustran el relleno de un perfil autosómico normalizado para una muestra de WI con trisomía 13. La Fig. 114A es un ejemplo de un perfil sin relleno. La Fig. 114B es un ejemplo de un perfil con relleno. La Fig. 114C es un ejemplo de una corrección de relleno (por ejemplo, un perfil ajustado, una elevación ajustada).

Las Fig. 115A-C ilustran el relleno de un perfil autosómico normalizado para una muestra de WI con trisomía 18. La Fig. 115A es un ejemplo de un perfil sin relleno. La Fig. 115B es un ejemplo de un perfil con relleno. La Fig. 115C es un ejemplo de una corrección de relleno (por ejemplo, un perfil ajustado, una elevación ajustada).

Las Fig. 116-120, 122, 123, 126, 128, 129 y 131 muestran una duplicación materna dentro de un perfil.

Las Fig. 121, 124, 125, 127 y 130 muestran una deleción materna dentro de un perfil.

Descripción detallada

Se proporcionan métodos, procedimientos y aparatos útiles para identificar una variación genética. Identificar una variación genética comprende a veces detectar una variación del número de copias y/o comprende a veces ajustar una elevación que comprende una variación del número de copias. En algunas implementaciones, se ajusta una elevación que proporciona una identificación de una o más varianzas o variaciones genéticas con una probabilidad reducida de un diagnóstico de falso positivo o falso negativo. En algunas implementaciones, identificar una variación genética mediante un método descrito en el presente documento puede conducir a un diagnóstico de, o determinar una predisposición a, una afección médica particular. La identificación de una varianza genética puede dar como resultado que se facilite una decisión médica y/o se emplee un procedimiento médico útil.

Muestras

En el presente documento se proporcionan métodos y composiciones para analizar ácido nucleico. En algunas implementaciones, se analizan los fragmentos de ácido nucleico en una mezcla de fragmentos de ácido nucleico. Una mezcla de ácidos nucleicos puede comprender dos o más especies de fragmento de ácido nucleico que tienen secuencias de nucleótidos diferentes, longitudes de fragmento diferentes, orígenes diferentes (por ejemplo, orígenes genómicos, orígenes fetales frente a maternos, orígenes celulares o tisulares, orígenes de muestra, orígenes de sujeto, y similares), o combinaciones de los mismos.

El ácido nucleico o una mezcla de ácido nucleico usados en los métodos y aparatos descritos en el presente documento se aísla a menudo de una muestra obtenida de un sujeto. Un sujeto puede ser cualquier organismo vivo o no vivo incluyendo, pero sin limitarse a, un ser humano, un animal no humano, una planta, una bacteria, un hongo o un protista. Puede seleccionarse cualquier animal humano o no humano incluyendo, pero sin limitarse a, mamífero, reptil, ave, anfibio, pez, ungulado, rumiante, bovino (por ejemplo, ganado vacuno), equino (por ejemplo, caballo), caprino y ovino (por ejemplo, oveja, cabra), porcino (por ejemplo, cerdo), camélido (por ejemplo, camello, llama, alpaca), mono, simio (por ejemplo, gorila, chimpancé), úrsido (por ejemplo, oso), ave de corral, perro, gato, ratón, rata, pescado, delfín, ballena y tiburón. Un sujeto puede ser de sexo masculino o femenino (por ejemplo, mujer).

El ácido nucleico puede aislarse de cualquier tipo de espécimen o muestra biológica adecuada (por ejemplo, una muestra de prueba). Una muestra o muestra de prueba puede ser cualquier espécimen aislado u obtenido de un sujeto (por ejemplo, un sujeto humano, una mujer embarazada). Los ejemplos no limitativos de muestras incluyen líquido o tejido de un sujeto incluyendo, sin limitación, sangre de cordón umbilical, vellosidades coriónicas, líquido amniótico, líquido cefalorraquídeo, líquido espinal, líquido de lavado (por ejemplo broncoalveolar, gástrico, peritoneal, canalicular, del oído, artroscópico), muestra de biopsia (por ejemplo, de embrión antes de la implantación), muestra de celocentesis, células nucleadas fetales o restos celulares fetales, lavados del aparato reproductor femenino, orina, heces, esputo, saliva, mucosidad nasal, líquido prostático, lavado, semen, líquido linfático, bilis, lágrimas, sudor, leche materna, líquido mamario, células embrionarias y células fetales (por ejemplo, células placentarias). En algunas implementaciones, una muestra biológica es un hisopo cervicouterino de un sujeto. En algunas implementaciones, una muestra biológica puede

ser sangre y, algunas veces, plasma o suero. Tal como se usa en el presente documento, el término "sangre" abarca sangre completa o cualquier fracción de sangre, tal como suero y plasma tal como se definen convencionalmente, por ejemplo. La sangre o fracciones de la misma comprenden a menudo nucleosomas (por ejemplo, nucleosomas maternos y/o fetales). Los nucleosomas comprenden ácidos nucleicos y, algunas veces, están libres de células o son intracelulares. La sangre comprende además capas leucocíticas. Algunas veces, las capas leucocíticas se aíslan utilizando un gradiente de Ficoll. Las capas leucocíticas pueden comprender glóbulos blancos (por ejemplo, leucocitos, linfocitos T, linfocitos B, plaquetas, y similares). Algunas veces, las capas leucocíticas comprenden ácido nucleico materno y/o fetal. Plasma sanguíneo se refiere a la fracción de sangre completa que resulta de la centrifugación de sangre tratada con anticoagulantes. Suero sanguíneo se refiere a la porción acuosa del fluido que queda después de que se ha coagulado una muestra de sangre. A menudo, se recogen muestras de líquido o tejido según protocolos convencionales que siguen generalmente hospitales o clínicas. A menudo, para la sangre se recoge una cantidad adecuada de sangre periférica (por ejemplo, entre 3-40 mililitros) y puede almacenarse según procedimientos convencionales antes o después de la preparación. Una muestra de líquido o tejido de la que se extrae ácido nucleico puede ser acelular (por ejemplo, libre de células). En algunas implementaciones, una muestra de líquido o tejido puede contener elementos celulares o restos celulares. En algunas implementaciones, pueden incluirse células fetales o células cancerosas en la muestra.

Una muestra es a menudo heterogénea, lo que significa que más de un tipo de especie de ácido nucleico está presente en la muestra. Por ejemplo, el ácido nucleico heterogéneo puede incluir, pero no se limita a, (i) ácido nucleico derivado fetal y derivado materno, (ii) ácido nucleico de cáncer y distinto de cáncer, (iii) ácido nucleico patógeno y de huésped y, más generalmente, (iv) ácido nucleico mutado y de tipo natural. Una muestra puede ser heterogénea porque más de un tipo de célula está presente, tal como una célula fetal y una célula materna, una célula cancerosa y no cancerosa, o una célula patógena y de huésped. En algunas implementaciones, están presente una especie minoritaria de ácido nucleico y una ucleico.

Para las aplicaciones prenatales de la tecnología descrita en el presente documento, la muestra de líquido o tejido puede recogerse de una mujer a una edad gestacional adecuada para la prueba, o de una mujer que está sometiéndose a prueba para un posible embarazo. La edad gestacional adecuada puede variar dependiendo de la prueba prenatal que se realiza. En determinadas implementaciones, un sujeto de sexo femenino gestante algunas veces está en el primer trimestre de embarazo, a veces en el segundo trimestre de embarazo o, algunas veces, en el tercer trimestre de embarazo. En determinadas implementaciones, se recoge un líquido o tejido de una mujer embarazada de aproximadamente 1 a aproximadamente 45 semanas de gestación fetal (por ejemplo, a las 1-4, 4-8, 8-12, 12-16, 16-20, 20-24, 24-28, 28-32, 32-36, 36-40 o 40-44 semanas de gestación fetal), y a veces de aproximadamente 5 a aproximadamente 28 semanas de gestación fetal (por ejemplo, a las 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26 o 27 semanas de gestación fetal). Algunas veces, se recoge una muestra de líquido o tejido de una mujer embarazada durante o justo después (por ejemplo, de 0 a 72 horas después) de dar a luz (por ejemplo, parto vaginal o no vaginal (por ejemplo, parto quirúrgico)).

Aislamiento y procesamiento de ácidos nucleicos

El ácido nucleico puede derivar de una o más fuentes (por ejemplo, células, suero, plasma, capa leucocítica, líquido linfático, piel, suelo, y similares) mediante métodos conocidos en la técnica. Los procedimientos y reactivos de lisis celular se conocen en la técnica y pueden realizarse generalmente mediante métodos químicos (por ejemplo, detergentes, disoluciones hipotónicas, procedimientos enzimáticos, y similares, o una combinación de los mismos), físicos (por ejemplo, prensa francesa, sonicación, y similares) o electrolíticos de lisis. Puede utilizarse cualquier procedimiento de lisis adecuado. Por ejemplo, los métodos químicos emplean generalmente agentes de lisis para perturbar las células y extraer los ácidos nucleicos de las células, seguido por el tratamiento con sales caotrópicas. Además, son útiles los métodos físicos, tales como congelación/descongelación seguida de trituración, el uso de prensas celulares y similares. Además, se usan habitualmente procedimientos de lisis con alto contenido de sal. Por ejemplo, puede utilizarse un procedimiento de lisis alcalina. Este último procedimiento incorpora tradicionalmente el uso de disoluciones de fenol-cloroformo, y puede usarse un procedimiento alternativo libre de fenol-cloroformo que involucra tres disoluciones. En estos últimos procedimientos, una disolución puede contener Tris 15 mM, pH 8,0; EDTA 10 mM y ARNasa A 100 ug/ml; una segunda disolución puede contener NaOH 0,2 N y SDS al 1 %; y una tercera disolución puede contener KOAc 3 M, pH 5,5. Estos procedimientos pueden encontrarse en Current Protocols in Molecular Biology, John Wiley & Sons, N.Y., 6.3.1-6.3.6 (1989).

Los términos "ácido nucleico" y "molécula de ácido nucleico" se usan indistintamente. Los términos se refieren a ácidos nucleicos de cualquier forma de composición, tales como ácido desoxirribonucleico (ADN, por ejemplo, ADN complementario (ADNc), ADN genómico (ADNg) y similares), ácido ribonucleico (ARN, por ejemplo, ARN mensajero (ARNm), ARN inhibidor corto (ARNIc), ARN ribosómico (ARNr), ARN de transferencia (ARNt), microARN, ARN altamente expresado por el feto o la placenta, y similares), y/o análogos de ADN o ARN (por ejemplo, que contienen análogos de base, análogos de azúcar y/o una cadena principal no nativa y similares), híbridos de ARN/ADN y ácidos nucleicos de poliamida (PNA), todos los cuales pueden estar en forma monocatenaria o bicatenaria. A menos que se limite de cualquier otra manera, un ácido nucleico puede comprender análogos conocidos de nucleótidos naturales, algunos de los cuales pueden funcionar de manera similar a los nucleótidos que se producen de manera natural. Un ácido nucleico puede estar en cualquier forma útil para realizar los procedimientos de la presente invención (por ejemplo, lineal, circular, superenrollado, monocatenario, bicatenario y similares). Un ácido

nucleico puede ser, o puede proceder de, un plásmido, fago, una secuencia de replicación autónoma (ARS), un centrómero, cromosoma artificial, cromosoma u otro ácido nucleico capaz de replicar o replicarse *in vitro* o en una célula huésped, una célula, un núcleo celular o un citoplasma de una célula en determinadas implementaciones. Un ácido nucleico en algunas implementaciones puede ser de un solo cromosoma o fragmento del mismo (por ejemplo, una muestra de ácido nucleico puede proceder de un cromosoma de una muestra obtenida de un organismo diploide). Algunas veces, los ácidos nucleicos comprenden nucleosomas, fragmentos o partes de nucleosomas o estructuras similares a nucleosomas. Los ácidos nucleicos comprenden, algunas veces, proteína (por ejemplo, histonas, proteínas de unión a ADN y similares). Los ácidos nucleicos analizados mediante procedimientos descritos en el presente documento, algunas veces, están sustancialmente aislados y no están sustancialmente asociados con proteínas u otras moléculas. Los ácidos nucleicos incluyen además derivados, variantes y análogos de ARN o ADN sintetizados, replicados o amplificados a partir de polinucleótidos monocatenarios ("sentido" o "antisentido", hebra "positiva" o hebra "negativa", marco de lectura "directo" o marco de lectura "inverso") y polinucleótidos bicatenarios. Los desoxirribonucleótidos incluyen desoxiadenosina, desoxicitidina, desoxiguanosina y desoxitimidina. Para el ARN, la base citosina se reemplaza por uracilo y la posición 2' del azúcar incluye un resto hidroxilo. Un ácido nucleico puede prepararse con el uso de un ácido nucleico obtenido de un sujeto como molde.

El ácido nucleico puede aislarse en un punto de tiempo diferente en comparación con otro ácido nucleico, en el que cada una de las muestras procede de la misma fuente o de una fuente diferente. Un ácido nucleico puede ser de una biblioteca de ácido nucleico, tal como una biblioteca de ADNc o ARN, por ejemplo. Un ácido nucleico puede ser un resultado de la purificación o el aislamiento de ácido nucleico y/o la amplificación de moléculas de ácido nucleico de la muestra.

El ácido nucleico proporcionado para los procedimientos descritos en el presente documento puede contener ácido nucleico de una muestra o de dos o más muestras (por ejemplo, de 1 o más, 2 o más, 3 o más, 4 o más, 5 o más, 6 o más, 7 o más, 8 o más, 9 o más, 10 o más, 11 o más, 12 o más, 13 o más, 14 o más, 15 o más, 16 o más, 17 o más, 18 o más, 19 o más, o 20 o más muestras).

Los ácidos nucleicos pueden incluir ácido nucleico extracelular en determinadas implementaciones. El término "ácido nucleico extracelular", tal como se usa en el presente documento, puede referirse a ácido nucleico aislado de una fuente que no tiene sustancialmente células y también se denomina ácido nucleico "libre de células" y/o ácido nucleico "circulante, libre de células". El ácido nucleico extracelular puede estar presente en y obtenerse de sangre (por ejemplo, de la sangre de una mujer embarazada). El ácido nucleico extracelular incluye a menudo células no detectables y puede contener elementos celulares o restos celulares. Los ejemplos no limitativos de fuentes acelulares de ácido nucleico extracelular son sangre, plasma sanguíneo, suero sanguíneo y orina. Tal como se usa en el presente documento, la expresión "obtener ácido nucleico de muestra circulante, libre de células" incluye obtener directamente una muestra (por ejemplo, recoger una muestra, por ejemplo, una muestra de prueba) u obtener una muestra de otra que ha recogido una muestra. Sin desear limitarse por la teoría, el ácido nucleico extracelular puede ser un producto de la apoptosis celular y la descomposición celular, lo que proporciona una base para el ácido nucleico extracelular que tiene a menudo una serie de longitudes a través de un espectro (por ejemplo, una "escalera").

El ácido nucleico extracelular puede incluir diferentes especies de ácido nucleico y, por tanto, se denomina en el presente documento "heterogéneo" en determinadas implementaciones. Por ejemplo, el suero o plasma sanguíneo de una persona que tiene cáncer puede incluir ácido nucleico de células cancerosas y ácido nucleico de células no cancerosas. En otro ejemplo, el suero o plasma sanguíneo de una mujer embarazada puede incluir ácido nucleico materno y ácido nucleico fetal. En algunos casos, el ácido nucleico fetal algunas veces es de aproximadamente el 5 % a aproximadamente el 50 % del ácido nucleico global (por ejemplo, aproximadamente el 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48 o el 49 % del ácido nucleico total es ácido nucleico fetal). En algunas implementaciones, la mayoría del ácido nucleico fetal en el ácido nucleico tiene una longitud de aproximadamente 500 pares de bases o menos (por ejemplo, aproximadamente el 80, 85, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 o el 100 % del ácido nucleico fetal tiene una longitud de aproximadamente 500 pares de bases o menos). En algunas implementaciones, la mayoría del ácido nucleico fetal en el ácido nucleico tiene una longitud de aproximadamente 250 pares de bases o menos (por ejemplo, aproximadamente el 80, 85, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 o el 100 % del ácido nucleico fetal tiene una longitud de aproximadamente 250 pares de bases o menos). En algunas implementaciones, la mayoría del ácido nucleico fetal en el ácido nucleico tiene una longitud de aproximadamente 200 pares de bases o menos (por ejemplo, aproximadamente el 80, 85, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 o el 100 % del ácido nucleico fetal tiene una longitud de aproximadamente 200 pares de bases o menos). En algunas implementaciones, la mayoría del ácido nucleico fetal en el ácido nucleico tiene una longitud de aproximadamente 150 pares de bases o menos (por ejemplo, aproximadamente el 80, 85, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 o el 100 % del ácido nucleico fetal tiene una longitud de aproximadamente 150 pares de bases o menos). En algunas implementaciones, la mayoría del ácido nucleico fetal en el ácido nucleico tiene una longitud de aproximadamente 100 pares de bases o menos (por ejemplo, aproximadamente el 80, 85, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 o el 100 % del ácido nucleico fetal tiene una longitud de aproximadamente 100 pares de bases o menos). En algunas implementaciones, la mayoría del ácido nucleico fetal en el ácido nucleico tiene una longitud de aproximadamente 50 pares de bases o menos (por ejemplo, aproximadamente el 80, 85, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 o el 100 % del ácido nucleico fetal tiene una longitud de aproximadamente 50 pares de bases o menos). En algunas implementaciones, la mayoría del ácido nucleico fetal en el ácido nucleico tiene una longitud de aproximadamente 25 pares de bases o menos (por ejemplo, aproximadamente el 80, 85, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 o el 100 % del ácido nucleico fetal tiene una longitud de aproximadamente 25 pares de bases o menos).

El ácido nucleico podría proporcionarse para realizar los métodos descritos en el presente documento sin procesar la(s) muestra(s) que contiene(n) el ácido nucleico, en determinadas implementaciones. En algunas implementaciones, se proporciona ácido nucleico para llevar a cabo los métodos descritos en el presente documento después del procesamiento de la(s) muestra(s) que contiene(n) el ácido nucleico. Por ejemplo, un ácido nucleico puede extraerse, aislarse, purificarse, purificarse parcialmente o amplificarse a partir de la(s) muestra(s). El término ‘‘aislado’’, tal como se usa en el presente documento, se refiere a un ácido nucleico retirado de su entorno original (por ejemplo, el entorno natural si se produce de manera natural o una célula huésped si se expresa de manera exógena) y, por tanto, se altera por intervención humana (por ejemplo, ‘‘por la mano del hombre’’) de su entorno original. El término ‘‘ácido nucleico aislado’’, tal como se usa en el presente documento, puede referirse a un ácido nucleico retirado de un sujeto (por ejemplo, un sujeto humano). Un ácido nucleico aislado puede proporcionarse con menos componentes distintos de ácido nucleico (por ejemplo, proteína, lípido) que la cantidad de componentes presentes en una muestra de origen. Una composición que comprende ácido nucleico aislado puede estar libre en de aproximadamente el 50 % a más del 99 % de componentes distintos de ácido nucleico. Una composición que comprende ácido nucleico aislado puede estar libre en aproximadamente el 90 %, 91 %, 92 %, 93 %, 94 %, 95 %, 96 %, 97 %, 98 %, 99 % o más del 99 % de componentes distintos de ácido nucleico. El término ‘‘purificado’’, tal como se usa en el presente documento, puede referirse a un ácido nucleico siempre que contenga menos componentes distintos de ácido nucleico (por ejemplo, proteína, lípido, hidrato de carbono) que la cantidad de componentes distintos de ácido nucleico presentes antes de someter el ácido nucleico a un procedimiento de purificación. Una composición que comprende ácido nucleico purificado puede estar libre en aproximadamente el 80 %, 81 %, 82 %, 83 %, 84 %, 85 %, 86 %, 87 %, 88 %, 89 %, 90 %, 91 %, 92 %, 93 %, 94 %, 95 %, 96 %, 97 %, 98 %, 99 % o más del 99 % de otros componentes distintos de ácido nucleico. El término ‘‘purificado’’, tal como se usa en el presente documento, puede referirse a un ácido nucleico siempre que contenga menos especies de ácido nucleico que en la fuente de muestra de la que se deriva el ácido nucleico. Una composición que comprende ácido nucleico purificado puede estar libre en aproximadamente el 90 %, 91 %, 92 %, 93 %, 94 %, 95 %, 96 %, 97 %, 98 %, 99 % o más del 99 % de otras especies de ácido nucleico. Por ejemplo, el ácido nucleico fetal puede purificarse a partir de una mezcla que comprende ácido nucleico materno y fetal. En determinados ejemplos, los nucleosomas que comprenden fragmentos pequeños de ácido nucleico fetal pueden purificarse a partir de una mezcla de complejos de nucleosomas más grandes que comprenden fragmentos más grandes de ácido nucleico materno.

El término ‘‘amplificado’’, tal como se usa en el presente documento, se refiere a someter un ácido nucleico diana en una muestra a un procedimiento que genera lineal o exponencialmente ácidos nucleicos de amplicón que tienen la misma secuencia de nucleótidos o prácticamente la misma secuencia de nucleótidos que el ácido nucleico diana o segmento del mismo. El término ‘‘amplificado’’, tal como se usa en el presente documento, puede referirse a someter un ácido nucleico diana (por ejemplo, en una muestra que comprende otros ácidos nucleicos) a un procedimiento que genera selectiva y lineal o exponencialmente ácidos nucleicos de amplicón que tienen la misma o prácticamente la misma secuencia de nucleótidos que el ácido nucleico diana, o segmento del mismo. El término ‘‘amplificado’’, tal como se usa en el presente documento, puede referirse a someter una población de ácidos nucleicos a un procedimiento que genera de manera no selectiva y lineal o exponencial ácidos nucleicos de amplicón que tienen la misma secuencia de nucleótidos o sustancialmente la misma secuencia de nucleótidos que los ácidos nucleicos, o porciones de los mismos, que estaban presentes en la muestra antes de la amplificación. Algunas veces, el término ‘‘amplificado’’ se refiere a un método que comprende una reacción en cadena de la polimerasa (PCR).

El ácido nucleico puede procesarse además al someter el ácido nucleico a un método que genera fragmentos de ácido nucleico, en determinadas implementaciones, antes de proporcionar ácido nucleico para un procedimiento descrito en el presente documento. En algunas implementaciones, el ácido nucleico sometido a fragmentación o escisión puede tener una longitud nominal, promedio o media de aproximadamente 5 a aproximadamente 10.000 pares de bases, de aproximadamente 100 a aproximadamente 1000 pares de bases, de aproximadamente 100 a aproximadamente 500 pares de bases, o aproximadamente 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000 o 9000 pares de bases. Los fragmentos pueden generarse mediante un método adecuado conocido en la técnica, y la longitud promedio, media o nominal de los fragmentos de ácido nucleico puede controlarse mediante la selección de un procedimiento de generación de fragmentos adecuado. En determinadas implementaciones, puede usarse un ácido nucleico de una longitud relativamente más corta para analizar secuencias que contienen poca variación de secuencia y/o contienen cantidades relativamente grandes de información de secuencia de nucleótidos conocida. En algunas implementaciones, puede usarse un ácido nucleico de una longitud relativamente más larga para analizar secuencias que contienen una mayor variación de secuencia y/o cantidades relativamente pequeñas de información de secuencia de nucleótidos.

Los fragmentos de ácido nucleico pueden contener secuencias de nucleótidos solapantes, y esas secuencias solapantes pueden facilitar la construcción de una secuencia de nucleótidos del ácido nucleico homólogo no fragmentado o un segmento del mismo. Por ejemplo, un fragmento puede tener subsecuencias x e y y otro fragmento puede tener subsecuencias y y z, en las que x, y y z son secuencias de nucleótidos que pueden tener 5 nucleótidos de longitud o más. La secuencia solapante y puede utilizarse para facilitar la construcción de la secuencia de nucleótidos x-y-z en ácido nucleico a partir de una muestra en determinadas implementaciones. El ácido nucleico puede estar parcialmente fragmentado (por ejemplo, a partir de una reacción de escisión específica incompleta o terminada) o completamente fragmentado en determinadas implementaciones.

El ácido nucleico puede fragmentarse mediante diversos métodos conocidos en la técnica, incluyendo, pero sin limitación, procedimientos físicos, químicos y enzimáticos. Se describen ejemplos no limitativos de tales procedimientos en la publicación de solicitud de patente estadounidense n.° 20050112590 (publicada el 26 de mayo de 2005, titulada "Fragmentation-based methods and systems for sequence variation detection and discovery", que nombra a Van Den Boom *et al.*). Determinados procedimientos pueden seleccionarse para generar fragmentos escindidos de manera inespecífica o fragmentos escindidos de manera específica. Los ejemplos no limitativos de procedimientos que pueden generar ácido nucleico de fragmentos escindidos de manera inespecífica incluyen, sin limitación, poner en contacto ácido nucleico con el aparato que expone el ácido nucleico a una fuerza de cizallamiento (por ejemplo, haciendo pasar el ácido nucleico a través de una aguja de jeringa; uso de una prensa francesa); exponer el ácido nucleico a irradiación (por ejemplo, irradiación gamma, rayos X, UV; los tamaños de los fragmentos pueden controlarse mediante la intensidad de irradiación); el ácido nucleico en ebullición en agua (por ejemplo, produce fragmentos de aproximadamente 500 pares de bases) y exponer el ácido nucleico a un procedimiento de hidrólisis con ácido y base.

Tal como se usa en el presente documento, "fragmentación" o "escisión" se refiere a un procedimiento o condiciones en los que una molécula de ácido nucleico, tal como una molécula de gen molde de ácido nucleico o producto amplificado de la misma, puede cortarse en dos o más moléculas de ácido nucleico más pequeñas. Tal fragmentación o escisión puede ser específica de secuencia, específica de base o inespecífica, y puede lograrse mediante cualquiera de una variedad de métodos, reactivos o condiciones incluyendo, por ejemplo, fragmentación química, enzimática o física.

Tal como se usa en el presente documento, "fragmentos", "productos de escisión", "productos escindidos" o variantes gramaticales de los mismos, se refieren a moléculas de ácido nucleico resultantes de una fragmentación o escisión de una molécula de gen molde de ácido nucleico o producto amplificado de la misma. Aunque tales fragmentos o productos escindidos pueden referirse a todas las moléculas de ácido nucleico resultantes de una reacción de escisión normalmente tales fragmentos o productos escindidos se refieren solamente a moléculas de ácido nucleico resultantes de una fragmentación o escisión de una molécula de gen molde de ácido nucleico o el segmento de un producto amplificado de la misma que contiene la secuencia de nucleótidos correspondiente de una molécula de gen molde de ácido nucleico. Por ejemplo, un producto amplificado puede contener uno o más nucleótidos más que la región de nucleótidos amplificada de una secuencia molde de ácido nucleico (por ejemplo, un cebador puede contener nucleótidos "extra" tales como una secuencia de iniciación de la transcripción, además de los nucleótidos complementarios a una molécula de gen molde de ácido nucleico, lo que da como resultado un producto amplificado que contiene nucleótidos "extra" o nucleótidos que no corresponden a la región de nucleótidos amplificada de la molécula de gen molde de ácido nucleico). En consecuencia, los fragmentos pueden incluir fragmentos que surgen de porciones de moléculas de ácido nucleico amplificadas que contienen, al menos en parte, información de secuencia de nucleótidos de o basada en la molécula molde de ácido nucleico representativa.

Tal como se usa en el presente documento, la expresión "reacciones de escisión complementarias" se refiere a las reacciones de escisión que se llevan a cabo en el mismo ácido nucleico con el uso de diferentes reactivos de escisión o mediante la alteración de la especificidad de escisión del mismo reactivo de escisión de tal manera que se generan patrones de escisión alternativos del mismo ácido nucleico o proteína diana o de referencia. En determinadas implementaciones, el ácido nucleico puede tratarse con uno o más agentes de escisión específicos (por ejemplo, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 o más agentes de escisión específicos) en uno o más recipientes de reacción (por ejemplo, el ácido nucleico se trata con cada agente de escisión específico en un recipiente independiente).

El ácido nucleico puede escindirse de manera específica o escindirse de manera inespecífica al poner en contacto el ácido nucleico con uno o más agentes de escisión enzimática (por ejemplo, nucleasas, enzimas de restricción). La expresión "agente de escisión específico", tal como se usa en el presente documento, se refiere a un agente, algunas veces, una sustancia química o una enzima que puede escindir un ácido nucleico en uno o más sitios específicos. Los agentes de escisión específicos se escinden a menudo de manera específica según una secuencia de nucleótidos particular en un sitio particular. Los agentes de escisión inespecíficos escinden a menudo ácidos nucleicos en sitios inespecíficos o degradan ácidos nucleicos. Los agentes de escisión inespecíficos degradan a menudo ácidos nucleicos mediante la eliminación de nucleótidos del extremo (ya sea el extremo 5', el extremo 3' o ambos) de una hebra de ácido nucleico.

Cualquier agente de escisión enzimática inespecífico o específico adecuado puede usarse para escindir o fragmentar ácidos nucleicos. Una enzima de restricción adecuada puede usarse para escindir ácidos nucleicos, en algunas implementaciones. Los ejemplos de agentes de escisión enzimática incluyen, pero no se limitan a, endonucleasas (por ejemplo, ADNasa (por ejemplo, ADNasa I, II); ARNasa (por ejemplo, ARNasa E, F, H, P); enzima Cleavase™; ADN polimerasa Taq; ADN polimerasa I de *E. coli* y endonucleasas específicas de estructura eucariotas; endonucleasas FEN-1 murinas; endonucleasas de restricción de tipo I, II o III tales como Acc I, Afi III, Alu I, Alw44 I, Apa I, Asn I, Ava I, Ava II, BamH I, Ban II, Bel I, Bgl I. Bgl II, Bln I, Bsm I, BssH II, BstE II, Cfo I, Cla I, Dde I, Dpn I, Dra I, EclX I, EcoR I, EcoR I, EcoR II, EcoR V, Hae II, Hae II, Hind II, Hind III, Hpa I, Hpa II, Kpn I, Ksp I, Miu I, MluN I, Msp I, Nci I, Neo I, Nde I, Nde II, Nhe I, Not I, Nru I, Nsi I, Pst I, Pvu I, Pvu II, Rsa I, Sac I, Sal I, Sau3A I, Sea I, ScrF I, Sfi I, Sma I, Spe I, Sph I, Ssp I, Stu I, Sty I, Swa I, Taq I, Xba I, Xho I; glicosilasas (por ejemplo, uracilo-ADN glicosilasa (UDG), ADN-3-metiladenina glicosilasa, ADN-3-metiladenina glicosilasa II, hidrato de pirimidina-ADN glicosilasa, FaPy-ADN glicosilasa, apareamiento erróneo de timina-ADN glicosilasa, hipoxantina-ADN glicosilasa, 5-hidroximetiluracilo-ADN glicosilasa (HmllDG), 5- hidroximetilcitosina-ADN glicosilasa, o 1,N6-eteno-adenina-ADN glicosilasa);

exonucleasas (por ejemplo, exonucleasa III); ribozimas y enzimas de ADN. El ácido nucleico puede tratarse con un agente químico, y el ácido nucleico modificado puede escindirse. En ejemplos no limitativos, el ácido nucleico puede tratarse con (i) agentes alquilantes tales como metilnitrosourea que generan varias bases alquiladas, incluyendo N3-metiladenina y N3-metilguanina, reconocidas y escindidas por alquilpurina-ADN glicosilasa; (ii) bisulfito de sodio, que provoca la desaminación de residuos de citosina en el ADN para formar residuos de uracilo que pueden escindirse por uracilo N-glicosilasa; y (iii) un agente químico que convierte guanina en su forma oxidada, 8-hidroxiguanina, que puede escindirse por formamidopirimidina-ADN N-glicosilasa. Los ejemplos de procedimientos de escisión química incluyen, sin limitación, alquilación, (por ejemplo, alquilación de ácido nucleico modificado con fosforotioato); escisión de labilidad al ácido de ácido nucleico que contiene P3'-N5'-fosforoamidato; y tratamiento con tetróxido de osmio y piperidina de ácido nucleico.

El ácido nucleico puede exponerse además a un procedimiento que modifica determinados nucleótidos en el ácido nucleico antes de proporcionar ácido nucleico para un método descrito en el presente documento. Un procedimiento que modifica selectivamente el ácido nucleico basándose en el estado de metilación de nucleótidos en el mismo puede aplicarse al ácido nucleico, por ejemplo. Además, condiciones tales como alta temperatura, radiación ultravioleta, radiación de rayos X, pueden inducir cambios en la secuencia de una molécula de ácido nucleico. El ácido nucleico puede proporcionarse en cualquier forma útil para realizar un análisis de secuencias o procedimiento de fabricación descrito en el presente documento, tal como forma sólida o líquida, por ejemplo. En determinadas implementaciones, el ácido nucleico puede proporcionarse en una forma líquida que comprende opcionalmente uno o más de otros componentes incluyendo, sin limitación, uno o más tampones o sales.

El ácido nucleico puede ser monocatenario o bicatenario. El ADN monocatenario, por ejemplo, puede generarse mediante la desnaturalización de ADN bicatenario mediante calentamiento o mediante tratamiento con álcali, por ejemplo. En algunos casos, el ácido nucleico se encuentra en una estructura de bucle D, formada mediante invasión de hebra de una molécula de ADN dúplex por un oligonucleótido o una molécula similar a ADN, tal como ácido nucleico peptídico (PNA). La formación del bucle D puede facilitarse mediante la adición de la proteína RecA de *E. coli* y/o mediante la alteración de la concentración de sal, por ejemplo, usando métodos conocidos en la técnica.

Determinación del contenido de ácido nucleico fetal

La cantidad de ácido nucleico fetal (por ejemplo, concentración, cantidad relativa, cantidad absoluta, número de copias y similares) en ácido nucleico se determina en algunas implementaciones. En algunos casos, la cantidad de ácido nucleico fetal en una muestra se denomina "fracción fetal". Algunas veces, "fracción fetal" se refiere a la fracción de ácido nucleico fetal en ácido nucleico circulante, libre de células en una muestra (por ejemplo, una muestra de sangre, una muestra de suero, una muestra de plasma) obtenida de una mujer embarazada. En determinadas implementaciones, la cantidad de ácido nucleico fetal se determina según marcadores específicos para un feto de sexo masculino (por ejemplo, marcadores STR de cromosoma Y (por ejemplo, marcadores DYS 19, DYS 385, DYS 392); marcador RhD en mujeres negativas para RhD), razones alélicas de secuencias polimórficas, o según uno o más marcadores específicos de ácido nucleico fetal y no de ácido nucleico materno (por ejemplo, biomarcadores epigenéticos diferenciales (por ejemplo, metilación; descritos con mayor detalle más adelante) entre la madre y el feto, o marcadores de ARN fetal en plasma sanguíneo materno (véase por ejemplo, Lo, 2005, Journal of Histochemistry and Cytochemistry 53 (3): 293-296)).

La determinación del contenido de ácido nucleico fetal (por ejemplo, fracción fetal) se realiza, algunas veces, usando un ensayo cuantificador fetal (FQA) tal como se describe, por ejemplo, en la publicación de solicitud de patente estadounidense n.° 2010/0105049. Este tipo de ensayo permite la detección y cuantificación de ácido nucleico fetal en una muestra materna basándose en el estado de metilación del ácido nucleico en la muestra. En algunos casos, la cantidad de ácido nucleico fetal de una muestra materna puede determinarse con respecto a la cantidad total de ácido nucleico presente, lo que proporciona de ese modo el porcentaje de ácido nucleico fetal en la muestra. En algunos casos, el número de copias de ácido nucleico fetal puede determinarse en una muestra materna. En algunos casos, la cantidad de ácido nucleico fetal puede determinarse de manera específica de secuencia (o específica de locus) y algunas veces con suficiente sensibilidad para permitir un análisis preciso de la dosificación cromosómica (por ejemplo, para detectar la presencia o ausencia de una aneuploidía fetal).

Un ensayo cuantificador fetal (FQA) puede realizarse junto con cualquiera de los métodos descritos en el presente documento. Tal ensayo puede realizarse mediante cualquier método conocido en la técnica y/o descrito en la publicación de solicitud de patente estadounidense n.° 2010/0105049, tal como, por ejemplo, mediante un método que puede distinguir entre ADN materno y fetal basándose en el estado de metilación diferencial, y cuantificar (es decir, determinar la cantidad de) el ADN fetal. Los métodos para diferenciar ácido nucleico basados en el estado de metilación incluyen, pero no se limitan a, captura sensible a la metilación, por ejemplo, usando un fragmento Fc de MBD2 en el cual el dominio de unión a metilo de MBD2 se fusiona al fragmento Fc de un anticuerpo (MBD-FC) (Gebhard *et al*. (2006) Cancer Res. 66(12):6118-28); anticuerpos específicos de metilación; métodos de conversión de bisulfito, por ejemplo, MSP (PCR sensible a la metilación), COBRA, extensión de cebadores de un solo nucleótido sensible a la metilación (Ms-SNuPE) o la tecnología Sequenom MassCLEAVE™; y el uso de enzimas de restricción sensibles a la metilación (por ejemplo, digestión del ADN materno en una muestra materna usando una o más enzimas de restricción sensibles a la metilación, enriqueciendo de ese modo el ADN fetal). Las enzimas sensibles a metilo pueden usarse además para diferenciar ácido nucleico basándose en el estado de metilación

que, por ejemplo, puede escindir o digerir preferente o sustancialmente en su secuencia de reconocimiento de ADN si esta última no está metilada. Por tanto, se corta una muestra de ADN no metilado en fragmentos más pequeños que una muestra de ADN metilado y no se escinde una muestra de ADN hipermetilado. Excepto cuando se indique explícitamente, puede usarse cualquier método para diferenciar ácido nucleico basado en el estado de metilación con las composiciones y métodos de la tecnología en el presente documento. La cantidad de ADN fetal puede determinarse, por ejemplo, introduciendo uno o más competidores a concentraciones conocidas durante una reacción de amplificación. Además, la determinación de la cantidad de ADN fetal puede realizarse, por ejemplo, mediante RT-PCR, extensión de cebadores, secuenciación y/o recuento. En determinados casos, la cantidad de ácido nucleico puede determinarse usando la tecnología BEAMing tal como se describe en la publicación de solicitud de patente estadounidense n.° 2007/0065823. En algunos casos, puede determinarse la eficiencia de restricción y el índice de eficiencia se usa para determinar además la cantidad de ADN fetal.

En algunos casos, puede usarse un ensayo cuantificador fetal (FQA) para determinar la concentración de ADN fetal en una muestra materna, por ejemplo, mediante el siguiente método: a) determinar la cantidad total de ADN presente en una muestra materna; b) digerir selectivamente el ADN materno en una muestra materna usando una o más enzimas de restricción sensibles a la metilación enriqueciendo de ese modo el ADN fetal; c) determinar la cantidad de ADN fetal de la etapa b); y d) comparar la cantidad de ADN fetal de la etapa c) con la cantidad total de ADN de la etapa a), determinándose de ese modo la concentración de ADN fetal en la muestra materna. En algunos casos, el número absoluto de copias de ácido nucleico fetal en una muestra materna puede determinarse, por ejemplo, mediante espectrometría de masas y/o un sistema que usa un enfoque competitivo de PCR para las mediciones absolutas del número de copias. Véase, por ejemplo, Ding y Cantor (2003) Proc.Natl.Acad.Sci. USA 100:3059-3064 y la publicación de solicitud de patente estadounidense n.° 2004/0081993.

En algunos casos, la fracción fetal puede determinarse basándose en las razones alélicas de secuencias polimórficas (por ejemplo, polimorfismos de un solo nucleótido (SNP)) tal como, por ejemplo, usando un método descrito en la publicación de solicitud de patente estadounidense n.° 2011/0224087. En tal método, se obtienen lecturas de secuencia de nucleótidos para una muestra materna y se determina la fracción fetal comparando el número total de lecturas de secuencia de nucleótidos que se mapean en un primer alelo y el número total de lecturas de secuencia de nucleótidos que se mapean en un segundo alelo en un sitio polimórfico informativo (por ejemplo, SNP) en un genoma de referencia. En algunos casos, los alelos fetales se identifican, por ejemplo, por su contribución minoritaria relativa a la mezcla de ácidos nucleicos fetales y maternos en la muestra cuando se compara con la contribución mayoritaria a la mezcla por los ácidos nucleicos maternos. En consecuencia, la abundancia relativa de ácido nucleico fetal en una muestra materna puede determinarse como un parámetro del número total de lecturas de secuencia únicas mapeadas en una secuencia de ácido nucleico diana en un genoma de referencia para cada uno de los dos alelos de un sitio polimórfico.

La cantidad de ácido nucleico fetal en ácido nucleico extracelular puede cuantificarse y usarse junto con un método proporcionado en el presente documento. Por tanto, en determinadas implementaciones, los métodos de la tecnología descrita en el presente documento comprenden una etapa adicional para determinar la cantidad de ácido nucleico fetal. La cantidad de ácido nucleico fetal puede determinarse en una muestra de ácido nucleico de un sujeto antes o después del procesamiento para preparar el ácido nucleico de muestra. En determinadas implementaciones, la cantidad de ácido nucleico fetal se determina en una muestra después de procesar y preparar el ácido nucleico de muestra, cantidad que se usa para una evaluación adicional. En algunas implementaciones, un resultado comprende factorizar la fracción de ácido nucleico fetal en el ácido nucleico de muestra (por ejemplo, ajustar recuentos, eliminar muestras, realizar una identificación o no realizar una identificación).

La etapa de determinación puede realizarse antes, durante, en un punto cualquiera de un método descrito en el presente documento, o después de determinados métodos (por ejemplo, detección de aneuploidía, determinación del sexo del feto) descritos en el presente documento. Por ejemplo, para lograr un método de determinación de aneuploidía o sexo del feto con una sensibilidad o especificidad dadas, puede implementarse un método de cuantificación de ácido nucleico fetal antes de, durante o después de la determinación de aneuploidía o sexo del feto para identificar aquellas muestras con más de aproximadamente el 2 %, 3 %, 4 %, 5 %, 6 %, 7 %, 8 %, 9 %, 10 %, 11 %, 12 %, 13 %, 14 %,15 %,16 %, 17 %, 18 %, 19 %, 20 %, 21 %, 22 %, 23 %, 24 %, 25 % o más de ácido nucleico fetal. En algunas implementaciones, las muestras que se ha determinado que tienen una determinada cantidad umbral de ácido nucleico fetal (por ejemplo, aproximadamente el 15 % o más de ácido nucleico fetal; aproximadamente el 4 % o más de ácido nucleico fetal) se analizan además para determinar aneuploidía o el sexo del feto, o la presencia o ausencia de aneuploidía o variación genética, por ejemplo. En determinadas implementaciones, las determinaciones de, por ejemplo, sexo del feto o la presencia o ausencia de aneuploidía se seleccionan (por ejemplo, se seleccionan y comunican a un paciente) solo para muestras que tienen una determinada cantidad umbral de ácido nucleico fetal (por ejemplo, aproximadamente el 15 % o más de ácido nucleico fetal; aproximadamente el 4 % o más de ácido nucleico fetal).

En algunas implementaciones, la determinación de fracción fetal o la determinación de la cantidad de ácido nucleico fetal no es necesaria o se requiere para identificar la presencia o ausencia de una aneuploidía cromosómica. En algunas implementaciones, la identificación de la presencia o ausencia de una aneuploidía cromosómica no requiere la diferenciación de secuencias de ADN fetal en comparación con el ADN materno. En algunos casos esto se debe a que se analiza la contribución sumada de las secuencias materna y fetal en un cromosoma particular, porción cromosómica o

segmento de los mismos. En algunas implementaciones, la identificación de la presencia o ausencia de una aneuploidía cromosómica no depende de una información de secuencia a priori que distinguiría el ADN fetal del ADN materno.

Enriquecimiento de una subpoblación de ácido nucleico

En algunas implementaciones, el ácido nucleico (por ejemplo, ácido nucleico extracelular) está enriquecido o relativamente enriquecido para una subpoblación o especie de ácido nucleico. Las subpoblaciones de ácidos nucleicos pueden incluir, por ejemplo, ácido nucleico fetal, ácido nucleico materno, ácido nucleico que comprende fragmentos de una longitud o rango de longitudes particular, o ácido nucleico de una región particular del genoma (por ejemplo, cromosoma único, conjunto de cromosomas, y/o determinadas regiones cromosómicas). Tales muestras enriquecidas pueden usarse junto con un método proporcionado en el presente documento. Por tanto, en determinadas implementaciones, los métodos de la tecnología comprenden una etapa adicional de enriquecimiento en una subpoblación de ácido nucleico en una muestra, tal como, por ejemplo, ácido nucleico fetal. En algunos casos, un método para determinar la fracción fetal descrita anteriormente puede usarse además para enriquecer el ácido nucleico fetal. En determinadas implementaciones, el ácido nucleico materno se elimina selectivamente (parcial, sustancialmente, casi completamente o completamente) de la muestra. En algunos casos, el enriquecimiento de un ácido nucleico de especie de bajo número de copias particular (por ejemplo, ácido nucleico fetal) puede mejorar la sensibilidad cuantitativa. Los métodos para enriquecer una muestra para una especie particular de ácido nucleico se describen, por ejemplo, en la patente estadounidense n.° 6.927.028, la publicación de solicitud de patente internacional n.° WO2007/140417, la publicación de solicitud de patente internacional n.° WO2007/147063, la publicación de solicitud de patente internacional n.° WO2009/032779, la publicación de solicitud de patente internacional n.° WO2009/032781, la publicación de solicitud de patente internacional n.° WO2010/033639, la publicación de solicitud de patente internacional n.° WO2011/034631, la publicación de solicitud de patente internacional n.° WO2006/056480 y la publicación de solicitud de patente internacional n.° WO2011/143659.

En algunas implementaciones, el ácido nucleico se enriquece en determinadas especies de fragmento diana y/o especies de fragmento de referencia. En algunos casos, el ácido nucleico se enriquece en una longitud específica de fragmento de ácido nucleico o rango de longitudes de fragmento con el uso de uno o más métodos de separación basados en la longitud descritos a continuación. En algunos casos, el ácido nucleico se enriquece en fragmentos de una región genómica seleccionada (por ejemplo, cromosoma) con el uso de uno o más métodos de separación basados en secuencia descritos en el presente documento y/o conocidos en la técnica. Determinados métodos para enriquecer una subpoblación de ácido nucleico (por ejemplo, ácido nucleico fetal) en una muestra se describen con detalle más adelante.

Algunos métodos para enriquecer una subpoblación de ácido nucleico (por ejemplo, ácido nucleico fetal) que pueden usarse con un método descrito en el presente documento incluyen métodos que aprovechan diferencias epigenéticas entre ácido nucleico materno y fetal. Por ejemplo, el ácido nucleico fetal puede diferenciarse y separarse del ácido nucleico materno basándose en diferencias de metilación. Se describen métodos de enriquecimiento de ácido nucleico fetal basados en metilación en la publicación de solicitud de patente estadounidense n.° 2010/0105049. Esos métodos implican, algunas veces, unir un ácido nucleico de muestra a un agente de unión específico de metilación (proteína de unión a metil-CpG (MBD), anticuerpos específicos de metilación y similares) y separar el ácido nucleico unido del ácido nucleico no unido basándose en el estado de metilación diferencial. Tales métodos pueden incluir además el uso de enzimas de restricción sensibles a la metilación (tal como se describió anteriormente; por ejemplo, Hhal y Hpall), que permiten el enriquecimiento de regiones de ácido nucleico fetal en una muestra materna mediante la digestión selectiva de ácido nucleico de la muestra materna con una enzima que digiere selectiva y completa o sustancialmente el ácido nucleico materno para enriquecer la muestra en al menos una región de ácido nucleico fetal.

Otro método para enriquecer una subpoblación de ácido nucleico (por ejemplo, ácido nucleico fetal) que puede usarse con un método descrito en el presente documento es un enfoque de secuencia polimórfica potenciada por endonucleasas de restricción, tal como un método descrito en la publicación de solicitud de patente estadounidense n.° 2009/0317818. Tales métodos incluyen la escisión de ácido nucleico que comprende un alelo no diana con una endonucleasa de restricción que reconoce el ácido nucleico que comprende el alelo no diana pero no el alelo diana; y amplificación de ácido nucleico no escindido pero no de ácido nucleico escindido, en el que el ácido nucleico amplificado no escindido representa ácido nucleico diana enriquecido (por ejemplo, ácido nucleico fetal) con respecto a ácido nucleico no diana (por ejemplo, ácido nucleico materno). En algunos casos, el ácido nucleico puede seleccionarse de tal manera que comprenda un alelo que tiene un sitio polimórfico susceptible a la digestión selectiva por medio de un agente de escisión, por ejemplo.

Algunos métodos para enriquecer una subpoblación de ácido nucleico (por ejemplo, ácido nucleico fetal) que pueden usarse con un método descrito en el presente documento incluyen enfoques de degradación enzimática selectiva. Tales métodos implican proteger secuencias diana frente a la digestión con exonucleasas, facilitando de ese modo la eliminación en una muestra de secuencias no deseadas (por ejemplo, ADN materno). Por ejemplo, en un enfoque, el ácido nucleico de muestra se desnaturaliza para generar ácido nucleico monocatenario, el ácido nucleico monocatenario se pone en contacto con al menos un par de cebadores específicos de diana en condiciones de hibridación adecuadas, los cebadores hibridados se extienden mediante polimerización de nucleótidos que genera secuencias diana bicatenarias, y que digiere el ácido nucleico monocatenario usando una nucleasa que digiere ácido nucleico monocatenario (es decir, no diana). En algunos casos, el método puede repetirse durante al menos un ciclo adicional. En algunos casos, se usa el mismo par de cebadores

específicos de diana para cebar cada uno del primer y el segundo ciclos de extensión y, en algunos casos, se usan pares de cebadores específicos de diana diferentes para el primer y el segundo ciclos.

Algunos métodos para enriquecer una subpoblación de ácido nucleico (por ejemplo, ácido nucleico fetal) que puede usarse con un método descrito en el presente documento incluyen métodos de secuenciación masiva en paralelo de firmas (MPSS). Normalmente, MPSS es un método de fase sólida que usa ligamiento de adaptador (es decir, etiqueta), seguida de decodificación de adaptador, y lectura de la secuencia de ácido nucleico en pequeños incrementos. Los productos de PCR marcados se amplifican normalmente de tal manera que cada ácido nucleico genera un producto de PCR con una etiqueta única. Las etiquetas se usan a menudo para unir los productos de PCR a microperlas. Después de varias tandas de determinación de secuencias basada en ligamiento, por ejemplo, puede identificarse una firma de secuencia a partir de cada perla. Se analiza cada secuencia de firma (etiqueta de MPSS) en un conjunto de datos MPSS, en comparación con todas las demás firmas, y se cuentan todas las firmas idénticas.

En algunos casos, determinados métodos de enriquecimiento basados en MPSS pueden incluir enfoques basados en amplificación (por ejemplo, PCR). En algunos casos, pueden usarse métodos de amplificación específicos de locus (por ejemplo, usando cebadores de amplificación específicos de locus). En algunos casos, puede usarse un enfoque de PCR de alelos de SNP múltiplex. En algunos casos, puede usarse un enfoque de PCR de alelos de SNP múltiplex en combinación con secuenciación uniplex. Por ejemplo, tal método puede incluir el uso de PCR múltiplex (por ejemplo, el sistema MASSARRAY) y la incorporación de secuencias de sondas de captura en los amplicones seguido por secuenciación usando, por ejemplo, el sistema Illumina de MPSS. En algunos casos, puede usarse un método de PCR de alelos de SNP múltiplex en combinación con un sistema de tres cebadores y secuenciación indexada. Por ejemplo, tal método puede involucrar el uso de PCR múltiple (por ejemplo, sistema MASSARRAY) con cebadores que tienen una primera sonda de captura incorporada en determinados cebadores de PCR directos específicos de locus y secuencias adaptadoras incorporadas en los cebadores de PCR inversos específicos de locus, para generar de ese modo amplicones, seguido por una PCR secundaria para incorporar las secuencias de captura inversa y los códigos de barras de índice molecular para la secuenciación usando, por ejemplo, el sistema Illumina de MPSS. En algunos casos, puede usarse un método de PCR de alelos de SNP múltiplex en combinación con un sistema de cuatro cebadores y secuenciación indexada. Por ejemplo, tal enfoque que puede involucrar el uso de PCR múltiplex (por ejemplo, sistema MASSARRAY) con cebadores que tienen secuencias adaptadoras incorporadas en los cebadores de PCR directos específicos de locus e inversos específicos de locus, seguido de una PCR secundaria para incorporar las secuencias de captura directa e inversa y los códigos de barra de índice molecular para la secuenciación usando, por ejemplo, el sistema Illumina de MPSS. En algunos casos, puede usarse un enfoque microfluídico. En algunos casos, puede usarse un enfoque microfluídico basado en alineamientos. Por ejemplo, tal método puede incluir el uso de un alineamiento microfluídico (por ejemplo, Fluidigm) para la amplificación a una multiplicidad baja y la incorporación de sondas de índice y captura, seguido de secuenciación. En algunos casos, puede usarse un método microfluídico en emulsión, tal como, por ejemplo, PCR digital en gotas.

En algunos casos, pueden usarse métodos de amplificación universales (por ejemplo, usando cebadores de amplificación universales o inespecíficos de locus). En algunos casos, los métodos de amplificación universales pueden usarse en combinación con enfoques de detección de interacciones (“pull-down”). En algunos casos, un método puede incluir la detección de interacciones de Ultramer biotinilado (por ejemplo, ensayos de detección de interacciones biotiniladas de Agilent o IDT) a partir de una biblioteca de secuenciación amplificada de manera universal. Por ejemplo, tal método puede implicar la preparación de una biblioteca convencional, el enriquecimiento de regiones seleccionadas mediante un ensayo de detección de interacciones y una etapa secundaria de amplificación universal. En algunos casos, los enfoques de detección de interacciones pueden usarse en combinación con métodos basados en ligamiento. En algunos casos, un método puede incluir la detección de interacciones de Ultramer biotinilado con ligamiento de adaptador específico de secuencia (por ejemplo, PCR HALOPLEX, Halo Genomics). Por ejemplo, tal método puede incluir el uso de sondas selectoras para capturar fragmentos digeridos por enzimas de restricción, seguido del ligamiento de productos capturados a un adaptador, y la amplificación universal seguida de secuenciación. En algunos casos, los enfoques de detección de interacciones pueden usarse en combinación con métodos basados en extensión y ligamiento. En algunos casos, un método puede incluir la extensión y ligamiento de la sonda de inversión molecular (MIP). Por ejemplo, tal enfoque puede incluir el uso de sondas de inversión molecular en combinación con adaptadores de secuencia seguido de amplificación y secuenciación universal. En algunos casos, el ADN complementario puede sintetizarse y secuenciarse sin amplificación.

En algunos casos, los enfoques de extensión y ligamiento pueden realizarse sin un componente de detección de interacciones. En algunos casos, un método puede incluir hibridación de cebadores directos e inversos específicos de locus, extensión y ligamiento. Tales métodos pueden incluir además amplificación universal o síntesis de ADN complementario sin amplificación, seguida de secuenciación. Tales métodos pueden reducir o excluir secuencias de fondo durante el análisis, en algunos casos.

En algunos casos, los enfoques de detección de interacciones pueden usarse con un componente de amplificación opcional o sin componente de amplificación. En algunos casos, un método puede incluir un ensayo de detección de interacciones modificado y ligamiento con incorporación completa de sondas de captura sin amplificación universal. Por ejemplo, tal método puede incluir el uso de sondas selectoras modificadas para capturar fragmentos digeridos por enzimas de restricción, seguido por el ligamiento de productos capturados a un adaptador, amplificación opcional y secuenciación. En algunos casos, un método puede incluir un ensayo de detección de interacciones biotiniladas con extensión y ligamiento de

la secuencia adaptadora en combinación con ligamiento monocatenario circular. Por ejemplo, tal enfoque puede incluir el uso de sondas selectoras para capturar regiones de interés (es decir, secuencias diana), extensión de las sondas, ligamiento de adaptador, ligamiento monocatenario circular, amplificación opcional y secuenciación. En algunos casos, el análisis del resultado de la secuenciación puede separar las secuencias diana forma fondo.

En algunas implementaciones, el ácido nucleico se enriquece en fragmentos de una región genómica seleccionada (por ejemplo, cromosoma) por medio del uso de uno o más métodos de separación basados en secuencia descritos en el presente documento. La separación basada en secuencia se basa generalmente en las secuencias de nucleótidos presentes en los fragmentos de interés (por ejemplo, fragmentos diana y/o de referencia) y sustancialmente no presentes en otros fragmentos de la muestra o presentes en una cantidad insustancial de los otros fragmentos (por ejemplo, el 5 % o menos). En algunas implementaciones, la separación basada en secuencia puede generar fragmentos diana separados y/o fragmentos de referencia separados. Los fragmentos diana separados y/o fragmentos de referencia separados, normalmente se aíslan de los fragmentos restantes en la muestra de ácido nucleico. En algunos casos, los fragmentos diana separados y los fragmentos de referencia separados se aíslan además entre sí (por ejemplo, se aíslan en compartimentos de ensayo separados). En algunos casos, los fragmentos diana separados y los fragmentos de referencia separados se aíslan juntos (por ejemplo, se aíslan en el mismo compartimento de ensayo). En algunas implementaciones, los fragmentos no unidos pueden retirarse diferencialmente, degradarse o digerirse.

En algunas implementaciones, se usa un procedimiento de captura de ácido nucleico selectivo para separar los fragmentos diana y/o de referencia de la muestra de ácido nucleico. Los sistemas de captura de ácido nucleico disponibles comercialmente incluyen, por ejemplo, el sistema de captura de secuencia Nimblegen (Roche NimbleGen, Madison, WI); la plataforma BEADARRAY de Illumina (Illumina, San Diego, CA); la plataforma Affymetrix GENECHIP (Affymetrix, Santa Clara, CA); el sistema de enriquecimiento diana Agilent SureSelect (Agilent Technologies, Santa Clara, CA); y plataformas relacionadas. Tales métodos implican normalmente la hibridación de un oligonucleótido de captura a un segmento o toda la secuencia de nucleótidos de un fragmento diana o de referencia y pueden incluir el uso de una fase sólida (por ejemplo, alineamiento de fase sólida) y/o una plataforma a base de disolución. Los oligonucleótidos de captura (a veces denominados “cebo”) pueden seleccionarse o diseñarse de tal manera que se hibriden preferiblemente con fragmentos de ácido nucleico de regiones o loci genómicos seleccionados (por ejemplo, uno de los cromosomas 21, 18, 13, X o Y, o un cromosoma de referencia).

En algunas implementaciones, el ácido nucleico se enriquece en una longitud particular de fragmento de ácido nucleico, rango de longitudes o longitudes por debajo o por encima de un umbral o punto de corte particular usando uno o más métodos de separación basados en la longitud. La longitud del fragmento de ácido nucleico se refiere normalmente al número de nucleótidos en el fragmento. La longitud del fragmento de ácido nucleico se denomina además algunas veces tamaño del fragmento de ácido nucleico. En algunas implementaciones, se realiza un método de separación basado en la longitud sin medir las longitudes de los fragmentos individuales. En algunas implementaciones, se realiza un método de separación basado en la longitud junto con un método para determinar la longitud de los fragmentos individuales. En algunas implementaciones, la separación basada en longitud se refiere a un procedimiento de fraccionamiento por tamaño, en el que la totalidad o parte de la combinación fraccionada puede aislarse (por ejemplo, retenerse) y/o analizarse. Los procedimientos de fraccionamiento por tamaños se conocen en la técnica (por ejemplo, separación en un alineamiento, separación mediante un tamiz molecular, separación por electroforesis en gel, separación por cromatografía en columna (por ejemplo, columnas de exclusión molecular) y enfoques basados en microfluídica). En algunos casos, los métodos de separación basados en la longitud pueden incluir circularización de fragmentos, tratamiento químico (por ejemplo, con formaldehído, polietilenglicol (PEG)), espectrometría de masas y/o amplificación de ácido nucleico específica de tamaño, por ejemplo.

Determinados métodos de separación basados en la longitud que pueden usarse con los métodos descritos en el presente documento usan un enfoque de marcaje con etiqueta de secuencia selectivo, por ejemplo. La expresión “marcaje con etiqueta de secuencia” se refiere a incorporar una secuencia reconocible y distinta en un ácido nucleico o una población de ácidos nucleicos. La expresión “marcaje con etiqueta de secuencia”, tal como se usa en el presente documento, tiene un significado diferente a la expresión “etiqueta de secuencia” que se describe más adelante en el presente documento. En tales métodos de marcaje con etiqueta de secuencia, ácidos nucleicos de una especie de tamaño de fragmento (por ejemplo, fragmentos cortos) se someten a marcaje con etiqueta selectivo de secuencia en una muestra que incluye ácidos nucleicos largos y cortos. Tales métodos implican normalmente llevar a cabo una reacción de amplificación de ácido nucleico usando un conjunto de cebadores anidados que incluyen cebadores internos y cebadores externos. En algunos casos, uno o ambos interiores pueden marcarse con etiqueta para introducir de ese modo una etiqueta sobre el producto de amplificación diana. Generalmente, los cebadores externos no se hibridan con los fragmentos cortos que portan la secuencia diana (interna). Los cebadores internos pueden hibridarse con los fragmentos cortos y generar un producto de amplificación que porta una etiqueta y la secuencia diana. Normalmente, el marcaje con etiqueta de los fragmentos largos se inhibe a través de una combinación de mecanismos que incluyen, por ejemplo, extensión bloqueada de los cebadores internos por la hibridación y extensión previas de los cebadores externos. El enriquecimiento en fragmentos marcados puede lograrse mediante cualquiera de una gran variedad de métodos incluyendo, por ejemplo, digestión con exonucleasa de ácido nucleico monocatenario y amplificación de los fragmentos marcados con etiqueta usando cebadores de amplificación específicos para al menos una etiqueta.

Otro método de separación basado en la longitud que puede usarse con los métodos descritos en el presente documento implica someter una muestra de ácido nucleico a precipitación con polietilenglicol (PEG). Los ejemplos de métodos incluyen los descritos en las publicaciones de solicitud de patente internacional n.^{os} WO2007/140417 y WO2010/115016. Este método implica generalmente poner en contacto una muestra de ácido nucleico con PEG en presencia de una o más sales monovalentes en condiciones suficientes para precipitar sustancialmente ácidos nucleicos grandes sin precipitar sustancialmente ácidos nucleicos pequeños (por ejemplo, de menos de 300 nucleótidos).

Otro método de enriquecimiento basado en tamaño que puede usarse con los métodos descritos en el presente documento implica la circularización mediante ligamiento, por ejemplo, con el uso de CircLigase. Los fragmentos de ácido nucleico cortos pueden circularizarse normalmente con una mayor eficiencia que los fragmentos largos. Las secuencias no circularizadas pueden separarse de las secuencias circularizadas, y los fragmentos cortos enriquecidos pueden usarse para un análisis adicional.

Obtención de lecturas de secuencia

En algunas implementaciones, pueden secuenciarse ácidos nucleicos (por ejemplo, fragmentos de ácido nucleico, ácido nucleico de muestra, ácido nucleico libre de células). En algunos casos, se obtiene una secuencia completa o sustancialmente completa y, algunas veces, se obtiene una secuencia parcial. Los métodos de secuenciación, mapeo y analíticos relacionados se conocen en la técnica (por ejemplo, la publicación de solicitud de patente estadounidense US-2009/0029377). Determinados aspectos de tales procedimientos se describen más adelante en el presente documento.

Tal como se usa en el presente documento, "lecturas" (es decir, "una lectura", "una lectura de secuencia") son secuencias de nucleótidos cortas producidas mediante cualquier procedimiento de secuenciación descrito en el presente documento o conocido en la técnica. Las lecturas pueden generarse a partir de un extremo de los fragmentos de ácido nucleico ("lecturas de un solo extremo") y, algunas veces, se generan a partir de ambos extremos de los ácidos nucleicos (por ejemplo, lecturas de extremos apareados, lecturas de doble extremo).

En algunas implementaciones, la longitud nominal, promedio, media o absoluta de lecturas de un solo extremo es, algunas veces, de aproximadamente 20 nucleótidos contiguos a aproximadamente 50 nucleótidos contiguos, algunas veces de aproximadamente 30 nucleótidos contiguos a aproximadamente 40 nucleótidos contiguos y, algunas veces, de aproximadamente 35 nucleótidos contiguos o de aproximadamente 36 nucleótidos contiguos. Algunas veces, la longitud nominal, promedio, media o absoluta de lecturas de un solo extremo es de aproximadamente 20 a aproximadamente 30 bases de longitud. Algunas veces, la longitud nominal, promedio, media o absoluta de lecturas de un solo extremo es de aproximadamente 24 a aproximadamente 28 bases de longitud. Algunas veces la longitud nominal, promedio, media o absoluta de lecturas de un solo extremo es de aproximadamente 21, 22, 23, 24, 25, 26, 27, 28 o aproximadamente 29 bases de longitud.

En determinadas implementaciones, la longitud nominal, promedio, media o absoluta de las lecturas de extremos apareados es, algunas veces, de aproximadamente 10 nucleótidos contiguos a aproximadamente 25 nucleótidos contiguos (por ejemplo, aproximadamente 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23 o 24 nucleótidos de longitud), algunas veces tiene de aproximadamente 15 nucleótidos contiguos a aproximadamente 20 nucleótidos contiguos y, algunas veces, tiene aproximadamente 17 nucleótidos contiguos o aproximadamente 18 nucleótidos contiguos.

Las lecturas son generalmente representaciones de secuencias de nucleótidos en un ácido nucleico físico. Por ejemplo, en una lectura que contiene una representación ATGC de una secuencia, "A" representa un nucleótido adenina, "T" representa un nucleótido timina, "G" representa un nucleótido guanina y "C" representa un nucleótido citosina, en un ácido nucleico físico. Las lecturas de secuencia obtenidas a partir de la sangre de una mujer embarazada pueden ser lecturas de una mezcla de ácido nucleico fetal y materno. Una mezcla de lecturas relativamente cortas puede transformarse mediante procedimientos descritos en el presente documento en una representación de un ácido nucleico genómico presente en la mujer embarazada y/o en el feto. Una mezcla de lecturas relativamente cortas puede transformarse en una representación de una variación del número de copias (por ejemplo, una variación del número de copias materno y/o fetal), una variación genética o una aneuploidía, por ejemplo. Las lecturas de una mezcla de ácido nucleico materno y fetal pueden transformarse en una representación de un cromosoma compuesto o un segmento del mismo que comprende características de uno o ambos cromosomas maternos y fetales. En determinadas implementaciones, "obtener" lecturas de secuencia de ácido nucleico de una muestra de un sujeto y/u "obtener" lecturas de secuencia de ácido nucleico de un espécimen biológico de una o más personas de referencia pueden implicar la secuenciación directa del ácido nucleico para obtener la información de secuencia. En algunas implementaciones, "obtener" puede implicar recibir información de secuencia obtenida directamente de un ácido nucleico por otro.

Las lecturas de secuencia pueden mapearse y la cantidad de lecturas o etiquetas de secuencia que se mapean en una región específica de ácido nucleico (por ejemplo, un cromosoma, un bin, una sección genómica) se denominan recuentos. En algunas implementaciones, los recuentos pueden manipularse o transformarse (por ejemplo, normalizarse, combinarse, sumarse, filtrarse, seleccionarse, promediarse, derivarse como una media, o similar o una combinación de los mismos). En algunas implementaciones, los recuentos pueden transformarse para producir recuentos normalizados. Los recuentos normalizados para múltiples secciones genómicas pueden proporcionarse en un perfil (por ejemplo, un perfil genómico, un perfil cromosómico, un perfil de un segmento o una porción de un

cromosoma). También pueden manipularse o transformarse una o más elevaciones diferentes en un perfil (por ejemplo, pueden normalizarse los recuentos asociados con las elevaciones) y pueden ajustarse las elevaciones.

En algunas implementaciones, se secuencia una muestra de ácido nucleico de un individuo. En determinadas implementaciones, las muestras de ácido nucleico de dos o más muestras biológicas, en las que cada muestra biológica es de un individuo o dos o más individuos, se combinan y se secuencia la combinación. En las últimas implementaciones, una muestra de ácido nucleico de cada muestra biológica se identifica a menudo por medio de una o más etiquetas de identificación únicas.

En algunas implementaciones, una fracción del genoma se secuencia, la cual a veces se expresa en la cantidad del genoma cubierta por las secuencias de nucleótidos determinadas (por ejemplo, "veces" de cobertura menores de 1). Cuando se secuencia un genoma con una cobertura de aproximadamente 1 vez, aproximadamente el 100 % de la secuencia de nucleótidos del genoma está representado por lecturas. Un genoma también puede secuenciarse con redundancia, en el que una región dada del genoma puede cubrirse por dos o más lecturas o lecturas solapantes (por ejemplo, "veces" de cobertura mayores de 1). En algunas implementaciones, se secuencia un genoma con una cobertura de aproximadamente 0,1 veces a aproximadamente 100 veces, una cobertura de aproximadamente 0,2 veces a 20 veces, o una cobertura de aproximadamente 0,2 veces a aproximadamente 1 vez (por ejemplo, una cobertura de aproximadamente 0,2, 0,3, 0,4, 0,5, 0,6, 0,7, 0,8, 0,9, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90 veces).

En determinadas implementaciones, una fracción de una combinación de ácidos nucleicos que se secuencia en una ejecución se subselecciona adicionalmente antes de la secuenciación. En determinadas implementaciones, pueden usarse técnicas basadas en hibridación (por ejemplo, usando alineamientos de oligonucleótidos) para subseleccionar en primer lugar las secuencias de ácido nucleico de determinados cromosomas (por ejemplo, un cromosoma potencialmente aneuploide y otro(s) cromosoma(s) no involucrado(s) en la aneuploidía sometida a prueba). En algunas implementaciones, el ácido nucleico puede fraccionarse por tamaños (por ejemplo, mediante electroforesis en gel, cromatografía de exclusión molecular o mediante un enfoque basado en microfluídica) y, en determinados casos, el ácido nucleico fetal puede enriquecerse mediante la selección de ácido nucleico que tiene un menor peso molecular (por ejemplo, menos de 300 pares de bases, menos de 200 pares de bases, menos de 150 pares de bases, menos de 100 pares de bases). En algunas implementaciones, el ácido nucleico fetal puede enriquecerse suprimiendo el ácido nucleico de fondo materno, tal como mediante la adición de formaldehído. En algunas implementaciones, una porción o un subconjunto de una combinación preseleccionada de ácidos nucleicos se secuencia aleatoriamente. En algunas implementaciones, el ácido nucleico se amplifica antes de la secuenciación. En algunas implementaciones, una porción o un subconjunto del ácido nucleico se amplifica antes de la secuenciación.

En algunos casos, se prepara una biblioteca de secuenciación antes de o durante un procedimiento de secuenciación. Los métodos para preparar una biblioteca de secuenciación se conocen en la técnica y pueden usarse plataformas disponibles comercialmente para determinadas aplicaciones. Determinadas plataformas de bibliotecas disponibles comercialmente pueden ser compatibles con determinados procedimientos de secuenciación de nucleótidos descritos en el presente documento. Por ejemplo, una o más plataformas de bibliotecas disponibles comercialmente pueden ser compatibles con una secuenciación mediante un procedimiento de síntesis. En algunos casos, se usa un método de preparación de bibliotecas basado en ligamiento (por ejemplo, ILLUMINA TRUSEQ, Illumina, San Diego CA). Los métodos de preparación de bibliotecas basados en ligamiento usan normalmente un diseño de adaptador metilado que puede incorporar una secuencia índice en la etapa de ligamiento inicial y a menudo pueden usarse para preparar muestras para secuenciación de una sola lectura, secuenciación de extremos apareados y secuenciación multiplexada. En algunos casos, se usa un método de preparación de biblioteca basado en transposones (por ejemplo, EPICENTRE NEXTERA, Epicentre, Madison WI). Los métodos basados en transposones usan normalmente la transposición *in vitro* para fragmentar y marcar con etiqueta simultáneamente ADN en una reacción de un solo tubo (que a menudo permite la incorporación de etiquetas específicas de plataforma y códigos de barras opcionales), y preparar bibliotecas listas para el secuenciador.

Puede usarse cualquier método de secuenciación adecuado para realizar los métodos descritos en el presente documento. En algunas implementaciones, se usa un método de secuenciación de alto rendimiento. Los métodos de secuenciación de alto rendimiento involucran generalmente moldes de ADN amplificados de manera clonal o moléculas de ADN individuales que se secuencian de manera masiva en paralelo dentro de una celda de flujo (por ejemplo, tal como se describe en Metzker M Nature Rev 11:31-46 (2010); Volkerding *et al.* Clin.Chern. 55:641-658 [2009]). Tales métodos de secuenciación pueden proporcionar además información cuantitativa digital, en los que cada lectura de secuencia es una "etiqueta de secuencia" o "recuento" contable que representa un molde de ADN clonal individual, una molécula de ADN individual, un bin o cromosoma. Las técnicas de secuenciación de nueva generación capaces de secuenciar ADN de manera masiva en paralelo se denominan colectivamente en el presente documento "secuenciación masiva en paralelo" (MPS). Las tecnologías de secuenciación de alto rendimiento incluyen, por ejemplo, secuenciación por síntesis con terminadores con colorante reversibles, secuenciación por ligamiento con sonda de oligonucleótidos, pirosecuenciación y secuenciación en tiempo real. Los ejemplos no limitativos de MPS incluyen secuenciación masiva en paralelo de firmas (MPSS), secuenciación de Polony, pirosecuenciación, secuenciación con Illumina (Solexa), secuenciación SOLiD, secuenciación con semiconductores iónicos, secuenciación de nanobolas de ADN, secuenciación de una sola molécula con Helioscope, secuenciación en tiempo real de una sola molécula (SMRT), secuenciación por nanoporos, secuenciación de lON-Torrent y ARN polimerasa (ARNP).

Los sistemas usados para métodos de secuenciación de alto rendimiento están disponibles comercialmente e incluyen, por ejemplo, la plataforma Roche 454, la plataforma SOLID de Applied Biosystems, la tecnología de secuenciación de ADN de una sola molécula verdadera Helicos, la plataforma de secuenciación por hibridación de Affymetrix Inc., la tecnología de un sola molécula, en tiempo real (SMRT) de Pacific Biosciences, las plataformas de secuenciación por síntesis de 454 Life Sciences, Illumina/Solexa y Helicos Biosciences, y la plataforma de secuenciación por ligamiento de Applied Biosystems. La tecnología ION-TORRENT de Life technologies y secuenciación por nanoporos pueden usarse además en métodos de secuenciación de alto rendimiento.

En algunas implementaciones, la tecnología de primera generación, tal como, por ejemplo, secuenciación de Sanger que incluye la secuenciación automatizada de Sanger, puede usarse en un método proporcionado en el presente documento. Además, en el presente documento se contemplan tecnologías de secuenciación adicionales que incluyen el uso de tecnologías de obtención de imágenes de ácidos nucleicos en desarrollo (por ejemplo, microscopia electrónica de transmisión (TEM) y microscopia de fuerza atómica (AFM)). A continuación se describen ejemplos de diversas tecnologías de secuenciación.

Una tecnología de secuenciación de ácidos nucleicos que puede usarse en un método descrito en el presente documento es secuenciación por síntesis y secuenciación basada en el terminador reversible (por ejemplo, analizador genómico de Illumina; Genome Analyzer II; HISEC. 2000; HISEC 2500 (Illumina, San Diego CA)). Con esta tecnología, millones de fragmentos de ácido nucleico (por ejemplo, ADN) pueden secuenciarse en paralelo. En un ejemplo de este tipo de tecnología de secuenciación, se usa una celda de flujo que contiene un portaobjetos ópticamente transparente con 8 carriles individuales sobre cuyas superficies hay anclajes de oligonucleótidos unidos (por ejemplo, cebadores adaptadores). Una celda de flujo es a menudo un soporte sólido que puede configurarse para retener y/o permitir el paso ordenado de disoluciones de reactivos sobre analitos unidos. Las celdas de flujo tienen a menudo forma plana, son ópticamente transparentes, generalmente en la escala milimétrica o submilimétrica, y tienen a menudo canales o carriles en los que se produce la interacción analito/reactivo.

En determinados procedimientos de secuenciación por síntesis, por ejemplo, algunas veces puede fragmentarse ADN molde (por ejemplo, ADN circulante, libre de células (ADNclc)) en longitudes de varios cientos de pares de bases en la preparación para la generación de bibliotecas. En algunas implementaciones, la preparación de bibliotecas puede realizarse sin fragmentación adicional o selección de tamaño del ADN de molde (por ejemplo, ADNclc). El aislamiento de la muestra y la generación de la biblioteca pueden realizarse usando aparatos y métodos automatizados, en determinadas implementaciones. Brevemente, el ADN molde se repara en los extremos mediante una reacción de relleno, reacción de exonucleasa o una combinación de una reacción de relleno y reacción de exonucleasa. El ADN molde reparado de extremos romos resultante se extiende por un solo nucleótido, que es complementario a un solo nucleótido en proyección en el extremo 3’ de un cebador adaptador y a menudo aumenta la eficiencia de ligamiento. Puede usarse cualquier nucleótido complementario para los nucleótidos de proyección/extensión (por ejemplo, A/T, C/G); sin embargo, la adenina se usa a menudo para extender el ADN reparado en los extremos, y la timina se usa a menudo como el nucleótido de proyección en el extremo 3'.

En determinados procedimientos de secuenciación por síntesis, por ejemplo, los oligonucleótidos adaptadores son complementarios a los anclajes de celdas de flujo, y algunas veces se usan para asociar el ADN molde modificado (por ejemplo, reparado en los extremos y extendido con un solo nucleótido) con un soporte sólido, tal como la superficie interior de una celda de flujo, por ejemplo. En algunas implementaciones, el adaptador incluye además identificadores (es decir, nucleótidos indexadores, o nucleótidos de “código de barras” (por ejemplo, una secuencia única de nucleótidos que puede usarse como identificador para permitir la identificación inequívoca de una muestra y/o un cromosoma)), uno o más sitios de hibridación de cebadores de secuenciación (por ejemplo, secuencias complementarias a los cebadores de secuenciación universales, cebadores de secuenciación de un solo extremo, cebadores de secuenciación de extremos apareados, cebadores de secuenciación multiplexados, y similares), o combinaciones de los mismos (por ejemplo, adaptador/secuenciación, adaptador/identificador, adaptador/identificador/secuenciación). Los identificadores o nucleótidos contenidos en un adaptador tienen a menudo seis o más nucleótidos de longitud y frecuentemente se sitúan en el adaptador de tal manera que los nucleótidos identificadores son los primeros nucleótidos secuenciados durante la reacción de secuenciación. En determinadas implementaciones, los nucleótidos identificadores se asocian con una muestra pero se secuencian en una reacción de secuenciación independiente para evitar comprometer la calidad de las lecturas de secuencia. Posteriormente, las lecturas de la secuenciación identificadora y la secuenciación del molde de ADN se ligan entre sí y las lecturas se desmultiplexan. Después del ligamiento y la desmultiplexación, las lecturas y/o los identificadores de secuencia pueden ajustarse o procesarse adicionalmente tal como se describe en el presente documento.

En determinados procedimientos de secuenciación por síntesis, el uso de identificadores permite la multiplexación de reacciones de secuencia en un carril de celda de flujo, lo que permite el análisis de múltiples muestras por carril de celda de flujo. El número de muestras que pueden analizarse en un carril de celda de flujo dado depende a menudo del número de identificadores únicos usados durante la preparación de la biblioteca y/o el diseño de sonda. Los ejemplos no limitativos de kits de secuenciación múltiplex disponibles comercialmente incluyen el kit de oligonucleótidos para la preparación de muestras de multiplexación de Illumina y los cebadores de secuenciación de multiplexación y el kit de control PhiX (por ejemplo, números de catálogo de Illumina PE-400-1001 y PE-400-1002, respectivamente). Un método descrito en el presente documento puede realizarse usando cualquier número de identificadores únicos (por ejemplo, 4,

8, 12, 24, 48, 96 o más). Cuanto mayor sea el número de identificadores únicos, mayor será el número de muestras y/o cromosomas, por ejemplo, que pueden multiplexarse en un solo carril de celda de flujo.

La multiplexación usando 12 identificadores, por ejemplo, permite el análisis simultáneo de 96 muestras (por ejemplo, igual al número de pocillos en una placa de micropocillos de 96 pocillos) en una celda de flujo de 8 carriles. De manera similar, la multiplexación usando 48 identificadores, por ejemplo, permite el análisis simultáneo de 384 muestras (por ejemplo, igual al número de pocillos en una placa de micropocillos de 384 pocillos) en una celda de flujo de 8 carriles.

En determinados procedimientos de secuenciación por síntesis, se añade ADN molde monocatenario modificado con adaptador a la celda de flujo y se inmoviliza mediante hibridación a los anclajes en condiciones de dilución limitante. En contraste con la PCR en emulsión, los moldes de ADN se amplifican en la celda de flujo mediante amplificación de "puente", que depende de "arcos" de las hebras de ADN capturadas y que se hibridan a un oligonucleótido de anclaje adyacente. Múltiples ciclos de amplificación convierten el molde de ADN de una sola molécula en una "agrupación" de arcos amplificados de manera clonal y cada agrupación contiene aproximadamente 1000 moléculas clonales. Pueden generarse aproximadamente 50 x 10^6 agrupaciones independientes por celda de flujo. Para la secuenciación, se desnaturalizan las agrupaciones, y una reacción posterior de escisión química y el lavado dejan solamente hebras directas para la secuenciación de un solo extremo. La secuenciación de las hebras directas se inicia mediante la hibridación de un cebador complementario a las secuencias adaptadoras, seguido de la adición de polimerasa y una mezcla de cuatro terminadores de colorante reversibles fluorescentes de colores diferentes. Los terminadores se incorporan según la complementariedad de secuencia en cada hebra en una agrupación clonal. Después de la incorporación, el exceso de reactivos se elimina por lavado, las agrupaciones se interrogan ópticamente, y se registra la fluorescencia. Con etapas químicas sucesivas, los terminadores de colorante reversibles se desbloquean, las etiquetas fluorescentes se separan y lavan y se realiza el siguiente ciclo de secuenciación. Este procedimiento iterativo de secuenciación por síntesis a veces requiere aproximadamente 2,5 días para generar longitudes de lectura de 36 bases. Con 50 x 10^6 agrupaciones por celda de flujo, la salida de secuencia total puede ser mayor de mil millones de pares de bases (Gb) por ejecución analítica.

Otra tecnología de secuenciación de ácidos nucleicos que puede usarse con un método descrito en el presente documento es la secuenciación 454 (Roche). La secuenciación 454 usa un sistema de pirosecuenciación paralela a gran escala capaz de secuenciar aproximadamente 400-600 megabases de ADN por prueba. El procedimiento implica normalmente dos etapas. En la primera etapa, el ácido nucleico de muestra (por ejemplo, ADN) algunas veces se fracciona en fragmentos más pequeños (300-800 pares de bases) y se pulen (se hacen romos en cada extremo). Después, los adaptadores cortos se ligan a los extremos de los fragmentos. Estos adaptadores proporcionan secuencias cebadoras para la amplificación y secuenciación de los fragmentos de la biblioteca de muestras. Un adaptador (adaptador B) contiene una etiqueta de 5'-biotina para inmovilizar la biblioteca de ADN sobre perlas recubiertas con estreptavidina. Después de la reparación de la mella, la hebra no biotinilada se libera y se usa como biblioteca de ADN molde monocatenario (ADNmmc). La biblioteca de ADNmmc se evalúa para determinar su calidad y la cantidad óptima (copias de ADN por perla) necesaria para emPCR se determina mediante titulación. La biblioteca de ADNmmc se inmoviliza sobre perlas. Las perlas que contienen un fragmento de biblioteca portan una sola molécula de ADNmmc. La biblioteca unida a perlas se emulsiona con los reactivos de amplificación en una mezcla de agua en aceite. Cada perla se captura dentro de su propio microrreactor en el que se produce la amplificación por PCR. Esto produce fragmentos de ADN inmovilizados en perlas, amplificados de manera clonal.

En la segunda etapa de la secuenciación 454, se añaden perlas de biblioteca de ADN molde monocatenario a una mezcla de incubación que contiene ADN polimerasa y se estratifican con perlas que contienen sulfurilasa y luciferasa en un dispositivo que contiene pocillos de tamaño de picolitros. La pirosecuenciación se realiza en cada fragmento de ADN en paralelo. La adición de uno o más nucleótidos genera una señal de luz registrada por una cámara CCD en un instrumento de secuenciación. La intensidad de la señal es proporcional al número de nucleótidos incorporados. La pirosecuenciación aprovecha la liberación de pirofosfato (PPi) tras la adición de nucleótidos. PPi se convierte en ATP por la ATP sulfurilasa en presencia de adenosina 5'-fosfosulfato. La luciferasa usa ATP para convertir la luciferina en oxiluciferina, y esta reacción genera luz que se distingue y analiza (véase, por ejemplo, Margulies, M. *et al.* Nature 437:376-380 (2005)).

Otra tecnología de secuenciación de ácidos nucleicos que puede usarse en un método proporcionado en el presente documento es la tecnología SOLiD™ de Applied Biosystems. En la secuenciación por ligamiento SOLiD™, se prepara una biblioteca de fragmentos de ácido nucleico a partir de la muestra y se usa para preparar poblaciones de perlas clonales. Con este método, una especie de fragmento de ácido nucleico estará presente en la superficie de cada perla (por ejemplo, perla magnética). El ácido nucleico de muestra (por ejemplo, ADN genómico) se corta en fragmentos y, posteriormente, los adaptadores se unen a los extremos 5' y 3' de los fragmentos para generar una biblioteca de fragmentos. Los adaptadores son normalmente secuencias adaptadoras universales de modo que la secuencia inicial de cada fragmento es conocida e idéntica. La PCR en emulsión se lleva a cabo en microrreactores que contienen todos los reactivos necesarios para la PCR. Después, los productos de PCR resultantes unidos a las perlas se unen covalentemente a un portaobjetos de vidrio. Después, los cebadores se hibridan a la secuencia adaptadora dentro del molde de biblioteca. Un conjunto de cuatro sondas di-base marcadas con fluorescencia compiten por el ligamiento al cebador de secuenciación. La especificidad de la sonda di-base se logra al interrogar cada 1ª y 2ª bases en cada reacción de ligamiento. Se realizan múltiples ciclos de ligamiento, detección y escisión, determinando el número de ciclos la longitud de lectura eventual. Después de una serie de ciclos de ligamiento, se retira el producto de extensión y el molde

se restablece con un cebador complementario a la posición n-1 para una segunda tanda de ciclos de ligamiento. A menudo, se completan cinco tandas de restablecimiento del cebador para cada etiqueta de secuencia. Mediante el procedimiento de restablecimiento del cebador, cada base se interroga en dos reacciones de ligamiento independientes por medio de dos cebadores diferentes. Por ejemplo, la base en la posición de lectura 5 se somete a ensayo mediante el cebador número 2 en el ciclo de ligamiento 2 y mediante el cebador número 3 en el ciclo de ligamiento 1.

Otra tecnología de secuenciación de ácidos nucleicos que puede usarse en un método descrito en el presente documento es la secuenciación de una sola molécula verdadera (tSMS) Helicos. En la técnica tSMS, se añade una secuencia de poliA al extremo 3' de cada hebra de ácido nucleico (por ejemplo, ADN) de la muestra. Cada hebra se marca mediante la adición de un nucleótido de adenosina marcado con fluorescencia. Después, las hebras de ADN se hibridan a una celda de flujo, que contiene millones de sitios de captura de oligo-T que se inmovilizan en la superficie de la celda de flujo. Los moldes pueden estar a una densidad de aproximadamente 100 millones de moldes/cm^2. La celda de flujo se carga después en un aparato de secuenciación y un láser ilumina la superficie de la celda de flujo, lo que revela la posición de cada molde. Una cámara CCD puede mapear la posición de los moldes en la superficie de la celda de flujo. Después, el molde de etiqueta fluorescente se escinde y retira por lavado. La reacción de secuenciación comienza introduciendo una ADN polimerasa y un nucleótido marcado de manera fluorescente. El ácido nucleico de oligo-T funciona como cebador. La polimerasa incorpora los nucleótidos marcados al cebador de una manera dirigida por molde. La polimerasa y los nucleótidos no incorporados se retiran. Los moldes que tienen la incorporación dirigida del nucleótido marcado de manera fluorescente se detectan mediante la obtención de imágenes de la superficie de la celda de flujo. Después de la obtención de imágenes, una etapa de escisión retira la etiqueta fluorescente, y el procedimiento se repite con otros nucleótidos marcados con fluorescencia hasta que se logra la longitud de lectura deseada. Se recopila información de secuencia con cada etapa de adición de nucleótidos (véase, por ejemplo, Harris T. D. *et al.*, Science 320: 106-109 (2008)).

Otra tecnología de secuenciación de ácidos nucleicos que puede usarse en un método proporcionado en el presente documento es la tecnología de secuenciación en tiempo real de una sola molécula (SMRT™) de Pacific Biosciences. Con este método, cada una de las cuatro bases de ADN se une a uno de los cuatro colorantes fluorescentes diferentes. Estos colorantes se fosfoligan. Una sola ADN polimerasa se inmoviliza con una sola molécula de ADN monocatenario molde en la parte inferior de una guía de ondas en modo cero (ZMW). Una ZMW es una estructura de confinamiento que permite la observación de la incorporación de un solo nucleótido por ADN polimerasa contra el fondo de nucleótidos fluorescentes que se difunden rápidamente dentro y fuera de la ZMW (en microsegundos). Lleva varios milisegundos incorporar un nucleótido en una hebra en crecimiento. Durante este tiempo, la etiqueta fluorescente se excita y produce una señal fluorescente, y la etiqueta fluorescente se escinde. La detección de la fluorescencia correspondiente del colorante indica qué base se incorporó. Luego se repite el procedimiento.

Otra tecnología de secuenciación de ácidos nucleicos que puede usarse en un método descrito en el presente documento es la secuenciación de una sola molécula ION-TORRENT (Life Technologies) que asocia la tecnología de semiconductores con una química de secuenciación simple para traducir directamente la información codificada químicamente (A, C, G, T) en información digital (0, 1) en un chip semiconductor. ION-TORRENT usa un alineamiento de alta densidad de pocillos micromecanizados para realizar la secuenciación de ácidos nucleicos de una manera masiva en paralelo. Cada pocillo contiene una molécula de ADN diferente. Debajo de los pocillos hay una capa sensible a iones y debajo de eso un sensor de iones. Normalmente, cuando una polimerasa incorpora un nucleótido en una hebra de ADN, se libera un ion de hidrógeno como subproducto. Si un nucleótido, por ejemplo, una C, se añade a un molde de ADN y, después, se incorpora en una hebra de ADN, se liberará un ion de hidrógeno. La carga de ese ion cambiará el pH de la disolución, que puede detectarse mediante un sensor de iones. Un secuenciador puede identificar la base, que va directamente de información química a información digital. Después, el secuenciador inunda secuencialmente el chip con un nucleótido después de otro. Si el siguiente nucleótido que inunda el chip no coincide, no se registrará ningún cambio de tensión y no se identificará ninguna base. Si hay dos bases idénticas en la hebra de ADN, la tensión será doble y el chip registrará dos bases idénticas identificadas. Debido a que esta es la detección directa (es decir, detección sin exploración, cámaras o luz), cada incorporación de nucleótidos se registra en segundos.

Otra tecnología de secuenciación de ácidos nucleicos que puede usarse en un método descrito en el presente documento es la matriz de transistores de efecto de campo sensible a sustancias químicas (CHEMFET). En un ejemplo de esta técnica de secuenciación, las moléculas de ADN se colocan en cámaras de reacción, y las moléculas de molde pueden hibridarse a un cebador de secuenciación unido a una polimerasa. La incorporación de uno o más trifosfatos en una nueva hebra de ácido nucleico en el extremo 3' del cebador de secuenciación puede detectarse mediante un cambio en la corriente mediante un sensor CHEMFET. Una matriz puede tener múltiples sensores CHEMFET. En otro ejemplo, los ácidos nucleicos sencillos se unen a perlas, y los ácidos nucleicos pueden amplificarse en la perla, y las perlas individuales pueden transferirse a cámaras de reacción individuales en una matriz CHEMFET, teniendo cada cámara un sensor CHEMFET, y los ácidos nucleicos pueden secuenciarse (véase, por ejemplo, la publicación de solicitud de patente estadounidense n.° 2009/0026082).

Otra tecnología de secuenciación de ácidos nucleicos que puede usarse en un método descrito en el presente documento es la microscopia electrónica. En un ejemplo de esta técnica de secuenciación, las moléculas de ácido nucleico individuales (por ejemplo, ADN) se marcan con el uso de marcadores metálicos que pueden distinguirse con el uso de un microscopio

electrónico. Después, estas moléculas se estiran en una superficie plana y se obtienen imágenes con el uso de un microscopio electrónico para medir secuencias (véase, por ejemplo, Moudrianakis E. N. y Beer M. Proc Natl Acad Sci USA. marzo de 1965; 53: 564-71). En algunos casos, se usa microscopia electrónica de transmisión (TEM) (por ejemplo, método TEM de Halcyon Molecular). Este método, denominado nanotransferencia rápida de colocación de moléculas individuales (IMPRNT), incluye el uso de obtención de imágenes con microscopio electrónico de transmisión de resolución de un solo átomo de alto peso molecular (por ejemplo, aproximadamente 150 kb o más) marcado selectivamente con marcadores de átomos pesados y la disposición de estas moléculas en películas ultradelgadas en matrices paralelas ultradensas (3 nm de hebra a hebra) con una separación constante de base a base. El microscopio electrónico se usa para captar imágenes de las moléculas en las películas para determinar la posición de los marcadores de átomos pesados y para extraer información de secuencia base del ADN (véase, por ejemplo, la solicitud de patente internacional n.° WO 2009/046445).

Otros métodos de secuenciación que pueden usarse para llevar a cabo los métodos en el presente documento incluyen PCR digital y secuenciación por hibridación. La reacción en cadena de la polimerasa digital (PCR digital o dPCR) puede usarse para identificar y cuantificar directamente los ácidos nucleicos en una muestra. La PCR digital puede llevarse a cabo en una emulsión, en algunas implementaciones. Por ejemplo, los ácidos nucleicos individuales se separan, por ejemplo, en un dispositivo de cámara microfluídico, y cada ácido nucleico se amplifica individualmente mediante PCR. Los ácidos nucleicos pueden separarse de tal manera que no haya más de un ácido nucleico por pocillo. En algunas implementaciones, pueden usarse sondas diferentes para distinguir diversos alelos (por ejemplo, alelos fetales y alelos maternos). Pueden enumerarse alelos para determinar el número de copias. En la secuenciación por hibridación, el método incluye poner en contacto una pluralidad de secuencias de polinucleótidos con una pluralidad de sondas de polinucleótidos, en el que cada una de la pluralidad de sondas de polinucleótidos puede anclarse, opcionalmente, a un sustrato. El sustrato puede ser una superficie plana con un alineamiento de secuencias de nucleótidos conocidas, en algunas implementaciones. El patrón de hibridación al alineamiento puede usarse para determinar las secuencias de polinucleótidos presentes en la muestra. En algunas implementaciones, cada sonda se conecta a una perla, por ejemplo, una perla magnética o similar. La hibridación a las perlas puede identificarse y usarse para identificar la pluralidad de secuencias de polinucleótidos dentro de la muestra.

En algunas implementaciones, la secuenciación por nanoporos puede usarse en un método descrito en el presente documento. La secuenciación por nanoporos es una tecnología de secuenciación de una sola molécula mediante la cual una sola molécula de ácido nucleico (por ejemplo, ADN) se secuencia directamente a medida que pasa a través de un nanoporo. Un nanoporo es un pequeño agujero o canal, del orden de 1 nanómetro de diámetro. Determinadas proteínas celulares transmembrana pueden actuar como nanoporos (por ejemplo, alfa-hemolisina). En algunos casos, pueden sintetizarse nanoporos (por ejemplo, usando una plataforma de silicio). La inmersión de un nanoporo en un fluido conductor y la aplicación de un potencial a través del mismo da como resultado una ligera corriente eléctrica debido a la conducción de iones a través del nanoporo. La cantidad de corriente que fluye es sensible al tamaño del nanoporo. A medida que una molécula de ADN pasa a través de un nanoporo, cada nucleótido en la molécula de ADN obstruye el nanoporo en un grado diferente y genera cambios característicos en la corriente. La cantidad de corriente que puede pasar a través del nanoporo en cualquier momento dado varía, por tanto, dependiendo de si el nanoporo está bloqueado por una A, una C, una G, una T o, en algunos casos, una metil-C. El cambio en la corriente a través del nanoporo a medida que la molécula de ADN pasa a través del nanoporo representa una lectura directa de la secuencia de ADN. En algunos casos puede usarse un nanoporo para identificar bases individuales de ADN a medida que pasan a través del nanoporo en el orden correcto (véase, por ejemplo, Soni GV y Meller A. Clin.Chern. 53: 1996-2001 (2007); solicitud de patente internacional n.° WO2010/004265).

Existen varias maneras en que pueden usarse nanoporos para secuenciar moléculas de ácido nucleico. En algunas implementaciones, se usa una enzima exonucleasa, tal como una desoxirribonucleasa. En este caso, la enzima exonucleasa se usa para separar secuencialmente nucleótidos de una molécula de ácido nucleico (por ejemplo, ADN). Luego se detectan los nucleótidos y se discriminan por el nanoporo en orden de su liberación, leyendo por tanto la secuencia de la hebra original. Para tal implementación, la enzima exonucleasa puede unirse al nanoporo de tal manera que una proporción de los nucleótidos liberados de la molécula de ADN puedan entrar e interaccionar con el canal del nanoporo. La exonucleasa puede unirse a la estructura del nanoporo en un sitio muy próximo a la parte del nanoporo que forma la abertura del canal. En algunos casos, la enzima exonucleasa puede unirse a la estructura del nanoporo de tal manera que su sitio de trayectoria de salida de nucleótidos se oriente hacia la parte del nanoporo que forma parte de la abertura.

En algunas implementaciones, la secuenciación por nanoporos de ácidos nucleicos implica el uso de una enzima que empuja o extrae la molécula de ácido nucleico (por ejemplo, ADN) a través del poro. En este caso, la corriente iónica fluctúa cuando un nucleótido en la molécula de ADN pasa a través del poro. Las fluctuaciones en la corriente son indicativas de la secuencia de ADN. Para tal implementación, la enzima puede unirse a la estructura del nanoporo de tal manera que pueda empujar o extraer el ácido nucleico diana a través del canal de un nanoporo sin interferir en el flujo de corriente iónica a través del poro. La enzima puede unirse a la estructura del nanoporo en un sitio muy próximo a la parte de la estructura que forma parte de la abertura. La enzima puede unirse a la subunidad, por ejemplo, de tal manera que su sitio activo se oriente hacia la parte de la estructura que forma parte de la abertura.

En algunas implementaciones, la secuenciación por nanoporos de ácidos nucleicos implica la detección de subproductos de polimerasa en estrecha proximidad con un detector de nanoporos. En este caso, los fosfatos de nucleósidos (nucleótidos) están marcados con etiquetas de modo que una especie marcada con fosfato se libera tras la adición de una polimerasa a la hebra de nucleótidos y el poro detecta la especie marcada con fosfato. Normalmente, la especie de fosfato contiene un marcador específico para cada nucleótido. A medida que los nucleótidos se añaden secuencialmente

a la hebra de ácido nucleico, se detectan los subproductos de la adición de bases. El orden en que se detectan las especies marcadas con fosfato puede usarse para determinar la secuencia de la hebra de ácido nucleico.

La longitud de la lectura de secuencia se asocia a menudo con la tecnología de secuenciación particular. Los métodos de alto rendimiento, por ejemplo, proporcionan lecturas de secuencia que pueden variar en tamaño desde decenas hasta cientos de pares de bases (pb). La secuenciación por nanoporos, por ejemplo, puede proporcionar lecturas de secuencia que pueden variar en tamaño desde decenas hasta cientos y hasta miles de pares de bases. En algunas implementaciones, las lecturas de secuencia tienen una longitud media, una mediana de longitud o longitud promedio de aproximadamente 15 pb a 900 pb de longitud (por ejemplo, aproximadamente 20 pb, aproximadamente 25 pb, aproximadamente 30 pb, aproximadamente 35 pb, aproximadamente 40 pb, aproximadamente 45 pb, aproximadamente 50 pb, aproximadamente 55 pb, aproximadamente 60 pb, aproximadamente 65 pb, aproximadamente 70 pb, aproximadamente 75 pb, aproximadamente 80 pb, aproximadamente 85 pb, aproximadamente 90 pb, aproximadamente 95 pb, aproximadamente 100 pb, aproximadamente 110 pb, aproximadamente 120 pb, aproximadamente 130 pb, aproximadamente 140 pb, aproximadamente 150 pb, aproximadamente 200 pb, aproximadamente 250 pb, aproximadamente 300 pb, aproximadamente 350 pb, aproximadamente 400 pb, aproximadamente 450 pb o aproximadamente 500 pb. En algunas implementaciones, las lecturas de secuencia tienen una longitud media, mediana de longitud o longitud promedio de aproximadamente 1000 pb o más.

En algunas implementaciones, se realiza la secuenciación específica de cromosoma. En algunas implementaciones, la secuenciación específica de cromosoma se realiza utilizando DANSR (análisis digital de regiones seleccionadas). El análisis digital de regiones seleccionadas permite la cuantificación simultánea de cientos de loci mediante catenación dependiente de ADNlc de dos oligonucleótidos específicos de locus por medio de un oligo intermedio 'puente' para formar un molde para PCR. En algunas implementaciones, la secuenciación específica de cromosoma se realiza mediante la generación de una biblioteca enriquecida en secuencias específicas de cromosoma. En algunas implementaciones, las lecturas de secuencia se obtienen solamente para un conjunto seleccionado de cromosomas. En algunas implementaciones, las lecturas de secuencia se obtienen solamente para los cromosomas 21, 18 y 13.

En algunas implementaciones, los ácidos nucleicos pueden incluir una señal fluorescente o información de etiqueta de secuencia. La cuantificación de la señal o etiqueta puede usarse en una variedad de técnicas tales como, por ejemplo, citometría de flujo, reacción en cadena de la polimerasa cuantitativa (qPCR), electroforesis en gel, análisis de chip génico, microalineamiento, espectrometría de masas, análisis citofluorimétrico, microscopia de fluorescencia, microscopia de barrido láser confocal, citometría de barrido láser, cromatografía de afinidad, separación en modo discontinuo manual, suspensión en campo eléctrico, secuenciación, y combinaciones de los mismos.

Módulo de secuenciación

La secuenciación y la obtención de lecturas de secuenciación pueden proporcionarse por un módulo de secuenciación o por un aparato que comprende un módulo de secuenciación. Un ‘‘módulo de recepción de secuencia’’, tal como se usa en el presente documento, es lo mismo que un ‘‘módulo de secuenciación’’. Un aparato que comprende un módulo de secuenciación puede ser cualquier aparato que determina la secuencia de un ácido nucleico a partir de una tecnología de secuenciación conocida en la técnica. En determinadas implementaciones, un aparato que comprende un módulo de secuenciación realiza una reacción de secuenciación conocida en la técnica. Generalmente, un módulo de secuenciación proporciona una secuencia de ácido nucleico leída según los datos de una reacción de secuenciación (por ejemplo, señales generadas a partir de un aparato de secuenciación). En algunas implementaciones, se requiere un módulo de secuenciación o un aparato que comprende un módulo de secuenciación para proporcionar lecturas de secuenciación. En algunas implementaciones, un módulo de secuenciación puede recibir, obtener, acceder o recuperar lecturas de secuencia de otro módulo de secuenciación, periférico de ordenador, operador, servidor, disco duro, aparato o de una fuente adecuada. Algunas veces, un módulo de secuenciación puede manipular las lecturas de secuencia. Por ejemplo, un módulo de secuenciación puede alinear, ensamblar, fragmentar, complementar, complementar de manera inversa, comprobar errores o corregir errores de lecturas de secuencia. Un aparato que comprende un módulo de secuenciación puede comprender al menos un procesador. En algunas implementaciones, las lecturas de secuenciación se proporcionan por un aparato que incluye un procesador (por ejemplo, uno o más procesadores), procesador que puede realizar y/o implementar una o más instrucciones (por ejemplo, procesos, rutinas y/o subrutinas) desde el módulo de secuenciación. En algunas implementaciones, las lecturas de secuenciación se proporcionan por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunas implementaciones, un módulo de secuenciación funciona con uno o más procesadores externos (por ejemplo, una red interna o externa, un servidor, dispositivo de almacenamiento y/o una red de almacenamiento (por ejemplo, una nube)). Algunas veces, un módulo de secuenciación recopila, ensambla y/o recibe información y/o datos de otro módulo, aparato, periférico, componente o componente especializado (por ejemplo, un secuenciador). En algunas implementaciones, las lecturas de secuenciación se proporcionan por un aparato que comprende uno o más de los siguientes: una o más celdas de flujo, una cámara, un fotodetector, una fotocelda, componentes de manipulación de fluidos, una impresora, una pantalla de visualización (por ejemplo, un LED, LCT o CRT) y similares. A menudo, un módulo de secuenciación recibe, recopila y/o ensambla lecturas de secuencia. Algunas veces, un módulo de secuenciación acepta y recopila información y/o datos de entrada de un operador de un aparato. Por ejemplo, algunas veces un operador de un aparato proporciona instrucciones, una constante, un valor umbral, una fórmula o un valor predeterminado a un módulo. Algunas veces, un módulo de secuenciación puede transformar información y/o datos que recibe en una secuencia de ácido nucleico contigua. En algunas implementaciones, se imprime o visualiza una

secuencia de ácido nucleico proporcionada por un módulo de secuenciación. En algunas implementaciones, las lecturas de secuencia se proporcionan por un módulo de secuenciación y se transfieren de un módulo de secuenciación a un aparato o un aparato que comprende cualquier periférico, componente o componente especializado adecuado. En algunas implementaciones, se proporcionan información y/o datos desde un módulo de secuenciación a un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunos casos, la información y/o los datos relacionados con lecturas de secuencia pueden transferirse de un módulo de secuenciación a cualquier otro módulo adecuado. En algunas implementaciones, un módulo de secuenciación puede transferir lecturas de secuencia a un módulo de mapeo o módulo de recuento.

Lecturas de mapeo

El mapeo de lecturas de secuencia de nucleótidos (es decir, información de secuencia de un fragmento cuya posición genómica física es desconocida) puede realizarse de varias maneras y a menudo comprende la alineación de las lecturas de secuencia obtenidas con una secuencia coincidente en un genoma de referencia (por ejemplo, Li *et al*., "Mapping short DNA sequencing reads and calling variants using mapping quality score", Genome Res., 19 de agosto de 2008.) En tales alineaciones, las lecturas de secuencia se alinean generalmente con una secuencia de referencia y aquellas que se alinean se designan como "mapeadas" o una "etiqueta de secuencia". En algunos casos, una lectura de secuencia mapeada se denomina "coincidencia" o "recuento". En algunas implementaciones, las lecturas de secuencia mapeadas se agrupan entre sí según diversos parámetros y se asignan a secciones genómicas particulares, que se describen con mayor detalle más adelante.

Tal como se usa en el presente documento, los términos "alineado", "alineación" o "alinear" se refieren a dos o más secuencias de ácido nucleico que pueden identificarse como apareamiento (por ejemplo, identidad del 100 %) o apareamiento parcial. Las alineaciones pueden realizarse manualmente o mediante un algoritmo informático, los ejemplos de los mismos incluyen el programa informático de alineación local eficiente de datos de nucleótidos (ELAND) distribuido como parte del flujo de trabajo de análisis genómico de Illumina. La alineación de una lectura de secuencia puede tener un 100 % de apareamiento de secuencia. En algunos casos, una alineación es menor del 100 % de apareamiento de secuencias (es decir, apareamiento no perfecto, apareamiento parcial, alineación parcial). En algunas implementaciones, una alineación es de aproximadamente el 99 %, 98 %, 97 %, 96 %, 95 %, 94 %, 93 %, 92 %, 91 %, 90 %, 89 %, 88 %, 87 %, 86 %, 85 %, 84 %, 83 %, 82 %, 81 %, 80 %, 79 %, 78 %, 77 %, 76 % o 75 % de apareamiento. En algunas implementaciones, una alineación comprende un apareamiento erróneo. En algunas implementaciones, una alineación comprende 1,2, 3, 4 o 5 apareamientos erróneos. Dos o más secuencias pueden alinearse usando cualquier hebra. En algunos casos, una secuencia de ácido nucleico se alinea con el complemento inverso de otra secuencia de ácido nucleico.

Pueden usarse diversos métodos computacionales para mapear cada secuencia leída en una sección genómica. Los ejemplos no limitativos de algoritmos informáticos que pueden usarse para alinear secuencias incluyen, sin limitación, BLAST, BLITZ, FASTA, BOWTIE 1, BOWTIE 2, ELAND, MAQ, PROBEMATCH, SOAP o SEQMAP, o variaciones de los mismos o combinaciones de los mismos. En algunas implementaciones, las lecturas de secuencia pueden alinearse con las secuencias en un genoma de referencia. En algunas implementaciones, las lecturas de secuencia pueden encontrarse y/o alinearse con secuencias en bases de datos de ácidos nucleicos conocidas en la técnica incluyendo, por ejemplo, GenBank, dbEST, dbSTS, EMBL (Laboratorio Europeo de Biología Molecular) y DDBJ (Banco de datos de ADN de Japón). Puede usarse Blast o herramientas similares para buscar las secuencias identificadas en una base de datos de secuencias. Entonces pueden usarse las coincidencias de búsqueda para clasificar las secuencias identificadas en secciones genómicas adecuadas (descritas de más adelante en el presente documento), por ejemplo.

La expresión "etiqueta de secuencia" se usa en el presente documento indistintamente con la expresión "etiqueta de secuencia mapeada" para referirse a una lectura de secuencia que se ha asignado de manera específica, es decir, mapeado, en una secuencia mayor, por ejemplo, un genoma de referencia, mediante alineación. Las etiquetas de secuencia mapeadas se mapean de manera única en un genoma de referencia, es decir, se asignan a una sola ubicación en el genoma de referencia. Las etiquetas que pueden mapearse en más de una ubicación en un genoma de referencia, es decir, etiquetas que no se mapean de manera única, no se incluyen en el análisis. Una "etiqueta de secuencia" puede ser una secuencia de ácido nucleico (por ejemplo, ADN) (es decir, leída) asignada de manera específica a una sección genómica y/o un cromosoma particular (es decir, uno de los cromosomas 1-22, X o Y para un sujeto humano). Una etiqueta de secuencia puede ser repetitiva o no repetitiva dentro de un solo segmento del genoma de referencia (por ejemplo, un cromosoma). En algunas implementaciones, las etiquetas de secuencia repetitiva se eliminan del análisis adicional (por ejemplo, cuantificación). En algunas implementaciones, una lectura puede mapearse de manera única o no única en porciones en el genoma de referencia. Se considera que una lectura se "mapea de manera única" si se alinea con una sola secuencia en el genoma de referencia. Se considera que una lectura se "mapea de manera no única" si se alinea con dos o más secuencias en el genoma de referencia. En algunas implementaciones, las lecturas mapeadas de manera no única se eliminan del análisis posterior (por ejemplo, cuantificación). Puede permitirse que un determinado grado pequeño de apareamiento erróneo (0-1) tenga en cuenta los polimorfismos de un solo nucleótido que pueden existir entre el genoma de referencia y las lecturas de las muestras individuales que se mapean, en determinadas implementaciones. En algunas implementaciones, no se permite ningún grado de apareamiento erróneo para que una lectura se mapee en una secuencia de referencia.

Tal como se usa en el presente documento, la expresión "genoma de referencia" puede referirse a cualquier genoma conocido, secuenciado o caracterizado, ya sea parcial o completo, de cualquier organismo o virus que pueda usarse para referenciar secuencias identificadas de un sujeto. Por ejemplo, un genoma de referencia usado para sujetos humanos, así como muchos otros organismos, puede encontrarse en el Centro Nacional de Información Biotecnológica en www.ncbi.nlm.nih.gov. Un "genoma" se refiere a la información genética completa de un organismo o virus, expresada en secuencias de ácido nucleico. Tal como se usa en el presente documento, una secuencia de referencia o un genoma de referencia es a menudo una secuencia genómica ensamblada o parcialmente ensamblada de un individuo o múltiples individuos. En algunas implementaciones, un genoma de referencia es una secuencia genómica ensamblada o parcialmente ensamblada de uno o más individuos humanos. En algunas implementaciones, un genoma de referencia comprende secuencias asignadas a cromosomas.

En determinadas implementaciones, en las que un ácido nucleico de muestra proviene de una mujer embarazada, una secuencia de referencia algunas veces no es del feto, la madre del feto o el padre del feto, y se denomina en el presente documento "referencia externa". Una referencia materna puede prepararse y usarse en algunas implementaciones. Cuando se prepara una referencia de la mujer embarazada ("secuencia de referencia materna") basándose en una referencia externa, las lecturas de ADN de la mujer embarazada que no contiene sustancialmente ADN fetal se mapean en menudo en la secuencia de referencia externa y se ensamblan. En determinadas implementaciones, la referencia externa es de ADN de un individuo que tiene sustancialmente la misma etnia que la mujer embarazada. Una secuencia de referencia materna puede no cubrir completamente el ADN genómico materno (por ejemplo, puede cubrir aproximadamente el 50 %, 60 %, 70 %, 80 %, 90 % o más del ADN genómico materno), y la referencia materna puede no aparearse perfectamente con la secuencia de ADN genómico materno (por ejemplo, la secuencia de referencia materna puede incluir múltiples apareamientos erróneos).

En algunos casos, se evalúa la capacidad de mapeo para una región genómica (por ejemplo, sección genómica, porción genómica, bin). La capacidad de mapeo es la capacidad de alinear de manera inequívoca una secuencia de nucleótidos leída con una porción de un genoma de referencia normalmente hasta un número especificado de apareamientos erróneos, incluyendo, por ejemplo, 0, 1, 2 o más apareamientos erróneos. Para una región genómica dada, la capacidad de mapeo esperada puede estimarse usando un enfoque de ventana deslizante de una longitud de lectura predeterminada y promediando los valores de capacidad de mapeo a nivel de lectura resultantes. Las regiones genómicas que comprenden tramos de secuencia de nucleótidos única tienen, algunas veces, un alto valor de capacidad de mapeo.

Módulo de mapeo

Las lecturas de secuencia pueden mapearse por un módulo de mapeo o por un aparato que comprende un módulo de mapeo, módulo de mapeo que mapea generalmente lecturas en un genoma de referencia o segmento del mismo. Un módulo de mapeo puede mapear lecturas de secuenciación mediante un método adecuado conocido en la técnica. En algunas implementaciones, se requiere un módulo de mapeo o un aparato que comprende un módulo de mapeo para proporcionar lecturas de secuencia mapeadas. Un aparato que comprende un módulo de mapeo puede comprender al menos un procesador. En algunas implementaciones, las lecturas de secuenciación mapeadas se proporcionan por un aparato que incluye un procesador (por ejemplo, uno o más procesadores), procesador que puede realizar y/o implementar una o más instrucciones (por ejemplo, procesos, rutinas y/o subrutinas) desde el módulo de mapeo. En algunas implementaciones, las lecturas de secuenciación se mapean por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunas implementaciones, un módulo de mapeo funciona con uno o más procesadores externos (por ejemplo, una red interna o externa, un servidor, dispositivo de almacenamiento y/o una red de almacenamiento (por ejemplo, una nube)). Un aparato puede comprender un módulo de mapeo y un módulo de secuenciación. En algunas implementaciones, las lecturas de secuencia se mapean por un aparato que comprende uno o más de los siguientes: una o más celdas de flujo, una cámara, componentes de manipulación de fluido, una impresora, una pantalla de visualización (por ejemplo, un LED, LCT o CRT) y similares. En algunas implementaciones, un módulo de mapeo puede recibir lecturas de secuencia de un módulo de secuenciación. Las lecturas de secuenciación mapeadas pueden transferirse de un módulo de mapeo a un módulo de recuento o un módulo de normalización, en algunas implementaciones.

Secciones genómicas

En algunas implementaciones, las lecturas de secuencia mapeadas (es decir, etiquetas de secuencia) se agrupan entre sí según diversos parámetros y se asignan a secciones genómicas particulares. A menudo, las lecturas de secuencia mapeadas individuales pueden usarse para identificar una cantidad de una sección genómica presente en una muestra. En algunas implementaciones, la cantidad de una sección genómica puede ser indicativa de la cantidad de una secuencia mayor (por ejemplo, un cromosoma) en la muestra. La expresión "sección genómica" puede denominarse además en el presente documento "ventana de secuencia", "sección", "bin", "locus", "región", "división" o "porción". En algunas implementaciones, una sección genómica es un cromosoma completo, un segmento de un cromosoma, un segmento de un genoma de referencia, múltiples porciones cromosómicas, múltiples cromosomas, porciones de múltiples cromosomas, y/o combinaciones de los mismos. A veces se predefine una sección genómica basándose en parámetros específicos. A veces, una sección genómica se define arbitrariamente basándose en la división de un genoma (por ejemplo, dividida por tamaño, segmentos, regiones contiguas, regiones contiguas de un tamaño definido arbitrariamente, y similares). En algunos casos, se delinea una sección genómica basándose en uno o más parámetros que incluyen, por ejemplo, longitud o una

característica o características particulares de la secuencia. Las secciones genómicas pueden seleccionarse, filtrarse y/o eliminarse de la consideración usando cualquier criterio adecuado conocido en la técnica o descrito en el presente documento. En algunas implementaciones, una sección genómica se basa en una longitud particular de la secuencia genómica. En algunas implementaciones, un método puede incluir el análisis de múltiples lecturas de secuencia mapeadas en una pluralidad de secciones genómicas. Las secciones genómicas pueden tener aproximadamente la misma longitud o las secciones genómicas pueden tener longitudes diferentes. Algunas veces, las secciones genómicas tienen aproximadamente la misma longitud. En algunos casos, se ajustan o ponderan secciones genómicas de longitudes diferentes. En algunas implementaciones, una sección genómica es de aproximadamente 10 kilobases (kb) a aproximadamente 100 kb, de aproximadamente 20 kb a aproximadamente 80 kb, de aproximadamente 30 kb a aproximadamente 70 kb, de aproximadamente 40 kb a aproximadamente 60 kb y, a veces, de aproximadamente 50 kb. En algunas implementaciones, una sección genómica es de aproximadamente 10 kb a aproximadamente 20 kb. Una sección genómica no se limita a ejecuciones contiguas de secuencia. Por tanto, las secciones genómicas pueden componerse de secuencias contiguas y/o no contiguas. Una sección genómica no se limita a un solo cromosoma. En algunas implementaciones, una sección genómica incluye la totalidad o parte de un cromosoma o la totalidad o parte de dos o más cromosomas. En algunos casos, las secciones genómicas pueden abarcar uno, dos o más cromosomas completos. Adicionalmente, las secciones genómicas pueden abarcar porciones de unión o desunidas de múltiples cromosomas.

En algunas implementaciones, las secciones genómicas pueden ser segmentos cromosómicos particulares en un cromosoma de interés, tales como, por ejemplo, cromosomas en los que se evalúa una variación genética (por ejemplo, una aneuploidía de los cromosomas 13, 18 y/o 21 o un cromosoma sexual). Una sección genómica puede ser además un genoma patógeno (por ejemplo, bacteriano, fúngico o viral) o fragmento del mismo. Las secciones genómicas pueden ser genes, fragmentos génicos, secuencias reguladoras, intrones, exones y similares.

En algunas implementaciones, un genoma (por ejemplo, genoma humano) se divide en secciones genómicas basándose en el contenido de información de las regiones. Las regiones genómicas resultantes pueden contener secuencias para múltiples cromosomas y/o pueden contener secuencias para porciones de múltiples cromosomas. En algunos casos, la división puede eliminar ubicaciones similares a través del genoma y mantener solamente regiones únicas. Las regiones eliminadas pueden estar dentro de un solo cromosoma o pueden abarcar múltiples cromosomas. Por tanto, el genoma resultante se recorta y se optimiza para una alineación más rápida, lo que permite a menudo concentrarse en secuencias identificables de manera única. En algunos casos, la división puede ponderar por disminución regiones similares. El procedimiento para ponderar por disminución una sección genómica se describe con mayor detalle más adelante. En algunas implementaciones, la división del genoma en regiones que transcienden a los cromosomas puede basarse en la ganancia de información producida en el contexto de la clasificación. Por ejemplo, el contenido de información puede cuantificarse usando el perfil de valor de p que mide la significación de ubicaciones genómicas particulares para distinguir entre grupos de sujetos normales y anómalos confirmados (por ejemplo, sujetos euploides y con trisomía, respectivamente). En algunas implementaciones, la división del genoma en regiones que transcienden a los cromosomas puede basarse en cualquier otro criterio, tal como, por ejemplo, velocidad/conveniencia mientras se alinean las etiquetas, alto o bajo contenido de GC, uniformidad del contenido de GC, otras medidas de contenido de secuencia (por ejemplo, fracción de nucleótidos individuales, fracción de pirimidinas o purinas, fracción de ácidos nucleicos naturales frente a no naturales, fracción de nucleótidos metilados y contenido de CpG), estado de metilación, temperatura de fusión del dúplex, predisposición a secuenciación o PCR, valor de incertidumbre asignado a bins individuales y/o una búsqueda dirigida para características particulares.

Densidad de etiqueta de secuencia

"Densidad de etiqueta de secuencia" se refiere al valor normalizado de etiquetas o lecturas de secuencia para una sección genómica definida, en el que la densidad de etiqueta de secuencia se usa para comparar diferentes muestras y para el análisis posterior. El valor de la densidad de la etiqueta de secuencia se normaliza a menudo dentro de una muestra. En algunas implementaciones, la normalización puede realizarse contando el número de etiquetas que se encuentran dentro de cada sección genómica; obteniendo una mediana del valor del recuento total de etiquetas de secuencia para cada cromosoma; obteniendo una mediana del valor de todos los valores autosómicos; y usando este valor como constante de normalización para representar las diferencias en el número total de etiquetas de secuencia obtenidas para diferentes muestras. Una densidad de etiqueta de secuencia algunas veces es de aproximadamente 1 para un cromosoma disómico. Las densidades de etiqueta de secuencia pueden variar según los artefactos de secuenciación, más particularmente sesgo de G/C, que puede corregirse usando un patrón externo o una referencia interna (por ejemplo, derivado de sustancialmente todas las etiquetas de secuencia (secuencias genómicas), que pueden ser, por ejemplo, un solo cromosoma o un valor calculado de todos los autosomas, en algunas implementaciones). Por tanto, el desequilibrio de dosificación de un cromosoma o regiones cromosómicas puede deducirse a partir de la representación porcentual del locus entre otras etiquetas secuenciadas mapeables del espécimen. Por tanto, el desequilibrio de dosificación de regiones cromosómicas o cromosomas particulares puede determinarse cuantitativamente y normalizarse. Los métodos para la normalización y cuantificación de la densidad de etiqueta de secuencia se describen con mayor detalle más adelante.

En algunas implementaciones, una proporción de todas las lecturas de secuencia son de un cromosoma involucrado en una aneuploidía (por ejemplo, el cromosoma 13, el cromosoma 18, el cromosoma 21), y otras lecturas de secuencia son de otros cromosomas. Teniendo en cuenta el tamaño relativo del cromosoma involucrado en la aneuploidía (por ejemplo,

"cromosoma diana": cromosoma 21) en comparación con otros cromosomas, podría obtenerse una frecuencia normalizada, dentro de un rango de referencia, de secuencias específicas del cromosoma diana, en algunas implementaciones. Si el feto tiene una aneuploidía en un cromosoma diana, entonces la frecuencia normalizada de las secuencias derivadas del cromosoma diana es estadísticamente mayor que la frecuencia normalizada de las secuencias derivadas del cromosoma no diana, lo que permite así la detección de la aneuploidía. El grado de cambio en la frecuencia normalizada dependerá de la concentración fraccional de ácidos nucleicos fetales en la muestra analizada, en algunas implementaciones.

Recuentos

Las lecturas de secuencia que se mapean o dividen basándose en una característica o variable seleccionada pueden cuantificarse para determinar el número de lecturas que se mapean en una sección genómica (por ejemplo, bin, división, porción genómica, porción de un genoma de referencia, porción de un cromosoma y similares), en algunas implementaciones. Algunas veces, la cantidad de lecturas de secuencia que se mapean en una sección genómica se denominan recuentos (por ejemplo, un recuento). A menudo, un recuento se asocia con una sección genómica. Algunas veces, los recuentos para dos o más secciones genómicas (por ejemplo, un conjunto de secciones genómicas) se manipulan matemáticamente (por ejemplo, se promedian, suman, normalizan, similares o combinaciones de los mismos). En algunas implementaciones, se determina un recuento a partir de algunas o todas las lecturas de secuencia mapeadas en (es decir, asociadas con) una sección genómica. En determinadas implementaciones, se determina un recuento a partir de un subconjunto predefinido de lecturas de secuencia mapeadas. Los subconjuntos predefinidos de lecturas de secuencia mapeadas pueden definirse o seleccionarse con el uso de cualquier característica o variable adecuada. En algunas implementaciones, los subconjuntos predefinidos de lecturas de secuencia mapeadas pueden incluir de 1 a n lecturas de secuencia, en el que n representa un número igual a la suma de todas las lecturas de secuencia generadas a partir de una muestra de sujeto de prueba o de sujeto de referencia.

Algunas veces, un recuento se deriva de lecturas de secuencia que se procesan o manipulan mediante un método, una operación o un procedimiento matemático adecuado conocido en la técnica. Algunas veces, un recuento se deriva de lecturas de secuencia asociadas con una sección genómica en la que algunas o todas las lecturas de secuencia se ponderan, eliminan, filtran, normalizan, ajustan, promedian, derivan como una media, se suman, o se restan o procesan mediante una combinación de los mismos. En algunas implementaciones, un recuento se deriva de lecturas de secuencia sin procesar y/o lecturas de secuencia filtradas. Un recuento (por ejemplo, recuentos) puede determinarse mediante un método, una operación o un procedimiento matemático adecuado. Algunas veces, un valor de recuento se determina mediante un procedimiento matemático. Algunas veces, un valor de recuento es un promedio, una media o suma de lecturas de secuencia mapeadas en una sección genómica. A menudo, un recuento es un número medio de recuentos. En algunas implementaciones, un recuento se asocia con un valor de incertidumbre. Los recuentos pueden procesarse (por ejemplo, normalizarse) mediante un método conocido en la técnica y/o tal como se describe en el presente documento (por ejemplo, normalización basada en bins, normalización por contenido de GC, regresión lineal y no lineal por mínimos cuadrados, LOESS de GC, LOWESS, PERUN, RM, GCRM, cQn y/o combinaciones de los mismos).

Los recuentos (por ejemplo, recuentos sin procesar, filtrados y/o normalizados) pueden procesarse y normalizarse a una o más elevaciones. Las elevaciones y los perfiles se describen con mayor detalle más adelante. Algunas veces, los recuentos pueden procesarse y/o normalizarse a una elevación de referencia. Las elevaciones de referencia se abordan más adelante en el presente documento. Los recuentos procesados según una elevación (por ejemplo, recuentos procesados) pueden asociarse con un valor de incertidumbre (por ejemplo, una varianza calculada, un error, desviación estándar, valor de p, desviación media absoluta, etc.). Un valor de incertidumbre define normalmente un rango por encima y por debajo de una elevación. Puede usarse un valor de desviación en lugar de un valor de incertidumbre y los ejemplos no limitativos de medidas de desviación incluyen desviación estándar, desviación absoluta promedio, mediana de desviación absoluta, puntuación estándar (por ejemplo, puntuación Z, valor de Z, puntuación normal, variable normalizada) y similares.

Los recuentos se obtienen a menudo a partir de una muestra de ácido nucleico de una mujer embarazada que porta un feto. Los recuentos de lecturas de secuencia de ácido nucleico mapeadas en una sección genómica son a menudo recuentos representativos tanto del feto como de la madre del feto (por ejemplo, un sujeto femenino gestante). Algunas veces, algunos de los recuentos mapeados en una sección genómica son de un genoma fetal y algunos de los recuentos mapeados en la misma sección genómica son del genoma materno.

Módulo de recuento

Los recuentos pueden proporcionarse por un módulo de recuento o por un aparato que comprende un módulo de recuento. Un módulo de recuento puede determinar, ensamblar y/o visualizar recuentos según un método de recuento conocido en la técnica. Generalmente, un módulo de recuento determina o ensambla recuentos según una metodología de recuento conocida en la técnica. En algunas implementaciones, se requiere un módulo de recuento o un aparato que comprende un módulo de recuento para proporcionar recuentos. Un aparato que comprende un módulo de recuento puede comprender al menos un procesador. En algunas implementaciones, los recuentos se proporcionan por un aparato que incluye un procesador (por ejemplo, uno o más procesadores), procesador que puede realizar y/o implementar una o más instrucciones (por ejemplo, procesos, rutinas y/o subrutinas) del módulo de recuento. En algunas implementaciones, las

lecturas se cuentan con un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunas implementaciones, un módulo de recuento funciona con uno o más procesadores externos (por ejemplo, una red interna o externa, un servidor, dispositivo de almacenamiento y/o una red de almacenamiento (por ejemplo, una nube)). En algunas implementaciones, las lecturas se cuentan por un aparato que comprende uno o más de los siguientes: un módulo de secuenciación, un módulo de mapeo, una o más celdas de flujo, una cámara, componentes de manipulación de fluidos, una impresora, una pantalla de visualización (por ejemplo, un LED, LCT o CRT) y similares. Un módulo de recuento puede recibir información y/o datos de un módulo de secuenciación y/o un módulo de mapeo, transformar la información y/o los datos y proporcionar recuentos (por ejemplo, recuentos mapeados en secciones genómicas). Un módulo de recuento puede recibir lecturas de secuencia mapeadas de un módulo de mapeo. Un módulo de recuento puede recibir lecturas de secuencia mapeadas normalizadas de un módulo de mapeo o de un módulo de normalización. Un módulo de recuento puede transferir información y/o datos relacionados con recuentos (por ejemplo, recuentos, recuentos ensamblados y/o visualizaciones de recuentos) a cualquier otro aparato, periférico o módulo adecuado. Algunas veces, la información y/o los datos relacionados con recuentos se transfieren de un módulo de recuento a un módulo de normalización, un módulo de representación gráfica, un módulo de categorización y/o un módulo de resultados.

Procesamiento de datos

Las lecturas de secuencia mapeadas que se han contado se denominan en el presente documento datos sin procesar, ya que los datos representan recuentos sin manipular (por ejemplo, recuentos sin procesar). En algunas implementaciones, los datos de lectura de secuencia en un conjunto de datos pueden procesarse adicionalmente (por ejemplo, manipularse matemática y/o estadísticamente) y/o visualizarse para facilitar la provisión de un resultado. En determinadas implementaciones, los conjuntos de datos, incluyendo conjuntos de datos más grandes, pueden beneficiarse del preprocesamiento para facilitar un análisis adicional. El preprocesamiento de conjuntos de datos a veces implica la eliminación de secciones o bins genómicos redundantes y/o no informativos (por ejemplo, bins con datos no informativos, lecturas mapeadas redundantes, secciones genómicas o bins con mediana de recuentos cero, secuencias sobrerrepresentadas o subrepresentadas). Sin desear limitarse por la teoría, el procesamiento y/o preprocesamiento de datos pueden (i) eliminar datos con ruido, (ii) eliminar datos no informativos, (iii) eliminar datos redundantes, (iv) reducir la complejidad de conjuntos de datos más grandes y/o (v) facilitar la transformación de los datos de una forma en una o más formas diferentes. Los términos "preprocesamiento" y "procesamiento", cuando se usan con respecto a los datos o conjuntos de datos, se denominan colectivamente en el presente documento "procesamiento". El procesamiento puede hacer que los datos sean más susceptibles de análisis adicional, y puede generar un resultado en algunas implementaciones.

La expresión "datos con ruido", tal como se usa en el presente documento, se refiere a (a) datos que tienen una varianza significativa entre los puntos de datos cuando se analizan o representan gráficamente, (b) datos que tienen una desviación estándar significativa (por ejemplo, mayor de 3 desviaciones estándar), (c) datos que tienen un error estándar de la media significativo, similares y combinaciones de los anteriores. Los datos con ruido suceden, algunas veces, debido a la cantidad y/o calidad del material de partida (por ejemplo, muestra de ácido nucleico) y, algunas veces, suceden como parte de los procedimientos para preparar o replicar el ADN usado para generar lecturas de secuencia. En determinadas implementaciones, el ruido resulta de determinadas secuencias sobrerrepresentadas cuando se preparan usando métodos basados en PCR. Los métodos descritos en el presente documento pueden reducir o eliminar la contribución de los datos con ruido y, por tanto, reducir el efecto de los datos con ruido sobre el resultado proporcionado.

Las expresiones "datos no informativos", "bins no informativos" y "secciones genómicas no informativas", tal como se usan en el presente documento, se refieren a secciones genómicas o datos derivados de las mismas que tienen un valor numérico que es significativamente diferente de un valor umbral predeterminado o se encuentran fuera de un rango de valores de punto de corte predeterminado. Las expresiones "umbral" y "valor umbral" en el presente documento se refieren a cualquier número que se calcula usando un conjunto de datos calificados y sirve como límite de diagnóstico de una variación genética (por ejemplo, una variación en el número de copias, una aneuploidía, una aberración cromosómica y similares). Algunas veces se supera un umbral por los resultados obtenidos mediante los métodos descritos en el presente documento y a un sujeto se le diagnostica una variación genética (por ejemplo, trisomía 21). A menudo, se calcula un valor umbral o rango de valores al manipular matemática y/o estadísticamente los datos leídos de secuencias (por ejemplo, a partir de una referencia y/o sujeto), en algunas implementaciones, y en determinadas implementaciones, los datos leídos de secuencias manipulados para generar un valor umbral o rango de valores son datos leídos de secuencias (por ejemplo, a partir de una referencia y/o sujeto). En algunas implementaciones, se determina un valor de incertidumbre. Un valor de incertidumbre es generalmente una medida de varianza o error y puede ser cualquier medida de varianza o error adecuada. Un valor de incertidumbre puede ser una desviación estándar, error estándar, varianza calculada, valor de p o desviación media absoluta (D.M.A.), en algunas implementaciones. En algunas implementaciones un valor de incertidumbre puede calcularse según una fórmula en el ejemplo 6.

Puede usarse cualquier procedimiento adecuado para procesar conjuntos de datos descritos en el presente documento. Los ejemplos no limitativos de procedimientos adecuados para su uso para procesar conjuntos de datos incluyen filtrado, normalización, ponderación, monitorización de alturas de pico, monitorización de áreas de pico, monitorización de bordes de pico, determinación de razones de área, procesamiento matemático de datos, procesamiento estadístico de datos, aplicación de algoritmos estadísticos, análisis con variables fijas, análisis con variables optimizadas, representación

gráfica de datos para identificar patrones o tendencias para el procesamiento adicional, similares y combinaciones de los anteriores. En algunas implementaciones, los conjuntos de datos se procesan basándose en diversas características (por ejemplo, contenido de GC, lecturas mapeadas redundantes, regiones de centrómero, regiones de telómero, similares y combinaciones de los mismos) y/o variables (por ejemplo, sexo del feto, edad materna, ploidía materna, contribución porcentual de ácido nucleico fetal, similares o combinaciones de los mismos). En determinadas implementaciones, el procesamiento de conjuntos de datos tal como se describe en el presente documento puede reducir la complejidad y/o dimensionalidad de conjuntos de datos grandes y/o complejos. Un ejemplo no limitativo de un conjunto de datos complejos incluye datos de lectura de secuencia generados a partir de uno o más sujetos de prueba y una pluralidad de sujetos de referencia de diferentes edades y orígenes étnicos. En algunas implementaciones, los conjuntos de datos pueden incluir de miles a millones de lecturas de secuencia para cada sujeto de prueba y/o referencia.

El procesamiento de datos puede realizarse en cualquier número de etapas, en determinadas implementaciones. Por ejemplo, los datos pueden procesarse usando solamente un único procedimiento de procesamiento en algunas implementaciones, y en determinadas implementaciones los datos pueden procesarse usando 1 o más, 5 o más, 10 o más o 20 o más etapas de procesamiento (por ejemplo, 1 o más etapas de procesamiento, 2 o más etapas de procesamiento, 3 o más etapas de procesamiento, 4 o más etapas de procesamiento, 5 o más etapas de procesamiento, 6 o más etapas de procesamiento, 7 o más etapas de procesamiento, 8 o más etapas de procesamiento, 9 o más etapas de procesamiento, 10 o más etapas de procesamiento, 11 o más etapas de procesamiento, 12 o más etapas de procesamiento, 13 o más etapas de procesamiento, 14 o más etapas de procesamiento, 15 o más etapas de procesamiento, 16 o más etapas de procesamiento, 17 o más etapas de procesamiento, 18 o más etapas de procesamiento, 19 o más etapas de procesamiento o 20 o más etapas de procesamiento). En algunas implementaciones, las etapas de procesamiento pueden ser la misma etapa repetida dos o más veces (por ejemplo, filtrar dos o más veces, normalizar dos o más veces) y, en determinadas implementaciones, las etapas de procesamiento pueden ser dos o más etapas de procesamiento diferentes (por ejemplo, filtrado, normalización; normalización, monitorización de alturas y bordes de pico; filtrado, normalización, normalización con respecto a una referencia, manipulación estadística para determinar valores de p, y similares), llevadas a cabo simultánea o secuencialmente. En algunas implementaciones, puede usarse cualquier número y/o combinación adecuada de las mismas o diferentes etapas de procesamiento para procesar datos de lectura de secuencia para facilitar la provisión de un resultado. En determinadas implementaciones, el procesamiento de conjuntos de datos por los criterios descritos en el presente documento puede reducir la complejidad y/o dimensionalidad de un conjunto de datos.

En algunas implementaciones, una o más etapas de procesamiento pueden comprender una o más etapas de filtrado. El término "filtrar", tal como se usa en el presente documento, se refiere a eliminar las secciones o bins genómicos de la consideración. Los bins pueden seleccionarse para la eliminación basándose en cualquier criterio adecuado incluyendo, pero sin limitarse a, datos redundantes (por ejemplo, lecturas mapeadas redundantes o solapantes), datos no informativos (por ejemplo, bins con mediana de recuentos cero), bins con secuencias sobrerrepresentadas o subrepresentadas, datos con ruidos, similares o combinaciones de los anteriores. Un procedimiento de filtrado implica a menudo eliminar uno o más bins de la consideración y restar los recuentos en el uno o más bins seleccionados para eliminar de los recuentos contados o sumados para los bins, cromosoma o cromosomas o genoma en consideración. En algunas implementaciones, los bins pueden eliminarse sucesivamente (por ejemplo, uno cada vez para permitir la evaluación del efecto de eliminación de cada bin individual), y en determinadas implementaciones todos los bins marcados para eliminación pueden eliminarse al mismo tiempo. En algunas implementaciones, las secciones genómicas caracterizadas por una varianza por encima o por debajo de un determinado nivel se eliminan, lo que algunas veces se denominan en el presente documento filtración de secciones genómicas "con ruido". En determinadas implementaciones, un procedimiento de filtrado comprende obtener puntos de datos de un conjunto de datos que se desvían de la elevación de perfil media de una sección genómica, un cromosoma o segmento de un cromosoma en un múltiplo predeterminado de la varianza del perfil, y en determinadas implementaciones, un procedimiento de filtrado comprende eliminar puntos de datos de un conjunto de datos que no se desvían de la elevación de perfil media de una sección genómica, un cromosoma o segmento de un cromosoma en un múltiplo predeterminado de la varianza del perfil. En algunas implementaciones, se usa un procedimiento de filtrado para reducir el número de secciones genómicas candidatas analizadas para determinar la presencia o ausencia de una variación genética.

Reducir el número de secciones genómicas candidatas analizadas para determinar la presencia o ausencia de una variación genética (por ejemplo, microdeleción, microduplicación) reduce a menudo la complejidad y/o dimensionalidad de un conjunto de datos y, algunas veces, aumenta la velocidad de búsqueda y/o identificación de variaciones genéticas y/o aberraciones genéticas en dos o más órdenes de magnitud.

En algunas implementaciones, una o más etapas de procesamiento pueden comprender una o más etapas de normalización. La normalización puede realizarse mediante un método adecuado conocido en la técnica. Algunas veces, la normalización comprende ajustar los valores medidos en diferentes escalas a una escala teóricamente común. Algunas veces, la normalización comprende un ajuste matemático sofisticado para llevar a alineación distribuciones de probabilidad de valores ajustados. En algunos casos, la normalización comprende alinear distribuciones a una distribución normal. Algunas veces, la normalización comprende ajustes matemáticos que permiten la comparación de los valores normalizados correspondientes para diferentes conjuntos de datos, de manera que se eliminen los efectos de determinadas influencias macroscópicas (por ejemplo, error y anomalías). Algunas veces, la normalización comprende escalamiento. La normalización comprende a veces la división de uno o más conjuntos de datos por una variable o fórmula predeterminada. Los ejemplos no limitativos de métodos de normalización incluyen normalización por bins, normalización por contenido de GC, regresión lineal y no lineal por

mínimos cuadrados, LOESS, LOESS de GC, LOWESS (suavizado de diagrama de dispersión ponderado localmente), PERUN, enmascaramiento de repetición (RM), normalización de GC y enmascaramiento de repetición (GCRM), cQn y/o combinaciones de los mismos. En algunas implementaciones, la determinación de una presencia o ausencia de una variación genética (por ejemplo, una aneuploidía) utiliza un método de normalización (por ejemplo, normalización basada en bins, normalización por contenido de GC, regresión lineal y no lineal por mínimos cuadrados, LOESS, LOESS de GC, LOWESS (suavizado de diagrama de dispersión ponderado localmente), PERUN, enmascaramiento de repetición (RM), normalización de GC y enmascaramiento de repetición (GCRM), cQn, un método de normalización conocido en la técnica y/o una combinación de los mismos).

Por ejemplo, LOESS es un método de modelado por regresión conocido en la técnica que combina modelos de regresión múltiple en un metamodelo basado en k vecinos más cercanos. LOESS se denomina, algunas veces, regresión polinómica ponderada localmente. LOESS de GO, en algunas implementaciones, aplica un modelo de LOESS a la relación entre el recuento de fragmentos (por ejemplo, lecturas de secuencia, recuentos) y composición de GO para secciones genómicas. Representar gráficamente una curva suave a través de un conjunto de puntos de datos usando LOESS se denomina, algunas veces, curva de LOESS, particularmente cuando cada valor suavizado viene dado por una regresión cuadrática de mínimos cuadrados ponderados sobre el tramo de valores de la variable de criterio de diagrama de dispersión del eje y. Para cada punto en un conjunto de datos, el método de LOESS ajusta un polinomio de bajo grado a un subconjunto de los datos, con valores de variables explicativas cerca del punto cuya respuesta se estima. El polinomio se ajusta usando mínimos cuadrados ponderados, lo que da más peso a puntos cerca del punto cuya respuesta se estima y menos peso a puntos más lejos. Después, se obtiene el valor de la función de regresión para un punto mediante la evaluación del polinomio local usando los valores de variables explicativas para ese punto de datos. Algunas veces, el ajuste de LOESS se considera completo después de que se hayan calculado los valores de la función de regresión para cada uno de los puntos de datos. Muchos de los detalles de este método, tales como el grado del modelo polinómico y los pesos, son flexibles.

Puede usarse cualquier número adecuado de normalizaciones. En algunas implementaciones, los conjuntos de datos pueden normalizarse 1 o más, 5 o más, 10 o más o incluso 20 o más veces. Los conjuntos de datos pueden normalizarse a valores (por ejemplo, valor de normalización) representativos de cualquier característica o variable adecuada (por ejemplo, datos de muestra, datos de referencia, o ambos). Los ejemplos no limitativos de tipos de normalizaciones de datos que pueden usarse incluyen normalizar datos de recuento sin procesar para una o más secciones genómicas de referencia o de prueba seleccionadas con respecto al número total de recuentos mapeados en el cromosoma o en todo el genoma en el que se mapean la sección o secciones genómicas seleccionadas; normalizar datos de recuento sin procesar para una o más secciones genómicas seleccionadas con respecto a una mediana de recuento de referencia para una o más secciones genómicas o el cromosoma en el que se mapea una sección o segmentos genómicos seleccionados; normalizar datos de recuento sin procesar con respecto a datos normalizados previamente o derivados de los mismos; y normalizar los datos normalizados previamente con respecto a una o más de otras variables de normalización predeterminadas. Normalizar un conjunto de datos a veces tiene el efecto de aislar el error estadístico, dependiendo de la característica o propiedad seleccionada como la variable de normalización predeterminada. A veces, normalizar un conjunto de datos permite además comparar las características de datos de datos que tienen escalas diferentes, al llevar los datos a una escala común (por ejemplo, variable de normalización predeterminada). En algunas implementaciones, pueden usarse una o más normalizaciones con respecto a un valor derivado estadísticamente para minimizar las diferencias de los datos y disminuir la importancia de los datos atípicos. Normalizar secciones genómicas, o bins, con respecto a un valor de normalización a veces se denomina “normalización basada en bins”.

En determinadas implementaciones, una etapa de procesamiento que comprende la normalización incluye la normalización con respecto a una ventana estática y, en algunas implementaciones, una etapa de procesamiento que comprende normalización incluye la normalización con respecto a una ventana móvil o deslizante. El término “ventana”, tal como se usa en el presente documento, se refiere a una o más secciones genómicas elegidas para el análisis y, algunas veces, usadas como referencia para la comparación (por ejemplo, usadas para la normalización y/u otra manipulación matemática o estadística). La expresión “normalizar con respecto a una ventana estática”, tal como se usa en el presente documento, se refiere a un procedimiento de normalización que usa una o más secciones genómicas seleccionadas para la comparación entre un sujeto de prueba y el conjunto de datos del sujeto de referencia. En algunas implementaciones, las secciones genómicas seleccionadas se utilizan para generar un perfil. Una ventana estática incluye generalmente un conjunto predeterminado de secciones genómicas que no cambian durante las manipulaciones y/o el análisis. Las expresiones “normalizar con respecto a una ventana móvil” y “normalizar con respecto a una ventana deslizante”, tal como se usan en el presente documento, se refieren a normalizaciones realizadas a secciones genómicas localizadas en la región genómica (por ejemplo, alrededores genéticos inmediatos, sección o secciones genómicas adyacentes y similares) de una sección genómica de prueba seleccionada, en la que una o más secciones genómicas de prueba seleccionadas se normalizan con respecto a secciones genómicas que rodean inmediatamente la sección genómica de prueba seleccionada. En determinadas implementaciones, las secciones genómicas seleccionadas se utilizan para generar un perfil. Una normalización de ventana deslizante o móvil incluye a menudo mover o deslizar repetidamente hacia una sección genómica de prueba adyacente, y normalizar la sección genómica de prueba recién seleccionada con respecto a secciones genómicas que rodean inmediatamente o adyacentes a la sección genómica de prueba recién seleccionada, en la que las ventanas adyacentes tienen una o más secciones genómicas en común. En determinadas implementaciones, una pluralidad de secciones genómicas y/o cromosomas de prueba seleccionados pueden analizarse mediante un procedimiento de ventana deslizante.

En algunas implementaciones, la normalización con respecto a una ventana deslizante o móvil puede generar uno o más valores, en la que cada valor representa la normalización con respecto a un conjunto diferente de secciones genómicas de referencia seleccionadas de regiones diferentes de un genoma (por ejemplo, cromosoma). En determinadas implementaciones, el uno o más valores generados son sumas acumulativas (por ejemplo, una estimación numérica de la integral del perfil de recuento normalizado en la sección genómica seleccionada, dominio (por ejemplo, parte del cromosoma) o cromosoma). Los valores generados por el procedimiento de ventana deslizante o móvil pueden usarse para generar un perfil y facilitar que se llegue un resultado. En algunas implementaciones, las sumas acumulativas de una o más secciones genómicas pueden mostrarse en función de la posición genómica. El análisis de ventana móvil o deslizante a veces se usa para analizar un genoma para determinar la presencia o ausencia de microdeleciones y/o microinserciones. En determinadas implementaciones, la visualización de sumas acumulativas de una o más secciones genómicas se usa para identificar la presencia o ausencia de regiones de variación genética (por ejemplo, microdeleciones, microduplicaciones). En algunas implementaciones, el análisis de ventana móvil o deslizante se usa para identificar regiones genómicas que contienen microdeleciones y, en determinadas implementaciones, el análisis de ventana móvil o deslizante se usa para identificar regiones genómicas que contienen microduplicaciones.

Una metodología de normalización particularmente útil para reducir el error asociado con indicadores de ácido nucleico se denomina, en el presente documento, eliminación del error parametrizado y normalización no sesgada (PERUN). La metodología PERUN puede aplicarse a una gran variedad de indicadores de ácido nucleico (por ejemplo, lecturas de secuencia de ácido nucleico) con el propósito de reducir los efectos de error que confunden predicciones basadas en tales indicadores.

Por ejemplo, la metodología PERUN puede aplicarse a lecturas de secuencia de ácido nucleico de una muestra y reducir los efectos de error que pueden afectar a las determinaciones de elevación de ácido nucleico (por ejemplo, determinaciones de elevación de sección genómica). Tal aplicación es útil para usar lecturas de secuencia de ácido nucleico para evaluar la presencia o ausencia de una variación genética en un sujeto que se manifiesta como una elevación variable de una secuencia de nucleótidos (por ejemplo, sección genómica). Los ejemplos no limitativos de variaciones en las secciones genómicas son las aneuploidías cromosómicas (por ejemplo, trisomía 21, trisomía 18, trisomía 13) y la presencia o ausencia de un cromosoma sexual (por ejemplo, XX en mujeres frente a XY en hombres). Una trisomía de un autosoma (por ejemplo, un cromosoma distinto de un cromosoma sexual) puede denominarse autosoma afectado. Otros ejemplos no limitativos de variaciones en las elevaciones de sección genómica incluyen microdeleciones, microinserciones, duplicaciones y mosaicismo.

En determinadas aplicaciones, la metodología PERUN puede reducir el sesgo experimental mediante la normalización de indicadores de ácido nucleico para grupos genómicos particulares, denominándose estos últimos bins. Los bins incluyen una colección adecuada de indicadores de ácido nucleico, un ejemplo no limitativo de los mismos incluye una longitud de nucleótidos contiguos, a lo que se hace referencia en el presente documento como sección genómica o porción de un genoma de referencia. Los bins pueden incluir otros indicadores de ácido nucleico tal como se describe en el presente documento. En tales aplicaciones, la metodología PERUN normaliza generalmente los indicadores de ácido nucleico en bins particulares a través de varias muestras en tres dimensiones. Una descripción detallada de aplicaciones PERUN particulares se describe en el ejemplo 4 y el ejemplo 5 en el presente documento.

En determinadas implementaciones, la metodología PERUN incluye calcular una elevación de sección genómica para cada bin a partir de una relación ajustada entre (i) sesgo experimental para un bin de un genoma de referencia en el que se mapean las lecturas de secuencia y (ii) recuentos de lecturas de secuencia mapeadas en el bin. El sesgo experimental para cada uno de los bins puede determinarse a través de múltiples muestras según una relación ajustada para cada muestra entre (i) los recuentos de lecturas de secuencia asignadas a cada uno de los bins, y (ii) una característica de mapeo para cada uno de los bins. Esta relación ajustada para cada muestra puede ensamblarse para múltiples muestras en tres dimensiones. El conjunto puede ordenarse según el sesgo experimental en determinadas implementaciones (por ejemplo, Fig. 82, ejemplo 4), aunque la metodología PERUN puede ponerse en práctica sin ordenar el conjunto según el sesgo experimental.

Puede generarse una relación mediante un método adecuado conocido en la técnica. Puede generarse una relación en dos dimensiones para cada muestra en determinadas implementaciones, y puede seleccionarse una variable probatoria de error, o posiblemente probatoria de error, para una o más de las dimensiones. Una relación puede generarse, por ejemplo, usando un software de gráficos conocido en la técnica que representa gráficamente un gráfico usando valores de dos o más variables proporcionadas por un usuario. Puede ajustarse una relación utilizando un método conocido en la industria (por ejemplo, software de gráficos). Determinadas relaciones pueden ajustarse por regresión lineal, y la regresión lineal puede generar un valor de pendiente y un valor de ordenada en el origen. Determinadas relaciones a veces no son lineales y pueden ajustarse mediante una función no lineal, tal como una función parabólica, hiperbólica o exponencial, por ejemplo.

En la metodología PERUN, una o más de las relaciones ajustadas pueden ser lineales. Para un análisis del ácido nucleico circulante, libre de células de mujeres embarazadas, en el que el sesgo experimental es el sesgo de GC y la característica de mapeo es el contenido de GC, la relación ajustada para una muestra entre (i) los recuentos de lecturas de secuencia mapeadas en cada bin, y (ii) el contenido de GC para cada uno de los bins, puede ser lineal. Para esta última relación ajustada, la pendiente pertenece al sesgo de GC, y puede determinarse un coeficiente de sesgo de GC

para cada bin cuando las relaciones ajustadas se ensamblan a través de múltiples muestras. En tales implementaciones, la relación ajustada para múltiples muestras y un bin entre (i) el coeficiente de sesgo de GC para el bin, y (ii) los recuentos de lecturas de secuencia mapeadas en el bin, también puede ser lineal. Puede obtenerse una ordenada en el origen y pendiente a partir de esta última relación ajustada. En tales aplicaciones, la pendiente aborda el sesgo específico de muestra basándose en el contenido de GC y la ordenada en el origen aborda un patrón de atenuación específico de bin común a todas las muestras. La metodología PERUN puede reducir significativamente tal sesgo específico de muestra y atenuación específica de bin cuando se calculan las elevaciones de sección genómica para proporcionar un resultado (por ejemplo, presencia o ausencia de variación genética; determinación del sexo del feto).

Por tanto, la aplicación de la metodología PERUN a las lecturas de secuencia en múltiples muestras paralelas puede reducir significativamente el error provocado por (i) el sesgo experimental específico de muestra (por ejemplo, el sesgo de GC) y (ii) la atenuación específica de bin común a las muestras. Otros métodos en los cuales cada una de estas dos fuentes de error se abordan a menudo por separado o en serie no pueden reducirlas tan eficazmente como la metodología PERUN. Sin desear limitarse por la teoría, se espera que la metodología PERUN reduzca el error más eficazmente en parte porque sus procedimientos generalmente aditivos no aumentan la dispersión tanto como los procedimientos generalmente multiplicativos usados en otros enfoques de normalización (por ejemplo, LOESS de GC).

Pueden usarse técnicas de normalización y estadísticas adicionales en combinación con la metodología PERUN. Puede aplicarse un procedimiento adicional antes, después y/o durante el empleo de la metodología PERUN. Se describen más adelante en el presente documento ejemplos no limitativos de procedimientos que pueden usarse en combinación con la metodología PERUN.

En algunas implementaciones, puede usarse una normalización secundaria o ajuste de una elevación de sección genómica para el contenido de GC junto con la metodología PERUN. Puede usarse un procedimiento de normalización o ajuste del contenido de GC adecuado (por ejemplo, LOESS de GC, GCRM). En determinadas implementaciones, puede identificarse una muestra particular para la aplicación de un procedimiento de normalización de GC adicional. Por ejemplo, la aplicación de la metodología PERUN puede determinar el sesgo de GC para cada muestra, y una muestra asociada con un sesgo de GC por encima de determinado umbral puede seleccionarse para un procedimiento adicional de normalización de GC. En tales implementaciones, puede usarse una elevación umbral predeterminada para seleccionar tales muestras para la normalización de GC adicional.

En determinadas implementaciones, puede usarse un procedimiento de ponderación o filtrado de bins junto con la metodología PERUN. Puede usarse un procedimiento de ponderación o filtrado de bins adecuado y se describen en el presente documento ejemplos no limitativos. Los ejemplos 4 y 5 describen la utilización de medidas de error del factor R para el filtrado de bins.

Módulo de sesgo de GC

La determinación del sesgo de GC (por ejemplo, determinación del sesgo de GC para cada una de las porciones de un genoma de referencia (por ejemplo, secciones genómicas)) puede proporcionarse por un módulo de sesgo de GC (por ejemplo, por un aparato que comprende un módulo de sesgo de GC). En algunas implementaciones, se requiere un módulo de sesgo de GC para proporcionar una determinación de sesgo de GC. Algunas veces, un módulo de sesgo de GC proporciona una determinación de sesgo de GC a partir de una relación ajustada (por ejemplo, una relación lineal ajustada) entre los recuentos de lecturas de secuencia mapeadas en cada una de las porciones de un genoma de referencia y el contenido de GC de cada porción. Un aparato que comprende un módulo de sesgo de GC puede comprender al menos un procesador. En algunas implementaciones, los determinaciones de sesgo de GC (es decir, datos de sesgo de GC) se proporcionan por un aparato que incluye un procesador (por ejemplo, uno o más procesadores), procesador que puede realizar y/o implementar una o más instrucciones (por ejemplo, procesos, rutinas y/o subrutinas) del módulo de sesgo de GC. En algunas implementaciones, los datos de sesgo de GC se proporcionan por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunas implementaciones, un módulo de sesgo de GC funciona con uno o más procesadores externos (por ejemplo, una red interna o externa, un servidor, dispositivo de almacenamiento y/o una red de almacenamiento (por ejemplo, una nube)). En algunas implementaciones, los datos de sesgo de GC se proporcionan por un aparato que comprende uno o más de los siguientes: una o más celdas de flujo, una cámara, componentes de manipulación de fluido, una impresora, una pantalla de visualización (por ejemplo, un LED, LCT o CRT) y similares. Un módulo de sesgo de GC puede recibir información y/o datos de un aparato o módulo adecuado. Algunas veces, un módulo de sesgo de GC puede recibir información y/o datos de un módulo de secuenciación, un módulo de normalización, un módulo de ponderación, un módulo de mapeo o módulo de recuento. A veces, un módulo de sesgo de GC forma parte de un módulo de normalización (por ejemplo, módulo de normalización PERUN). Un módulo de sesgo de GC puede recibir lecturas de secuenciación de un módulo de secuenciación, lecturas de secuenciación mapeadas de un módulo de mapeo y/o recuentos de un módulo de recuento, en algunas implementaciones. A menudo, un módulo de sesgo de GC recibe información y/o datos de un aparato u otro módulo (por ejemplo, un módulo de recuento), transforma la información y/o los datos y proporciona información y/o datos de sesgo de GC (por ejemplo, una determinación de sesgo de GC, una relación ajustada lineal, y similares). La información y/o los datos de sesgo de GC pueden transferirse de un módulo de sesgo de GC a un módulo de nivel, módulo de filtrado, módulo de comparación, un módulo de normalización, un módulo de ponderación, un módulo de establecimiento de rango, un módulo de ajuste, un módulo de categorización y/o un módulo de resultados, en determinadas implementaciones.

Módulo de nivel

La determinación de los niveles (por ejemplo, elevaciones) y/o el cálculo de los niveles de sección genómica (por ejemplo, elevaciones de sección genómica) para porciones de un genoma de referencia puede proporcionarse por un módulo de nivel (por ejemplo, por un aparato que comprende un módulo de nivel). En algunas implementaciones, se requiere un módulo de nivel para proporcionar un nivel o un nivel de sección genómica calculado. Algunas veces, un módulo de nivel proporciona un nivel a partir de una relación ajustada (por ejemplo, una relación lineal ajustada) entre un sesgo de GC y recuentos de lecturas de secuencia mapeadas en cada una de las porciones de un genoma de referencia. Algunas veces, un módulo de nivel calcula un nivel de sección genómica como parte del PERUN. En algunas implementaciones, un módulo de nivel proporciona un nivel de sección genómica (es decir, L_i) según la ecuación $L_i = (m_i - G_iS) I^{-1}$ en el que G_i es el sesgo de GC, mi son recuentos medidos mapeados en cada porción de un genoma de referencia, i es una muestra, e I es la ordenada en el origen y S es la pendiente de la relación a ajustada (por ejemplo, una relación lineal ajustada) entre un sesgo de GC y recuentos de lecturas de secuencia mapeadas en cada una de las porciones de un genoma de referencia. Un aparato que comprende un módulo de nivel puede comprender al menos un procesador. En algunas implementaciones, se proporciona una determinación de nivel (es decir, datos de nivel) por un aparato que incluye un procesador (por ejemplo, uno o más procesadores), procesador que puede realizar y/o implementar una o más instrucciones (por ejemplo, procesos, rutinas y/o subrutinas) del módulo de recuento. En algunas implementaciones, los datos de nivel se proporcionan por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunas implementaciones, un módulo de nivel funciona con uno o más procesadores externos (por ejemplo, una red interna o externa, un servidor, dispositivo de almacenamiento y/o una red de almacenamiento (por ejemplo, una nube)). En algunas implementaciones, los datos de nivel se proporcionan por un aparato que comprende uno o más de los siguientes: una o más celdas de flujo, una cámara, componentes de manipulación de fluido, una impresora, una pantalla de visualización (por ejemplo, un LED, LCT o CRT) y similares. Un módulo de nivel puede recibir información y/o datos de un aparato o módulo adecuado. Algunas veces, un módulo de nivel puede recibir información y/o datos de un módulo de sesgo de GC, un módulo de secuenciación, un módulo de normalización, un módulo de ponderación, un módulo de mapeo o módulo de recuento. Un módulo de nivel puede recibir lecturas de secuenciación de un módulo de secuenciación, lecturas de secuenciación mapeadas de un módulo de mapeo y/o recuentos de un módulo de recuento, en algunas implementaciones. Algunas veces, un módulo de nivel forma parte de un módulo de normalización (por ejemplo, módulo de normalización PERUN). A menudo, un módulo de nivel recibe información y/o datos de un aparato u otro módulo (por ejemplo, un módulo de sesgo de GC), transforma la información y/o los datos y proporciona información y/o datos de nivel (por ejemplo, una determinación del nivel, una relación ajustada lineal, y similares). La información y/o los datos de nivel pueden transferirse de un módulo de nivel a un módulo de comparación, un módulo de normalización, un módulo de ponderación, un módulo de establecimiento de rango, un módulo de ajuste, un módulo de categorización, un módulo en un módulo de normalización y/o un módulo de resultados, en determinadas implementaciones.

Módulo de filtrado

El filtrado de secciones genómicas puede proporcionarse por un módulo de filtrado (por ejemplo, por un aparato que comprende un módulo de filtrado). En algunas implementaciones, se requiere un módulo de filtrado para proporcionar datos de sección genómica filtrados (por ejemplo, secciones genómicas filtradas) y/o para eliminar secciones genómicas de la consideración. Algunas veces, un módulo de filtrado elimina los recuentos mapeados en una sección genómica de la consideración. Algunas veces, un módulo de filtrado elimina los recuentos mapeados en una sección genómica de una determinación de una elevación o un perfil. Un módulo de filtrado puede filtrar los datos (por ejemplo, recuentos, recuentos mapeados en secciones genómicas, secciones genómicas, elevaciones de secciones genómicas, recuentos normalizados, recuentos sin procesar, y similares) mediante uno o más procedimientos de filtrado conocidos en la técnica o descritos en el presente documento. Un aparato que comprende un módulo de filtrado puede comprender al menos un procesador. En algunas implementaciones, los datos filtrados se proporcionan por un aparato que incluye un procesador (por ejemplo, uno o más procesadores), procesador que puede realizar y/o implementar una o más instrucciones (por ejemplo, procesos, rutinas y/o subrutinas) del módulo de filtrado. En algunas implementaciones, los datos filtrados se proporcionan por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunas implementaciones, un módulo de filtrado funciona con uno o más procesadores externos (por ejemplo, una red interna o externa, un servidor, dispositivo de almacenamiento y/o una red de almacenamiento (por ejemplo, una nube)). En algunas implementaciones, los datos filtrados se proporcionan por un aparato que comprende uno o más de los siguientes: una o más celdas de flujo, una cámara, componentes de manipulación de fluido, una impresora, una pantalla de visualización (por ejemplo, un LED, LCT o CRT) y similares. Un módulo de filtrado puede recibir información y/o datos de un aparato o módulo adecuado. Algunas veces, un módulo de filtrado puede recibir información y/o datos de un módulo de secuenciación, un módulo de normalización, un módulo de ponderación, un módulo de mapeo o módulo de recuento. Un módulo de filtrado puede recibir lecturas de secuenciación de un módulo de secuenciación, lecturas de secuenciación mapeadas de un módulo de mapeo y/o recuentos de un módulo de recuento, en algunas implementaciones. A menudo, un módulo de filtrado recibe información y/o datos de otro aparato o módulo, transforma la información y/o los datos y proporciona información y/o datos filtrados (por ejemplo, recuentos filtrados, valores filtrados, secciones genómicas filtradas y similares). La información y/o los datos filtrados pueden transferirse de un módulo de filtrado a un módulo de

comparación, un módulo de normalización, un módulo de ponderación, un módulo de establecimiento de rango, un módulo de ajuste, un módulo de categorización y/o un módulo de resultados, en determinadas implementaciones.

Módulo de ponderación

La ponderación de secciones genómicas puede proporcionarse por un módulo de ponderación (por ejemplo, por un aparato que comprende un módulo de ponderación). En algunas implementaciones, se requiere un módulo de ponderación para ponderar secciones genómicas y/o proporcionar valores de sección genómica ponderados. Un módulo de ponderación puede ponderar secciones genómicas mediante uno o más procedimientos de ponderación conocidos en la técnica o descritos en el presente documento. Un aparato que comprende un módulo de ponderación puede comprender al menos un procesador. En algunas implementaciones, las secciones genómicas ponderadas se proporcionan por un aparato que incluye un procesador (por ejemplo, uno o más procesadores), procesador que puede realizar y/o implementar una o más instrucciones (por ejemplo, procesos, rutinas y/o subrutinas) del módulo de ponderación. En algunas implementaciones, las secciones genómicas ponderadas se proporcionan por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunas implementaciones, un módulo de ponderación funciona con uno o más procesadores externos (por ejemplo, una red interna o externa, un servidor, dispositivo de almacenamiento y/o una red de almacenamiento (por ejemplo, una nube)). En algunas implementaciones, las secciones genómicas ponderadas se proporcionan por un aparato que comprende uno o más de los siguientes: una o más celdas de flujo, una cámara, componentes de manipulación de fluido, una impresora, una pantalla de visualización (por ejemplo, un LED, LCT o CRT) y similares. Un módulo de ponderación puede recibir información y/o datos de un aparato o módulo adecuado. Algunas veces, un módulo de ponderación puede recibir información y/o datos de un módulo de secuenciación, un módulo de normalización, un módulo de filtrado, un módulo de mapeo y/o un módulo de recuento. Un módulo de ponderación puede recibir lecturas de secuenciación de un módulo de secuenciación, lecturas de secuenciación mapeadas de un módulo de mapeo y/o recuentos de un módulo de recuento, en algunas implementaciones. En algunas implementaciones, un módulo de ponderación recibe información y/o datos de otro aparato o módulo, transforma la información y/o los datos y proporciona información y/o datos (por ejemplo, secciones genómicas ponderadas, valores ponderados y similares). La información y/o los datos de sección genómica ponderados pueden transferirse de un módulo de ponderación a un módulo de comparación, un módulo de normalización, un módulo de filtrado, un módulo de establecimiento de rango, un módulo de ajuste, un módulo de categorización y/o un módulo de resultados, en determinadas implementaciones.

En algunas implementaciones, se usa una técnica de normalización que reduce el error asociado con inserciones, duplicaciones y/o deleciones (por ejemplo, variaciones del número de copias materno y/o fetal) junto con la metodología PERUN.

Las elevaciones de sección genómica calculadas mediante la metodología PERUN pueden utilizarse directamente para proporcionar un resultado. En algunas implementaciones, las elevaciones de sección genómica pueden usarse directamente para proporcionar un resultado para las muestras en las cuales la fracción fetal es de aproximadamente el 2 % a aproximadamente el 6 % o más (por ejemplo, fracción fetal de aproximadamente el 4 % o más). Las elevaciones de sección genómica calculadas mediante la metodología PERUN, algunas veces, se procesan adicionalmente para proporcionar un resultado. En algunas implementaciones, las elevaciones de sección genómica calculadas están normalizadas. En determinadas implementaciones, la suma, media o mediana de elevaciones de sección genómica calculadas para una sección genómica de prueba (por ejemplo, el cromosoma 21) puede dividirse entre la suma, media o mediana de elevaciones de sección genómica calculadas para secciones genómicas distintas de la sección genómica de prueba (por ejemplo, autosomas distintos del cromosoma 21), para generar una elevación de sección genómica experimental. Una elevación de sección genómica experimental o una elevación de sección genómica sin procesar puede usarse como parte de un análisis de normalización, tal como el cálculo de una puntuación Z o valor de Z. Puede generarse una puntuación Z para una muestra restando una elevación de sección genómica esperada de una elevación de sección genómica experimental o elevación de sección genómica sin procesar y el valor resultante puede dividirse entre una desviación estándar para las muestras. Las puntuaciones Z resultantes pueden distribuirse para diferentes muestras y analizarse, o pueden relacionarse con otras variables, tales como fracción fetal y otras, y analizarse, para proporcionar un resultado, en determinadas implementaciones.

Tal como se indica en el presente documento, la metodología PERUN no se limita a la normalización según el sesgo de GC y el contenido de GC *per se*, y puede usarse para reducir el error asociado con otras fuentes de error. Un ejemplo no limitativo de una fuente de sesgo distinto del contenido de GC es la capacidad de mapeo. Cuando se abordan parámetros de normalización distintos del sesgo y el contenido de GC, una o más de las relaciones ajustadas pueden ser no lineales (por ejemplo, hiperbólica, exponencial). Cuando el sesgo experimental se determina a partir de una relación no lineal, por ejemplo, una estimación de la curvatura del sesgo experimental puede analizarse en algunas implementaciones.

La metodología PERUN puede aplicarse a una gran variedad de indicadores de ácido nucleico. Los ejemplos no limitativos de indicadores de ácido nucleico son lecturas de secuencia de ácido nucleico y elevaciones de ácido nucleico en una ubicación particular en un microalineamiento. Los ejemplos no limitativos de lecturas de secuencia incluyen aquellas obtenidas a partir de ADN circulante, libre de células, ARN circulante, libre de células, ADN celular y ARN celular. La metodología PERUN puede aplicarse a lecturas de secuencia mapeadas en secuencias de referencia adecuadas, tales

como ADN genómico de referencia, ARN celular de referencia (por ejemplo, transcriptoma) y porciones de los mismos (por ejemplo, parte(s) de un complemento genómico de transcriptoma de ADN o ARN, parte(s) de un cromosoma).

Por tanto, en determinadas implementaciones, un ácido nucleico celular (por ejemplo, ADN o ARN) puede servir como indicador de ácido nucleico. Las lecturas de ácidos nucleicos celulares mapeadas en porciones del genoma de referencia pueden normalizarse usando la metodología PERUN.

El ácido nucleico celular es, algunas veces, una asociación con una o más proteínas, y un agente que captura ácido nucleico asociado a proteínas puede usarse para enriquecer este último, en algunas implementaciones. Un agente en determinados casos es un anticuerpo o fragmento de anticuerpo que se une de manera específica a una proteína en asociación con ácido nucleico celular (por ejemplo, un anticuerpo que se une de manera específica a una proteína cromatina (por ejemplo, proteína histona)). Los procedimientos en los cuales se usa un anticuerpo o fragmento de anticuerpo para enriquecer el ácido nucleico celular unido a una proteína particular a veces se denominan procedimientos de inmunoprecipitación de cromatina (ChIP). El ácido nucleico enriquecido mediante ChIP es un ácido nucleico en asociación con una proteína celular, tal como ADN o ARN, por ejemplo. Las lecturas de ácido nucleico enriquecido mediante ChIP pueden obtenerse con el uso de tecnología conocida en la técnica. Las lecturas de ácido nucleico enriquecido mediante ChIP pueden mapearse en una o más porciones de un genoma de referencia, y los resultados pueden normalizarse usando la metodología PERUN para proporcionar un resultado.

Por tanto, en determinadas implementaciones, también se proporcionan métodos para calcular con sesgo reducido elevaciones de sección genómica para una muestra de prueba, que comprenden: (a) obtener recuentos de lecturas de secuencia mapeadas en bins de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico celular de una muestra de prueba obtenida mediante aislamiento de una proteína con la que estaba asociado el ácido nucleico; (b) determinar el sesgo experimental para cada uno de los bins a través de múltiples muestras a partir de una relación ajustada entre (i) los recuentos de las lecturas de secuencia mapeadas en cada uno de los bins, y (ii) una característica de mapeo para cada uno de los bins; y (c) calcular una elevación de sección genómica para cada uno de los bins a partir de una relación ajustada entre el sesgo experimental y los recuentos de las lecturas de secuencia mapeadas en cada uno de los bins, proporcionándose de ese modo elevaciones de sección genómica calculadas, mediante lo cual el sesgo en los recuentos de las lecturas de secuencia mapeadas en cada uno de los bins se reduce en las elevaciones de sección genómica calculadas.

En determinadas implementaciones, un ARN celular puede servir como indicadores de ácido nucleico. Las lecturas de ARN celular pueden mapearse en porciones de ARN de referencia y normalizar con el uso de la metodología PERUN para proporcionar un resultado. Las secuencias conocidas para ARN celular, denominado transcriptoma, o un segmento del mismo, pueden usarse como una referencia en la que pueden mapearse lecturas de ARN a partir de una muestra. Las lecturas de ARN de muestra pueden obtenerse usando tecnología conocida en la técnica. Los resultados de las lecturas de ARN mapeadas en una referencia pueden normalizarse con el uso de la metodología PERUN para proporcionar un resultado.

Por tanto, en algunas implementaciones, también se proporcionan métodos para calcular con sesgo reducido elevaciones de sección genómica para una muestra de prueba, que comprenden: (a) obtener recuentos de lecturas de secuencia mapeadas en bins de ARN de referencia (por ejemplo, transcriptoma de referencia o segmento(s) del mismo), lecturas de secuencia que son lecturas de ARN celular de una muestra de prueba; (b) determinar el sesgo experimental para cada uno de los bins a través de múltiples muestras a partir de una relación ajustada entre (i) los recuentos de las lecturas de secuencia mapeadas en cada uno de los bins, y (ii) una característica de mapeo para cada uno de los bins; y (c) calcular una elevación de sección genómica para cada uno de los bins a partir de una relación ajustada entre el sesgo experimental y los recuentos de las lecturas de secuencia mapeadas en cada uno de los bins, proporcionándose de ese modo elevaciones de sección genómica calculadas, mediante lo cual el sesgo en los recuentos de las lecturas de secuencia mapeadas en cada uno de los bins se reduce en las elevaciones de sección genómica calculadas.

En algunas implementaciones, los niveles de ácido nucleico de microalineamiento pueden servir como indicadores de ácido nucleico. Los niveles de ácido nucleico a través de las muestras para una dirección particular o ácido nucleico de hibridación en un alineamiento pueden analizarse con el uso de la metodología PERUN, normalizando de ese modo los indicadores de ácido nucleico proporcionados por el análisis de microalineamiento. De esta manera, una dirección particular o ácido nucleico de hibridación en una microalineamiento es análogo a un bin para lecturas de secuencia de ácido nucleico mapeadas, y puede usarse la metodología PERUN para normalizar los datos de microalineamiento para proporcionar un resultado mejorado.

Por tanto, en determinadas implementaciones se proporcionan métodos para reducir el error de nivel de ácido nucleico de microalineamiento para una muestra de prueba, que comprenden: (a) obtener niveles de ácido nucleico en una microalineamiento al que se ha asociado el ácido nucleico de muestra de prueba, microalineamiento que incluye un alineamiento de ácidos nucleicos de captura; (b) determinar el sesgo experimental para cada uno de los ácidos nucleicos de captura a través de múltiples muestras a partir de una relación ajustada entre (i) los niveles de ácido nucleico de muestra de prueba asociados con cada uno de los ácidos nucleicos de captura, y (ii) una característica de asociación para cada uno de los ácidos nucleicos de captura; y (c) calcular un nivel de ácido nucleico de muestra de prueba para cada uno de los ácidos nucleicos de captura a partir de una relación ajustada entre el sesgo experimental y los niveles de

ácido nucleico de muestra de prueba asociados con cada uno de los ácidos nucleicos de captura, proporcionándose de ese modo niveles calculados, mediante lo cual se reduce el sesgo en los niveles de ácido nucleico de muestra de prueba asociado con cada uno de los ácidos nucleicos de captura en los niveles calculados. La característica de asociación mencionada anteriormente puede ser cualquier característica correlacionada con la hibridación de un ácido nucleico de muestra de prueba con un ácido nucleico de captura que da lugar a, o puede dar lugar a, error en la determinación del nivel de ácido nucleico de muestra de prueba asociado con un ácido nucleico de captura.

Módulo de normalización

Los datos normalizados (por ejemplo, recuentos normalizados) pueden proporcionarse por un módulo de normalización (por ejemplo, por un aparato que comprende un módulo de normalización). En algunas implementaciones, se requiere un módulo de normalización para proporcionar datos normalizados (por ejemplo, recuentos normalizados) obtenidos a partir de lecturas de secuenciación. Un módulo de normalización puede normalizar datos (por ejemplo, recuentos, recuentos filtrados, recuentos sin procesar) mediante uno o más procedimientos de normalización conocidos en la técnica. Un aparato que comprende un módulo de normalización puede comprender al menos un procesador. En algunas implementaciones, los datos normalizados se proporcionan por un aparato que incluye un procesador (por ejemplo, uno o más procesadores), procesador que puede realizar y/o implementar una o más instrucciones (por ejemplo, procesos, rutinas y/o subrutinas) del módulo de recuento. En algunas implementaciones, los datos normalizados se proporcionan por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunas implementaciones, un módulo de normalización funciona con uno o más procesadores externos (por ejemplo, una red interna o externa, un servidor, dispositivo de almacenamiento y/o una red de almacenamiento (por ejemplo, una nube)). En algunas implementaciones, los datos normalizados se proporcionan por un aparato que comprende uno o más de los siguientes: una o más celdas de flujo, una cámara, componentes de manipulación de fluido, una impresora, una pantalla de visualización (por ejemplo, un LED, LCT o CRT) y similares. Un módulo de normalización puede recibir información y/o datos de un aparato o módulo adecuado. Algunas veces, un módulo de normalización puede recibir información y/o datos de un módulo de secuenciación, un módulo de normalización, un módulo de mapeo o módulo de recuento. Un módulo de normalización puede recibir lecturas de secuenciación de un módulo de secuenciación, lecturas de secuenciación mapeadas de un módulo de mapeo y/o recuentos de un módulo de recuento, en algunas implementaciones. A menudo, un módulo de normalización recibe información y/o datos de otro aparato o módulo, transforma la información y/o los datos y proporciona información y/o datos normalizados (por ejemplo, recuentos normalizados, valores normalizados, valores de referencia normalizados (NRV) y similares). La información y/o los datos normalizados pueden transferirse de un módulo de normalización a un módulo de comparación, un módulo de normalización, un módulo de establecimiento de rango, un módulo de ajuste, un módulo de categorización y/o un módulo de resultados, en determinadas implementaciones. Algunas veces, los recuentos normalizados (por ejemplo, recuentos mapeados normalizados) se transfieren a un módulo de representación esperado y/o a un módulo de representación experimental de un módulo de normalización.

En algunas implementaciones, una etapa de procesamiento comprende una ponderación. Las expresiones "ponderado", "ponderación" o "función de ponderación" o derivados gramaticales o equivalentes de las mismas, tal como se usan en el presente documento, se refieren a una manipulación matemática de una porción o la totalidad de un conjunto de datos utilizado a veces para alterar la influencia de determinadas características o variables del conjunto de datos con respecto a otras características o variables del conjunto de datos (por ejemplo, aumentar o disminuir la significación y/o contribución de los datos contenidos en una o más secciones o bins genómicos, basándose en la calidad o utilidad de los datos en el bin o bins seleccionados). Una función de ponderación puede usarse para aumentar la influencia de los datos con una varianza de medición relativamente pequeña, y/o para disminuir la influencia de los datos con una varianza de medición relativamente grande, en algunas implementaciones. Por ejemplo, los bins con datos de secuencia de baja calidad o subrepresentados pueden "ponderarse por disminución" para minimizar la influencia en un conjunto de datos, mientras que los bins seleccionados pueden "ponderarse por aumento" para aumentar la influencia en un conjunto de datos. Un ejemplo no limitativo de una función de ponderación es [1 */* (desviación estándar)^2]. A veces, una etapa de ponderación se realiza de una manera sustancialmente similar a una etapa de normalización. En algunas implementaciones, un conjunto de datos se divide entre una variable predeterminada (por ejemplo, variable de ponderación). A menudo, se selecciona una variable predeterminada (por ejemplo, función objetivo minimizada, Phi) para ponderar distintas partes de un conjunto de datos de manera diferente (por ejemplo, aumentar la influencia de determinados tipos de datos mientras se reduce la influencia de otros tipos de datos).

En determinadas implementaciones, una etapa de procesamiento puede comprender una o más manipulaciones matemáticas y/o estadísticas. Cualquier manipulación matemática y/o estadística adecuada, sola o en combinación, puede usarse para analizar y/o manipular un conjunto de datos descrito en el presente documento. Puede usarse cualquier número adecuado de manipulaciones matemáticas y/o estadísticas. En algunas implementaciones, un conjunto de datos puede manipularse matemática y/o estadísticamente 1 o más, 5 o más, 10 o más o incluso 20 o más veces. Los ejemplos no limitativos de manipulaciones matemáticas y estadísticas que pueden usarse incluyen suma, resta, multiplicación, división, funciones algebraicas, estimadores de mínimos cuadrados, ajuste de curvas, ecuaciones diferenciales, polinomios racionales, polinomios dobles, polinomios ortogonales, puntuaciones z, valores de p, valores de chi, valores de phi, análisis de elevaciones de pico, determinación de ubicaciones de bordes de pico, cálculo de razones de áreas de pico, análisis de la mediana de elevación cromosómica, cálculo de la desviación media absoluta, suma de residuos al cuadrado, media, desviación estándar, error estándar, similares o combinaciones de los

mismos. Puede realizarse una manipulación matemática y/o estadística en todos o en una parte de los datos de lectura de secuencia o productos procesados de los mismos. Los ejemplos no limitativos de variables o características del conjunto de datos que pueden manipularse estadísticamente incluyen recuentos sin procesar, recuentos filtrados, recuentos normalizados, alturas de pico, anchuras de pico, áreas de pico, bordes de pico, tolerancias laterales, valores de P, mediana de elevaciones, elevaciones medias, distribución de recuentos dentro de una región genómica, representación relativa de especies de ácido nucleico, similares o combinaciones de los mismos.

En algunas implementaciones, una etapa de procesamiento puede incluir el uso de uno o más algoritmos estadísticos. Cualquier algoritmo estadístico adecuado, solo o en combinación, puede usarse para analizar y/o manipular un conjunto de datos descrito en el presente documento. Puede usarse cualquier número adecuado de algoritmos estadísticos. En algunas implementaciones, un conjunto de datos puede analizarse usando 1 o más, 5 o más, 10 o más o incluso 20 o más algoritmos estadísticos. Los ejemplos no limitativos de algoritmos estadísticos adecuados para su uso con los métodos descritos en el presente documento incluyen árboles de decisión, valores contranulos, comparaciones múltiples, prueba ómnibus, problema de Behrens-Fisher, remuestreo de tipo *bootstrapping*, método de Fisher para combinar pruebas independientes de significación, hipótesis nula, error tipo I, error tipo II, prueba exacta, prueba Z de una muestra, prueba Z de dos muestras, prueba de la t de una muestra, prueba de la t para datos emparejados, prueba de la t agrupada de dos muestras que tienen varianzas iguales, prueba de la t no agrupada de dos muestras que tienen varianzas desiguales, prueba z de una proporción, prueba z de dos proporciones agrupadas, prueba z de dos proporciones no agrupadas, prueba de chi cuadrado de una muestra, prueba F de dos muestras para determinar la igualdad de varianzas, intervalo de confianza, intervalo creíble, significación, metaanálisis, regresión lineal simple, regresión lineal robusta, similares o combinaciones de los anteriores. Los ejemplos no limitativos de variables o características del conjunto de datos que pueden analizarse usando algoritmos estadísticos incluyen recuentos sin procesar, recuentos filtrados, recuentos normalizados, alturas de pico, anchuras de pico, bordes de pico, tolerancias laterales, valores de p, mediana de elevaciones, elevaciones medias, distribución de recuentos dentro de una región genómica, representación relativa de especies de ácido nucleico, similares o combinaciones de los mismos.

En determinadas implementaciones, un conjunto de datos puede analizarse utilizando algoritmos estadísticos múltiples (por ejemplo, 2 o más) (por ejemplo, regresión por mínimos cuadrados, análisis de componentes principales, análisis discriminante lineal, análisis discriminante cuadrático, agregación de tipo *bootstrap*, redes neurales, modelos de máquinas de vectores de soporte, bosques aleatorios, modelos de árboles de clasificación, K vecinos más cercanos, regresión logística y/o suavizado de pérdida) y/o manipulaciones matemáticas y/o estadísticas (por ejemplo, a las que se hace referencia en el presente documento como manipulaciones). El uso de múltiples manipulaciones puede generar un espacio N-dimensional que puede usarse para proporcionar un resultado, en algunas implementaciones. En determinadas implementaciones, el análisis de un conjunto de datos utilizando múltiples manipulaciones puede reducir la complejidad y/o dimensionalidad del conjunto de datos. Por ejemplo, el uso de múltiples manipulaciones en un conjunto de datos de referencia puede generar un espacio N-dimensional (por ejemplo, gráfico de probabilidad) que puede usarse para representar la presencia o ausencia de una variación genética, dependiendo del estado genético de las muestras de referencia (por ejemplo, positivo o negativo para una variación genética seleccionada). El análisis de las muestras de prueba usando un conjunto sustancialmente similar de manipulaciones puede usarse para generar un punto N-dimensional para cada una de las muestras de prueba. A veces, la complejidad y/o dimensionalidad de un conjunto de datos de un sujeto de prueba se reduce a un solo valor o punto N-dimensional que puede compararse fácilmente con el espacio N-dimensional generado a partir de los datos de referencia. Los datos de muestra de prueba que se encuentran dentro del espacio N-dimensional poblado por los datos del sujeto de referencia son indicativos de un estado genético prácticamente similar al de los sujetos de referencia. Los datos de muestra de prueba que se encuentran fuera del espacio N-dimensional poblado por los datos del sujeto de referencia son indicativos de un estado genético sustancialmente diferente al de los sujetos de referencia. En algunas implementaciones, las referencias son euploides o no tienen de cualquier otra manera una variación genética o afección médica.

Después de que los conjuntos de datos se han contado, opcionalmente filtrado y normalizado, los conjuntos de datos procesados pueden manipularse adicionalmente mediante uno o más procedimientos de filtrado y/o normalización, en algunas implementaciones. Un conjunto de datos que se ha manipulado adicionalmente mediante uno o más procedimientos de filtrado y/o normalización puede usarse para generar un perfil, en determinadas implementaciones. El uno o más procedimientos de filtrado y/o normalización a veces pueden reducir la complejidad y/o dimensionalidad del conjunto de datos, en algunas implementaciones. Puede proporcionarse un resultado basado en un conjunto de datos de complejidad y/o dimensionalidad reducidas.

Se proporciona ejemplos no limitativos de filtrado de sección genómica en el presente documento en el ejemplo 4 con respecto a los métodos PERUN. Las secciones genómicas pueden filtrarse basándose en, o basándose en parte en, una medida de error. Una medida de error que comprende valores absolutos de desviación, tal como un factor R, puede usarse para la eliminación o ponderación de la sección genómica en determinadas implementaciones. Un factor R, en algunas implementaciones, se define como la suma de las desviaciones absolutas de los valores de recuento predichos de las mediciones reales divididas entre los valores de recuento predichos de las mediciones reales (por ejemplo, la ecuación B en el presente documento). Aunque puede usarse una medida de error que comprende valores absolutos de desviación, puede usarse alternativamente una medida de error adecuada. En determinadas implementaciones puede usarse una medida de error que no comprende

valores absolutos de desviación, tal como una dispersión basada en cuadrados. En algunas implementaciones, las secciones genómicas se filtran o ponderan según una medida de capacidad de mapeo (por ejemplo, una puntuación de capacidad de mapeo; ejemplo 5). Algunas veces, una sección genómica se filtra o se pondera según un número relativamente bajo de lecturas de secuencia mapeadas en la sección genómica (por ejemplo, 0, 1, 2, 3, 4, 5 lecturas mapeadas en la sección genómica). Las secciones genómicas pueden filtrarse o ponderarse según el tipo de análisis que se realiza. Por ejemplo, para el análisis de aneuploidía del cromosoma 13, 18 y/o 21, los cromosomas sexuales pueden filtrarse, y pueden analizarse solo autosomas, o un subconjunto de autosomas.

En implementaciones particulares, puede emplearse el siguiente procedimiento de filtrado. Se seleccionan el mismo conjunto de secciones genómicas (por ejemplo, bins) dentro de un cromosoma dado (por ejemplo, el cromosoma 21) y se comparan el número de lecturas en muestras afectadas y no afectadas. La brecha relaciona muestras con trisomía 21 y euploides e involucra un conjunto de secciones genómicas que cubren la mayor parte del cromosoma 21. El conjunto de secciones genómicas es el mismo entre las muestras euploides y T21. La distinción entre un conjunto de secciones genómicas y una sola sección no es crucial, ya que puede definirse una sección genómica. La misma región genómica se compara en diferentes pacientes. Este procedimiento puede utilizarse para un análisis de trisomía, tal como para T13 o T18 además de, o en lugar de, T21.

Después de que los conjuntos de datos se han contado, opcionalmente filtrado y normalizado, los conjuntos de datos procesados pueden manipularse mediante ponderación, en algunas implementaciones. Una o más secciones genómicas pueden seleccionarse para la ponderación para reducir la influencia de los datos (por ejemplo, datos con ruido, datos no informativos) contenidos en las secciones genómicas seleccionadas, en determinadas implementaciones, y en algunas implementaciones, una o más secciones genómicas pueden seleccionarse para la ponderación para mejorar o aumentar la influencia de los datos (por ejemplo, datos con pequeña varianza medida) contenidos en las secciones genómicas seleccionadas. En algunas implementaciones, se pondera un conjunto de datos usando una única función de ponderación que disminuye la influencia de los datos con grandes varianzas y aumenta la influencia de los datos con pequeñas varianzas. A veces se usa una función de ponderación para reducir la influencia de los datos con grandes varianzas y aumentar la influencia de los datos con pequeñas varianzas (por ejemplo, [1/(desviación estándar)^2]). En algunas implementaciones, se genera un gráfico de perfiles de datos procesados manipulados adicionalmente mediante ponderación para facilitar la clasificación y/o proporcionar un resultado. Puede proporcionarse un resultado basado en un gráfico de perfiles de datos ponderados

El filtrado o la ponderación de secciones genómicas puede realizarse en uno o más puntos adecuados en un análisis. Por ejemplo, las secciones genómicas pueden filtrarse o ponderarse antes o después de mapear las lecturas de secuencia en porciones de un genoma de referencia. Las secciones genómicas pueden filtrarse o ponderarse antes o después de determinar un sesgo experimental para las porciones individuales del genoma en algunas implementaciones. En determinadas implementaciones, las secciones genómicas pueden filtrarse o ponderarse antes o después de calcular las elevaciones de sección genómica.

Después de que los conjuntos de datos se han contado, opcionalmente filtrado, normalizado y, opcionalmente, ponderado, los conjuntos de datos procesados pueden manipularse mediante una o más manipulaciones matemáticas y/o estadísticas (por ejemplo, funciones estadísticas o algoritmo estadístico), en algunas implementaciones. En determinadas implementaciones, los conjuntos de datos procesados pueden manipularse adicionalmente mediante el cálculo de puntuaciones Z para una o más secciones genómicas, cromosomas o porciones de cromosomas seleccionados. En algunas implementaciones, los conjuntos de datos procesados pueden manipularse adicionalmente mediante el cálculo de valores de p. Las fórmulas para calcular las puntuaciones Z y los valores de p se presentan en el ejemplo 1. En determinadas implementaciones, las manipulaciones matemáticas y/o estadísticas incluyen una o más suposiciones que pertenecen a ploidía y/o fracción fetal. En algunas implementaciones, se genera un gráfico de perfiles de datos procesados manipulados adicionalmente mediante una o más manipulaciones estadísticas y/o matemáticas para facilitar la clasificación y/o proporcionar un resultado. Puede proporcionarse un resultado basado en un gráfico de perfiles de datos manipulados estadística y/o matemáticamente. Un resultado proporcionado basado en un gráfico de perfiles de datos manipulados estadística y/o matemáticamente incluye a menudo una o más suposiciones que pertenecen a ploidía y/o fracción fetal.

En determinadas implementaciones, se realizan múltiples manipulaciones en conjuntos de datos procesados para generar un espacio N-dimensional y/o un punto N-dimensional, después de que los conjuntos de datos se han contado, opcionalmente, filtrado y normalizado. Puede proporcionarse un resultado basado en un gráfico de perfiles de conjuntos de datos analizados en N dimensiones.

En algunas implementaciones, los conjuntos de datos se procesan usando uno o más análisis de elevación de pico, análisis de anchura de pico, análisis de ubicación de borde de pico, tolerancias laterales de pico, similares, derivaciones de los mismos o combinaciones de los anteriores, como parte de o después de procesar y/o manipular los conjuntos de datos. En algunas implementaciones, se genera un gráfico de perfiles de datos procesados usando uno o más análisis de elevación de pico, análisis de anchura de pico, análisis de ubicación de borde de pico, tolerancias laterales de pico, similares, derivaciones de los mismos o combinaciones de los anteriores para facilitar la clasificación y/o proporcionar un resultado. Puede proporcionarse un resultado basado en un gráfico de perfiles de datos que se procesaron usando uno o más análisis de elevación de pico, análisis de anchura de pico, análisis de

ubicación de borde de pico, tolerancias laterales de pico, similares, derivaciones de los mismos o combinaciones de los anteriores.

En algunas implementaciones, el uso de una o más muestras de referencia conocidas por estar libres de una variación genética en cuestión puede usarse para generar una mediana de perfil de recuento de referencia, que puede dar como resultado un valor predeterminado representativo de la ausencia de la variación genética, y a menudo se desvía de un valor predeterminado en las áreas correspondientes a la ubicación genómica en la cual la variación genética está ubicada en el sujeto de prueba, si el sujeto de prueba presentase la variación genética. En los sujetos de prueba en riesgo de, o que padecen, una afección médica asociada con una variación genética, se espera que el valor numérico para la sección o secciones genómicas seleccionadas varíe significativamente con respecto al valor predeterminado para ubicaciones genómicas no afectadas. En determinadas implementaciones, el uso de una o más muestras de referencia que se sabe que portan la variación genética en cuestión puede usarse para generar una mediana de perfil de recuento de referencia, lo que puede dar como resultado un valor predeterminado representativo de la presencia de la variación genética y a menudo se desvía de un valor predeterminado en las áreas correspondientes a la ubicación genómica en la que un sujeto de prueba no porta la variación genética. En sujetos de prueba que no están en riesgo de o que padecen una afección médica asociada con una variación genética, se espera que el valor numérico para la sección o secciones genómicas seleccionadas varíe significativamente con respecto al valor predeterminado para las ubicaciones genómicas afectadas.

En algunas implementaciones, el análisis y procesamiento de datos pueden incluir el uso de una o más suposiciones. Puede usarse un número o tipo adecuado de suposiciones para analizar o procesar un conjunto de datos. Los ejemplos no limitativos de suposiciones que pueden usarse para el procesamiento y/o análisis de datos incluyen ploidía materna, contribución fetal, prevalencia de determinadas secuencias en una población de referencia, origen étnico, prevalencia de una afección médica seleccionada en miembros de la familia emparentados, paralelismo entre perfiles de recuento sin procesar de diferentes pacientes y/o ejecuciones después de la normalización de GC y enmascaramiento de repetición (por ejemplo, GCRM), los apareamientos idénticos representan artefactos de PCR (por ejemplo, posición de base idéntica), las suposiciones inherentes en un ensayo cuantificador fetal (por ejemplo, FQA), las suposiciones con respecto a gemelos (por ejemplo, si hay 2 gemelos y solo 1 se ve afectado, la fracción fetal efectiva es solo el 50 % de la fracción fetal total medida (de manera similar para tripletes, cuadrupletes y similares)), ADN fetal libre de células (por ejemplo, ADNlc) cubre uniformemente todo el genoma, similares y combinaciones de los mismos.

En los casos en los que la calidad y/o profundidad de las lecturas de secuencia mapeadas no permite una predicción de resultados de la presencia o ausencia de una variación genética a un nivel de confianza deseado (por ejemplo, del 95 % o mayor nivel de confianza), basándose en los perfiles de recuento normalizados, pueden usarse uno o más algoritmos de manipulación matemática y/o algoritmos de predicción estadística adicionales, para generar valores numéricos adicionales útiles para el análisis de datos y/o proporcionar un resultado. La expresión ‘‘perfil de recuento normalizado’’, tal como se usa en el presente documento, se refiere a un perfil generado usando recuentos normalizados. Los ejemplos de métodos que pueden usarse para generar recuentos normalizados y perfiles de recuento normalizados se describen en el presente documento. Tal como se indicó, las lecturas de secuencia mapeadas que se han contado pueden normalizarse con respecto a los recuentos de muestras de prueba o recuentos de muestras de referencia. En algunas implementaciones, un perfil de recuento normalizado puede presentarse como una representación gráfica.

Perfiles

En algunas implementaciones, una etapa de procesamiento puede comprender generar uno o más perfiles (por ejemplo, gráfico de perfiles) a partir de diversos aspectos de un conjunto de datos o derivación del mismo (por ejemplo, producto de una o más etapas de procesamiento de datos matemáticas y/o estadísticas conocidas en la técnica y/o descritas en el presente documento). El término ‘‘perfil’’, tal como se usa en el presente documento, se refiere a un producto de una manipulación matemática y/o estadística de datos que puede facilitar la identificación de patrones y/o correlaciones en grandes cantidades de datos. Un ‘‘perfil’’ incluye a menudo valores resultantes de una o más manipulaciones de datos o conjuntos de datos, basándose en uno o más criterios. Un perfil incluye a menudo múltiples puntos de datos. Cualquier número adecuado de puntos de datos puede incluirse en un perfil dependiendo de la naturaleza y/o complejidad de un conjunto de datos. En determinadas implementaciones, los perfiles pueden incluir 2 o más puntos de datos, 3 o más puntos de datos, 5 o más puntos de datos, 10 o más puntos de datos, 24 o más puntos de datos, 25 o más puntos de datos, 50 o más puntos de datos, 100 o más puntos de datos, 500 o más puntos de datos, 1000 o más puntos de datos, 5000 o más puntos de datos, 10.000 o más puntos de datos o 100.000 o más puntos de datos.

En algunas implementaciones, un perfil es representativo de la totalidad de un conjunto de datos y, en determinadas implementaciones, un perfil es representativo de una porción o subconjunto de un conjunto de datos. Es decir, un perfil a veces incluye o se genera a partir de puntos de datos representativos de datos que no se han filtrado para eliminar cualquier dato y, a veces, un perfil incluye o se genera a partir de puntos de datos representativos de datos que se han filtrado para eliminar datos no deseados. En algunas implementaciones, un punto de datos en un perfil representa los resultados de la manipulación de datos para una sección genómica. En determinadas implementaciones, un punto de datos en un perfil incluye resultados de la manipulación de datos para grupos de secciones genómicas. En algunas implementaciones, los grupos de secciones genómicas pueden ser adyacentes entre sí y, en determinadas implementaciones, los grupos de secciones genómicas pueden ser de diferentes partes de un cromosoma o genoma.

Los puntos de datos en un perfil derivado de un conjunto de datos pueden ser representativos de cualquier categorización de datos adecuada. Los ejemplos no limitativos de categorías en las que los datos pueden agruparse para generar puntos de datos de perfil incluyen: secciones genómicas basadas en el tamaño, secciones genómicas basadas en características de secuencia (por ejemplo, contenido de GC, contenido de AT, posición en un cromosoma (por ejemplo, brazo corto, brazo largo, centrómero, telómero) y similares), niveles de expresión, cromosoma, similares o combinaciones de los mismos. En algunas implementaciones, puede generarse un perfil a partir de puntos de datos obtenidos de otro perfil (por ejemplo, perfil de datos normalizado renormalizado a un valor de normalización diferente para generar un perfil de datos renormalizado). En determinadas implementaciones, un perfil generado a partir de los puntos de datos obtenidos a partir de otro perfil reduce el número de puntos de datos y/o la complejidad del conjunto de datos. Reducir el número de puntos de datos y/o la complejidad de un conjunto de datos facilita a menudo la interpretación de los datos y/o facilita proporcionar un resultado.

Un perfil es a menudo una colección de recuentos normalizados o no normalizados para dos o más secciones genómicas. Un perfil incluye a menudo al menos una elevación y comprende a menudo dos o más elevaciones (por ejemplo, un perfil tiene a menudo múltiples elevaciones). Una elevación es generalmente para un conjunto de secciones genómicas que tienen aproximadamente los mismos recuentos o recuentos normalizados. Las elevaciones se describen con mayor detalle en el presente documento. En algunos casos, un perfil comprende una o más secciones genómicas, secciones genómicas que pueden ponderarse, eliminarse, filtrarse, normalizarse, ajustarse, promediarse, derivarse como una media, sumarse, restarse, procesarse o transformarse mediante cualquier combinación de los mismos. Un perfil comprende a menudo recuentos normalizados mapeados en secciones genómicas que definen dos o más elevaciones, en el que los recuentos se normalizan además según una de las elevaciones mediante un método adecuado. A menudo, los recuentos de un perfil (por ejemplo, una elevación de perfil) se asocian con un valor de incertidumbre.

Un perfil que comprende una o más elevaciones puede incluir una primera elevación y una segunda elevación. Algunas veces, una primera elevación es diferente (por ejemplo, significativamente diferente) de una segunda elevación. En algunas implementaciones, una primera elevación comprende un primer conjunto de secciones genómicas, una segunda elevación comprende un segundo conjunto de secciones genómicas y el primer conjunto de secciones genómicas no es un subconjunto del segundo conjunto de secciones genómicas. En algunos casos, un primer conjunto de secciones genómicas es diferente de un segundo conjunto de secciones genómicas a partir de las cuales se determina una primera y una segunda elevación. A veces un perfil puede tener múltiples primeras elevaciones que son diferentes (por ejemplo, significativamente diferentes, por ejemplo, tienen un valor significativamente diferente) que una segunda elevación dentro del perfil. A veces, un perfil comprende una o más primeras elevaciones que son significativamente diferentes de una segunda elevación dentro del perfil y una o más de las primeras elevaciones se ajustan. A veces, un perfil comprende una o más primeras elevaciones que son significativamente diferentes de una segunda elevación dentro del perfil, cada una de las primeras elevaciones comprenden una variación del número de copias materno, una variación del número de copias fetal, o una variación del número de copias materno y una variación del número de copias fetal y una o más de las primeras elevaciones se ajustan. Algunas veces, una primera elevación dentro de un perfil se elimina del perfil o se ajusta (por ejemplo, rellena). Un perfil puede comprender múltiples elevaciones que incluyen una o más primeras elevaciones significativamente diferentes de una o más segundas elevaciones y a menudo la mayoría de las elevaciones en un perfil son segundas elevaciones, segundas elevaciones que son aproximadamente iguales entre sí. Algunas veces más del 50 %, más del 60 %, más del 70 %, más del 80 %, más del 90 % o más del 95 % de las elevaciones en un perfil son segundas elevaciones.

Algunas veces, un perfil se visualiza como una representación gráfica. Por ejemplo, pueden representarse gráficamente y visualizarse una o más elevaciones que representan recuentos (por ejemplo, recuentos normalizados) de secciones genómicas. Los ejemplos no limitativos de gráficos de perfiles que pueden generarse incluyen el recuento sin procesar (por ejemplo, perfil de recuento sin procesar o perfil sin procesar), recuento normalizado, ponderado en bins, puntuación z, valor de p, razón de área frente a ploidía ajustada, mediana de elevación frente a razón entre la fracción fetal ajustada y medida, componentes principales, similares o combinaciones de los mismos. Los gráficos de perfiles permiten la visualización de los datos manipulados, en algunas implementaciones. En determinadas implementaciones, puede usarse un gráfico de perfiles para proporcionar un resultado (por ejemplo, razón de área frente a ploidía ajustada, mediana de elevación frente a razón entre fracción fetal ajustada y medida, componentes principales). Las expresiones "gráfico de perfiles de recuento sin procesar" o "gráfico de perfiles sin procesar", tal como se usan en el presente documento, se refieren a una representación gráfica de recuentos en cada sección genómica en una región normalizada para recuentos totales en una región (por ejemplo, genoma, sección genómica, cromosoma, bins cromosómicos o un segmento de un cromosoma). En algunas implementaciones, puede generarse un perfil usando un procedimiento de ventana estática y, en determinadas implementaciones, puede generarse un perfil usando un procedimiento de ventana deslizante.

A veces, un perfil generado para un sujeto de prueba se compara con un perfil generado para uno o más sujetos de referencia, para facilitar la interpretación de manipulaciones matemáticas y/o estadísticas de un conjunto de datos y/o para proporcionar un resultado. En algunas implementaciones, se genera un perfil basándose en una o más suposiciones iniciales (por ejemplo, contribución materna de ácido nucleico (por ejemplo, fracción materna), contribución fetal de ácido nucleico (por ejemplo, fracción fetal), ploidía de la muestra de referencia, similares o combinaciones de los mismos). En

determinadas implementaciones, un perfil de prueba se centra a menudo alrededor de un valor predeterminado representativo de la ausencia de una variación genética y se desvía a menudo de un valor predeterminado en áreas correspondientes a la ubicación genómica en la que está ubicada la variación genética en el sujeto de prueba, si el sujeto de prueba presentase la variación genética. En los sujetos de prueba en riesgo de, o que padecen, una afección médica asociada con una variación genética, se espera que el valor numérico para una sección genómica seleccionada varíe significativamente con respecto al valor predeterminado para ubicaciones genómicas no afectadas. Dependiendo de las suposiciones iniciales (por ejemplo, ploidía fija o ploidía optimizada, fracción fetal fija o fracción fetal optimizada o combinaciones de las mismas) el umbral predeterminado o valor de punto de corte o rango umbral de valores indicativos de la presencia o ausencia de una variación genética puede variar mientras todavía proporciona un resultado útil para determinar la presencia o ausencia de una variación genética. En algunas implementaciones, un perfil es indicativo de y/o representativo de un fenotipo.

A modo de ejemplo no limitativo, los perfiles de muestras normalizadas y/o recuentos de referencia pueden obtenerse a partir de datos de lectura de secuencia sin procesar mediante (a) cálculo de la mediana de recuentos de referencia para cromosomas, secciones genómicas o segmentos de los mismos seleccionados de un conjunto de referencias que se sabe que no portan una variación genética, (b) eliminación de secciones genómicas no informativas de los recuentos sin procesar de muestra de referencia (por ejemplo, filtrar); (c) normalizar los recuentos de referencia para todos los bins restantes con respecto al número residual total de recuentos (por ejemplo, suma de los recuentos restantes después de eliminar los bins no informativos) para el cromosoma seleccionado de la muestra de referencia o ubicación genómica seleccionada, generándose de ese modo un perfil de sujeto de referencia normalizado; (d) eliminar las secciones genómicas correspondientes de la muestra de sujeto de prueba; y (e) normalizar los recuentos de sujetos de prueba restantes para una o más ubicaciones genómicas seleccionadas con respecto a la suma de la mediana de recuentos de referencia residuales para el cromosoma o los cromosomas que contienen las ubicaciones genómicas seleccionadas, generándose de ese modo un perfil de sujeto de prueba normalizado. En determinadas implementaciones, una etapa de normalización adicional con respecto a todo el genoma, reducida por las secciones genómicas filtradas en (b), puede incluirse entre (c) y (d).

Un perfil de conjunto de datos puede generarse mediante una o más manipulaciones de datos de lecturas de secuencia mapeadas contadas. Algunas implementaciones incluyen lo siguiente. Las lecturas de secuencia se mapean y se determina el número de etiquetas de secuencia que se mapean en cada bin genómico (por ejemplo, se cuentan). Se genera un perfil de recuento sin procesar a partir de las lecturas de secuencia mapeadas que se cuentan. Se proporciona un resultado comparando un perfil de recuento sin procesar de un sujeto de prueba con una mediana de perfil de recuento de referencia para cromosomas, secciones genómicas o segmentos de los mismos de un conjunto de sujetos de referencia que se sabe que no presentan una variación genética, en determinadas implementaciones.

En algunas implementaciones, los datos de lectura de secuencia se filtran, opcionalmente, para eliminar los datos con ruido o las secciones genómicas no informativas. Después del filtrado, los recuentos restantes se suman normalmente para generar un conjunto de datos filtrado. Un perfil de recuento filtrado se genera a partir de un conjunto de datos filtrado, en determinadas implementaciones.

Después de contar los datos de lectura de secuencia y, opcionalmente, filtrarse, los conjuntos de datos pueden normalizarse para generar elevaciones o perfiles. Un conjunto de datos puede normalizarse normalizando una o más secciones genómicas seleccionadas con respecto a un valor de referencia de normalización adecuado. En algunas implementaciones, un valor de referencia normalizado es representativo de los recuentos totales para el cromosoma o cromosomas de los cuales se seleccionan las secciones genómicas. En determinadas implementaciones, un valor de referencia de normalización es representativo de una o más secciones genómicas correspondientes, porciones de cromosomas o cromosomas de un conjunto de datos de referencia preparado a partir de un conjunto de sujetos de referencia que se sabe que no presentan una variación genética. En algunas implementaciones, un valor de referencia de normalización es representativo de una o más secciones genómicas correspondientes, porciones de cromosomas o cromosomas de un conjunto de datos del sujeto de prueba preparado a partir de un sujeto de prueba que se analiza para determinar la presencia o ausencia de una variación genética. En determinadas implementaciones, el procedimiento de normalización se realiza utilizando un enfoque de ventana estática y, en algunas implementaciones, el procedimiento de normalización se realiza utilizando un enfoque de ventana móvil o deslizante. En determinadas implementaciones, se genera un perfil que comprende recuentos normalizados para facilitar la clasificación y/o proporcionar un resultado. Puede proporcionarse un resultado basado en una representación gráfica de un perfil que comprende recuentos normalizados (por ejemplo, usando una representación gráfica de tal perfil).

Elevaciones

En algunas implementaciones, se atribuye un valor a una elevación (por ejemplo, un número). Una elevación puede determinarse mediante un método, una operación o un procedimiento matemático adecuado (por ejemplo, una elevación procesada). El término "nivel", tal como se usa en el presente documento, es sinónimo del término "elevación" tal como se usa en el presente documento. A menudo, una elevación es, o se deriva de, recuentos (por ejemplo, recuentos normalizados) para un conjunto de secciones genómicas. Algunas veces, una elevación de una sección genómica es sustancialmente igual al número total de recuentos mapeados en una sección genómica (por ejemplo, recuentos normalizados). Con frecuencia, una elevación se determina a partir de recuentos que se procesan o manipulan mediante

un método, una operación o un procedimiento matemático adecuado conocido en la técnica. Algunas veces, una elevación se deriva de recuentos que se procesan y los ejemplos no limitativos de recuentos procesados incluyen recuentos ponderados, eliminados, filtrados, normalizados, ajustados, promediados, derivados como una media (por ejemplo, elevación media), sumados, restados, transformados o combinaciones de los mismos. Algunas veces, una elevación comprende recuentos normalizados (por ejemplo, recuentos normalizados de secciones genómicas). Una elevación puede ser para recuentos normalizados mediante un procedimiento adecuado, los ejemplos no limitativos de los cuales incluyen normalización basada en bins, normalización por contenido de GC, regresión lineal y no lineal por mínimos cuadrados, LOESS de GC, LOWESS, PERUN, RM, GCRM, cQn, similares y/o combinaciones de los mismos. Una elevación puede comprender recuentos normalizados o cantidades relativas de recuentos. Algunas veces, una elevación es para recuentos o recuentos normalizados de dos o más secciones genómicas que se promedian y la elevación se denomina elevación promedio. Algunas veces, una elevación es para un conjunto de secciones genómicas que tienen un mediana de recuento o media de recuentos normalizados que se denomina elevación media. Algunas veces se deriva una elevación para secciones genómicas que comprenden recuentos sin procesar y/o filtrados.

En algunas implementaciones, una elevación se basa en recuentos que son sin procesar. Algunas veces, una elevación se asocia con un valor de incertidumbre. Una elevación para una sección genómica, o una "elevación de sección genómica", es sinónimo de un "nivel de sección genómica" en el presente documento.

Los recuentos normalizados o no normalizados para dos o más elevaciones (por ejemplo, dos o más elevaciones en un perfil) pueden, algunas veces, manipularse matemáticamente (por ejemplo, sumarse, multiplicarse, promediarse, normalizarse, similares o una combinación de los mismos) según las elevaciones. Por ejemplo, los recuentos normalizados o no normalizados para dos o más elevaciones pueden normalizarse según una, algunas o todas las elevaciones de un perfil. Algunas veces, los recuentos normalizados o no normalizados de todas las elevaciones en un perfil se normalizan según una elevación del perfil. Algunas veces, los recuentos normalizados o no normalizados de una primera elevación en un perfil se normalizan según los recuentos normalizados o no normalizados de una segunda elevación en el perfil.

Los ejemplos no limitativos de una elevación (por ejemplo, una primera elevación, una segunda elevación) son una elevación para un conjunto de secciones genómicas que comprenden recuentos procesados, una elevación para un conjunto de secciones genómicas que comprenden una media, mediana o promedio de recuentos, una elevación para un conjunto de secciones genómicas que comprenden recuentos normalizados, similares o cualquier combinación de los mismos. En algunas implementaciones, una primera elevación y una segunda elevación en un perfil se derivan de recuentos de secciones genómicas mapeadas en el mismo cromosoma. En algunas implementaciones, una primera elevación y una segunda elevación en un perfil se derivan de recuentos de secciones genómicas mapeadas en cromosomas diferentes.

En algunas implementaciones, una elevación se determina a partir de recuentos normalizados o no normalizados mapeados en una o más secciones genómicas. En algunas implementaciones, una elevación se determina a partir de recuentos normalizados o no normalizados mapeados en dos o más secciones genómicas, en el que los recuentos normalizados para cada sección genómica son a menudo aproximadamente iguales. Puede haber variación en los recuentos (por ejemplo, recuentos normalizados) en un conjunto de secciones genómicas para una elevación. En un conjunto de secciones genómicas para una elevación puede haber una o más secciones genómicas que tienen recuentos que son significativamente diferentes que en otras secciones genómicas del conjunto (por ejemplo, picos y/o depresiones). Cualquier número adecuado de recuentos normalizados o no normalizados asociados con cualquier número adecuado de secciones genómicas puede definir una elevación.

Algunas veces, una o más elevaciones pueden determinarse a partir de recuentos normalizados o no normalizados de todas o algunas de las secciones genómicas de un genoma. A menudo, puede determinarse una elevación a partir de todos o algunos de los recuentos normalizados o no normalizados de un cromosoma o segmento del mismo.

A veces, dos o más recuentos derivados de dos o más secciones genómicas (por ejemplo, un conjunto de secciones genómicas) determinan una elevación. A veces dos o más recuentos (por ejemplo, recuentos de dos o más secciones genómicas) determinan una elevación. En algunas implementaciones, los recuentos de 2 a aproximadamente 100.000 secciones genómicas determinan una elevación. En algunas implementaciones, recuentos de 2 a aproximadamente 50.000, de 2 a aproximadamente 40.000, de 2 a aproximadamente 30.000, de 2 a aproximadamente 20.000, de 2 a aproximadamente 10.000, de 2 a aproximadamente 5000, de 2 a aproximadamente 2500, de 2 a aproximadamente 1250, de 2 a aproximadamente 1000, de 2 a aproximadamente 500, de 2 a aproximadamente 250, de 2 a aproximadamente 100 o de 2 a aproximadamente 60 secciones genómicas determinan una elevación. En algunas implementaciones, los recuentos de aproximadamente 10 a aproximadamente 50 secciones genómicas determinan una elevación. En algunas implementaciones, los recuentos de aproximadamente 20 a aproximadamente 40 o más secciones genómicas determinan una elevación. En algunas implementaciones, una elevación comprende recuentos de aproximadamente 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 45, 50, 55, 60 o más secciones genómicas. En algunas implementaciones, una elevación corresponde a un conjunto de secciones genómicas (por ejemplo, un conjunto de secciones genómicas de un genoma de referencia, un conjunto de secciones genómicas de un cromosoma o un conjunto de secciones genómicas de un segmento de un cromosoma).

En algunas implementaciones, se determina una elevación para recuentos normalizados o no normalizados de secciones genómicas contiguas. Algunas veces, las secciones genómicas (por ejemplo, un conjunto de secciones

genómicas) que son contiguas representan segmentos vecinos de un genoma o segmentos vecinos de un cromosoma o gen. Por ejemplo, dos o más secciones genómicas contiguas, cuando se alinean al fusionar las secciones genómicas extremo con extremo, pueden representar un ensamblaje de secuencias de una secuencia de ADN más larga que cada sección genómica. Por ejemplo, dos o más secciones genómicas contiguas pueden representar un genoma, cromosoma, gen, intrón, exón o segmento de los mismos intacto. Algunas veces se determina una elevación a partir de una colección (por ejemplo, un conjunto) de secciones genómicas contiguas y/o secciones genómicas no contiguas.

Elevaciones significativamente diferentes

En algunas implementaciones, un perfil de recuentos normalizados comprende una elevación (por ejemplo, una primera elevación) significativamente diferente de otra elevación (por ejemplo, una segunda elevación) dentro del perfil. Una primera elevación puede ser mayor o menor que una segunda elevación. En algunas implementaciones, una primera elevación es para un conjunto de secciones genómicas que comprenden una o más lecturas que comprenden una variación del número de copias (por ejemplo, una variación del número de copias materno, una variación del número de copias fetal, o una variación del número de copias materno y una variación del número de copias fetal) y la segunda elevación es para un conjunto de secciones genómicas que comprenden lecturas que no tienen sustancialmente ninguna variación del número de copias. En algunas implementaciones, significativamente diferente se refiere a una diferencia observable. Algunas veces, significativamente diferente se refiere a una diferencia estadísticamente diferente o estadísticamente significativa. Una diferencia estadísticamente significativa es, algunas veces, una evaluación estadística de una diferencia observada. Una diferencia estadísticamente significativa puede evaluarse mediante un método adecuado en la técnica. Puede usarse cualquier umbral o rango adecuado para determinar que dos elevaciones son significativamente diferentes. En algunos casos, dos elevaciones (por ejemplo, elevaciones medias) que difieren en aproximadamente el 0,01 por ciento o más (por ejemplo, el 0,01 por ciento de uno o cualquiera de los valores de elevación) son significativamente diferentes. A veces dos elevaciones (por ejemplo, elevaciones medias) que difieren en aproximadamente el 0,1 por ciento o más son significativamente diferentes. En algunos casos, dos elevaciones (por ejemplo, elevaciones medias) que difieren en aproximadamente el 0,5 por ciento o más son significativamente diferentes. A veces dos elevaciones (por ejemplo, elevaciones medias) que difieren en aproximadamente el 0,5, 0,75, 1, 1,5, 2, 2,5, 3, 3,5, 4, 4,5, 5, 5,5, 6, 6,5, 7, 7,5, 8, 8,5, 9, 9,5 o más que aproximadamente el 10 % son significativamente diferentes. A veces, dos elevaciones (por ejemplo, elevaciones medias) son significativamente diferentes y no hay solapamiento en ninguna elevación y/o no hay solapamiento en un rango definido por un valor de incertidumbre calculado para una o ambas elevaciones. En algunos casos, el valor de incertidumbre es una desviación estándar expresada como sigma. A veces dos elevaciones (por ejemplo, elevaciones medias) son significativamente diferentes y difieren en aproximadamente 1 o más veces el valor de incertidumbre (por ejemplo, 1 sigma). A veces, dos elevaciones (por ejemplo, elevaciones medias) son significativamente diferentes y difieren en aproximadamente 2 o más veces el valor de incertidumbre (por ejemplo, 2 sigma), aproximadamente 3 o más, aproximadamente 4 o más, aproximadamente 5 o más, aproximadamente 6 o más, aproximadamente 7 o más, aproximadamente 8 o más, aproximadamente 9 o más, o aproximadamente 10 o más veces el valor de incertidumbre. Algunas veces, dos elevaciones (por ejemplo, elevaciones medias) son significativamente diferentes cuando difieren en aproximadamente 1,1, 1,2, 1,3, 1,4, 1,5, 1,6, 1,7, 1,8, 1,9, 2,0, 2,1, 2,2, 2,3, 2,4, 2,5, 2,6, 2,7, 2,8, 2,9, 3,0, 3,1, 3,2, 3,3, 3,4, 3,5, 3,6, 3,7, 3,8, 3,9 o 4,0 veces el valor de incertidumbre o más. En algunas implementaciones, el nivel de confianza aumenta a medida que lo hace la diferencia entre dos elevaciones. En algunos casos, el nivel de confianza disminuye a medida que disminuye la diferencia entre dos elevaciones y/o a medida que aumenta el valor de incertidumbre. Por ejemplo, algunas veces el nivel de confianza aumenta con la relación de la diferencia entre las elevaciones y la desviación estándar (por ejemplo, D.M.A.).

En algunas implementaciones, un primer conjunto de secciones genómicas incluye a menudo secciones genómicas que son diferentes de (por ejemplo, no solapantes con) un segundo conjunto de secciones genómicas. Por ejemplo, a veces una primera elevación de recuentos normalizados es significativamente diferente de una segunda elevación de recuentos normalizados en un perfil, y la primera elevación es para un primer conjunto de secciones genómicas, la segunda elevación es para un segundo conjunto de secciones genómicas y las secciones genómicas no se solapan en el primer conjunto ni en el segundo conjunto de secciones genómicas. En algunos casos, un primer conjunto de secciones genómicas no es un subconjunto de un segundo conjunto de secciones genómicas a partir de las cuales se determinan una primera elevación y una segunda elevación, respectivamente. Algunas veces, un primer conjunto de secciones genómicas es diferente y/o distinto de un segundo conjunto de secciones genómicas a partir de las cuales se determinan una primera elevación y una segunda elevación, respectivamente.

Algunas veces, un primer conjunto de secciones genómicas es un subconjunto de un segundo conjunto de secciones genómicas en un perfil. Por ejemplo, a veces una segunda elevación de recuentos normalizados para un segundo conjunto de secciones genómicas en un perfil comprende recuentos normalizados de un primer conjunto de secciones genómicas para una primera elevación en el perfil y el primer conjunto de secciones genómicas es un subconjunto del segundo conjunto de secciones genómicas en el perfil. Algunas veces, una elevación promedio, media o mediana de elevación se deriva de una segunda elevación en el que la segunda elevación comprende una primera elevación. A veces, una segunda elevación comprende un segundo conjunto de secciones genómicas que representan un cromosoma completo y una primera elevación comprende un primer conjunto de secciones genómicas, en la que el primer conjunto es un subconjunto del segundo conjunto de secciones genómicas y la primera elevación representa

una variación del número de copias materno, variación del número de copias fetal, o una variación del número de copias materno y una variación del número de copias fetal que está presente en el cromosoma.

En algunas implementaciones, un valor de una segunda elevación está más próximo al valor medio, promedio o mediana del valor de un perfil de recuento para un cromosoma, o segmento del mismo, que la primera elevación. En algunas implementaciones, una segunda elevación es una elevación media de un cromosoma, una porción de un cromosoma o un segmento de los mismos. En algunas implementaciones, una primera elevación es significativamente diferente de una elevación predominante (por ejemplo, una segunda elevación) que representa un cromosoma, o segmento del mismo. Un perfil puede incluir múltiples primeras elevaciones que difieren significativamente de una segunda elevación, y cada primera elevación puede ser, independientemente, mayor o menor que la segunda elevación. En algunas implementaciones, una primera elevación y una segunda elevación se derivan del mismo cromosoma y la primera elevación es mayor o menor que la segunda elevación, y la segunda elevación es la elevación predominante del cromosoma. A veces, una primera elevación y una segunda elevación se derivan del mismo cromosoma, una primera elevación es indicativa de una variación del número de copias (por ejemplo, una variación del número de copias materno y/o fetal, deleción, inserción, duplicación) y una segunda elevación es una elevación media o elevación predominante de las secciones genómicas para un cromosoma, o segmento del mismo.

En algunos casos, una lectura en un segundo conjunto de secciones genómicas para una segunda elevación no incluye sustancialmente una variación genética (por ejemplo, una variación del número de copias, una variación del número de copias materno y/o fetal). A menudo, un segundo conjunto de secciones genómicas para una segunda elevación incluye cierta variabilidad (por ejemplo, variabilidad en la elevación, variabilidad en los recuentos para secciones genómicas). A veces, una o más secciones genómicas en un conjunto de secciones genómicas para una elevación asociada sustancialmente con la ausencia de variación del número de copias incluyen una o más lecturas que tienen una variación del número de copias presente en un genoma materno y/o fetal. Por ejemplo, a veces un conjunto de secciones genómicas incluyen una variación del número de copias que está presente en un pequeño segmento de un cromosoma (por ejemplo, menos de 10 secciones genómicas) y el conjunto de secciones genómicas es para una elevación asociada sustancialmente con la ausencia de variación del número de copias. Por tanto, un conjunto de secciones genómicas que no incluyen sustancialmente ninguna variación del número de copias todavía puede incluir una variación del número de copias que está presente en menos de aproximadamente 10, 9, 8, 7, 6, 5, 4, 3, 2 o 1 secciones genómicas de una elevación.

Algunas veces, una primera elevación es para un primer conjunto de secciones genómicas y una segunda elevación es para un segundo conjunto de secciones genómicas y el primer conjunto de secciones genómicas y el segundo conjunto de secciones genómicas son contiguos (por ejemplo, adyacentes con respecto a la secuencia de ácido nucleico de un cromosoma o segmento del mismo). Algunas veces, el primer conjunto de secciones genómicas y el segundo conjunto de secciones genómicas no son contiguos.

Pueden usarse lecturas de secuencia relativamente cortas de una mezcla de ácido nucleico fetal y materno para proporcionar recuentos que pueden transformarse en una elevación y/o un perfil. Los recuentos, elevaciones y perfiles pueden representarse en forma electrónica o tangible y pueden visualizarse. Los recuentos mapeados en secciones genómicas (por ejemplo, representados como elevaciones y/o perfiles) pueden proporcionar una representación visual de un genoma fetal y/o materno, cromosoma o una porción o un segmento de un cromosoma que está presente en un feto y/o mujer embarazada.

Módulo de comparación

Una primera elevación puede identificarse como significativamente diferente de una segunda elevación por un módulo de comparación o por un aparato que comprende un módulo de comparación. En algunas implementaciones, se requiere un módulo de comparación o un aparato que comprende un módulo de comparación para proporcionar una comparación entre dos elevaciones. Un aparato que comprende un módulo de comparación puede comprender al menos un procesador. En algunas implementaciones, se determina que las elevaciones son significativamente diferentes por un aparato que incluye un procesador (por ejemplo, uno o más procesadores), procesador que puede realizar y/o implementar una o más instrucciones (por ejemplo, procesos, rutinas y/o subrutinas) del módulo de comparación. En algunas implementaciones, se determina que las elevaciones son significativamente diferentes por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunas implementaciones, un módulo de comparación funciona con uno o más procesadores externos (por ejemplo, una red interna o externa, un servidor, dispositivo de almacenamiento y/o una red de almacenamiento (por ejemplo, una nube)). En algunas implementaciones, se determina que las elevaciones son significativamente diferentes por un aparato que comprende uno o más de los siguientes: una o más celdas de flujo, una cámara, componentes de manipulación de fluido, una impresora, una pantalla de visualización (por ejemplo, un LED, LCT o CRT) y similares. Un módulo de comparación puede recibir información y/o datos de un módulo adecuado. Un módulo de comparación puede recibir información y/o datos de un módulo de secuenciación, un módulo de mapeo, un módulo de recuento o un módulo de normalización. Un módulo de comparación puede recibir información y/o datos normalizados de un módulo de normalización. La información y/o los datos derivados de, o transformados por, un módulo de comparación pueden transferirse de un módulo de comparación a un módulo de establecimiento de rango, un módulo de representación gráfica, un módulo de ajuste, un módulo de categorización o un módulo de resultados. Una comparación entre dos o más elevaciones y/o una identificación de una elevación significativamente diferente de otra elevación puede

transferirse de (por ejemplo, proporcionarse a) un módulo de comparación a un módulo de categorización, módulo de establecimiento de rango o módulo de ajuste.

Elevación de referencia y valor de referencia normalizado

A veces, un perfil comprende una elevación de referencia (por ejemplo, una elevación usada como referencia). A menudo, un perfil de recuentos normalizados proporciona una elevación de referencia a partir de la cual se determinan las elevaciones esperadas y los rangos esperados (véase la descripción más adelante de las elevaciones y los rangos esperados). Una elevación de referencia es a menudo para recuentos normalizados de secciones genómicas que comprenden lecturas mapeadas de una madre y de un feto. Una elevación de referencia es a menudo la suma de los recuentos normalizados de lecturas mapeadas de un feto y una madre (por ejemplo, una mujer embarazada). A veces, una elevación de referencia es para secciones genómicas que comprenden lecturas mapeadas de una madre euploide y/o un feto euploide. Algunas veces, una elevación de referencia es para secciones genómicas que comprenden lecturas mapeadas que tienen una variación genética fetal (por ejemplo, una aneuploidía (por ejemplo, una trisomía)), y/o lecturas que tienen una variación genética materna (por ejemplo, una variación del número de copias, inserción, deleción). Algunas veces, una elevación de referencia es para secciones genómicas que no incluyen sustancialmente ninguna variación del número de copias materno y/o fetal. A veces se usa una segunda elevación como elevación de referencia. En algunos casos, un perfil comprende una primera elevación de recuentos normalizados y una segunda elevación de recuentos normalizados, la primera elevación es significativamente diferente de la segunda elevación y la segunda elevación es la elevación de referencia. En algunos casos, un perfil comprende una primera elevación de recuentos normalizados para un primer conjunto de secciones genómicas, una segunda elevación de recuentos normalizados para un segundo conjunto de secciones genómicas, el primer conjunto de secciones genómicas incluye lecturas mapeadas que tienen una variación del número de copias materno y/o fetal, el segundo conjunto de secciones genómicas comprende lecturas mapeadas que no tienen sustancialmente variación del número de copias materno y/o variación del número de copias fetal, y la segunda elevación es una elevación de referencia.

En algunas implementaciones, los recuentos mapeados en secciones genómicas para una o más elevaciones de un perfil se normalizan según los recuentos de una elevación de referencia. En algunas implementaciones, normalizar recuentos de una elevación según recuentos de una elevación de referencia comprenden dividir los recuentos de una elevación entre los recuentos de una elevación de referencia o un múltiplo o fracción de los mismos. Los recuentos normalizados según los recuentos de una elevación de referencia se han normalizado a menudo según otro procedimiento (por ejemplo, PERUN) y además a menudo, los recuentos de una elevación de referencia se han normalizado (por ejemplo, mediante PERUN). A veces, los recuentos de una elevación se normalizan según los recuentos de una elevación de referencia y los recuentos de la elevación de referencia son escalables hasta un valor adecuado ya sea antes o después de la normalización. El procedimiento de escalar los recuentos de una elevación de referencia puede comprender cualquier constante adecuada (es decir, número) y cualquier manipulación matemática adecuada puede aplicarse a los recuentos de una elevación de referencia.

Un valor de referencia normalizado (NRV) se determina a menudo según los recuentos normalizados de una elevación de referencia. Determinar un NRV puede comprender cualquier procedimiento de normalización adecuado (por ejemplo, manipulación matemática) aplicado a los recuentos de una elevación de referencia en el que se usa el mismo procedimiento de normalización para normalizar los recuentos de otras elevaciones dentro del mismo perfil. Determinar un NRV comprende a menudo dividir una elevación de referencia entre sí misma. Determinar un NRV comprende a menudo dividir una elevación de referencia entre un múltiplo de sí misma. Determinar un NRV comprende a menudo dividir una elevación de referencia entre la suma o diferencia de la elevación de referencia y una constante (por ejemplo, cualquier número).

Un NRV se denomina, algunas veces, valor nulo. Un NRV puede ser cualquier valor adecuado. En algunas implementaciones, un NRV es cualquier valor distinto de cero. Algunas veces, un NRV es un número entero. Algunas veces, un NRV es un entero positivo. En algunas implementaciones, un NRV es 1, 10, 100 o 1000. A menudo, un NRV es igual a 1. Algunas veces, un NRV es igual a cero. Los recuentos de una elevación de referencia pueden normalizarse con respecto a cualquier NRV adecuado. En algunas implementaciones, los recuentos de una elevación de referencia se normalizan con respecto a un NRV de cero. A menudo, los recuentos de una elevación de referencia se normalizan con respecto a un NRV de 1.

Elevaciones esperadas

Una elevación esperada es algunas veces una elevación predefinida (por ejemplo, una elevación teórica, elevación predicha). Una "elevación esperada" se denomina algunas veces en el presente documento un "valor de elevación predeterminado". En algunas implementaciones, una elevación esperada es un valor predicho para una elevación de recuentos normalizados para un conjunto de secciones genómicas que incluyen una variación del número de copias. En algunos casos, se determina una elevación esperada para un conjunto de secciones genómicas que no incluyen sustancialmente ninguna variación del número de copias. Una elevación esperada puede determinarse para una ploidía cromosómica (por ejemplo, 0, 1, 2 (es decir, diploide), 3 o 4 cromosomas) o una microploidía (deleción homocigota o heterocigota, duplicación, inserción o ausencia de los mismos). A menudo, se determina una elevación esperada para una microploidía materna (por ejemplo, una variación del número de copias materno y/o fetal).

Una elevación esperada para una variación genética o una variación del número de copias puede determinarse de cualquier manera adecuada. A menudo, una elevación esperada se determina mediante una manipulación matemática adecuada de una elevación (por ejemplo, recuentos mapeados en un conjunto de secciones genómicas para una elevación). Algunas veces se determina una elevación esperada usando una constante algunas veces denominada constante de elevación esperada. Una elevación esperada para una variación del número de copias se calcula, a veces, al multiplicar una elevación de referencia, recuentos normalizados de una elevación de referencia o un NRV por una constante de elevación esperada, sumar una constante de elevación esperada, restar una constante de elevación esperada, dividir entre una constante de elevación esperada, o mediante una combinación de los mismos. A menudo, una elevación esperada (por ejemplo, una elevación esperada de una variación del número de copias materno y/o fetal) determinada para el mismo sujeto, muestra o grupo de prueba se determina según la misma elevación de referencia o NRV.

A menudo, una elevación esperada se determina multiplicando una elevación de referencia, recuentos normalizados de una elevación de referencia o un NRV por una constante de elevación esperada, en el que la elevación de referencia, recuentos normalizados de una elevación de referencia o NRV no es igual a cero. A veces se determina una elevación esperada sumando una constante de elevación esperada a la elevación de referencia, recuentos normalizados de una elevación de referencia o un NRV que es igual a cero. En algunas implementaciones, una elevación esperada, recuentos normalizados de una elevación de referencia, NRV y constante de elevación esperada son escalables. El procedimiento de escalado puede comprender cualquier constante adecuada (es decir, número) y cualquier manipulación matemática adecuada en el que el mismo procedimiento de escalado se aplica a todos los valores considerados.

Constante de elevación esperada

Una constante de elevación esperada puede determinarse mediante un método adecuado. A veces se determina arbitrariamente una constante de elevación esperada. A menudo, se determina empíricamente una constante de elevación esperada. Algunas veces se determina una constante de elevación esperada según una manipulación matemática. Algunas veces se determina una constante de elevación esperada según una referencia (por ejemplo, un genoma de referencia, una muestra de referencia, datos de prueba de referencia). En algunas implementaciones, una constante de elevación esperada se predetermina para una elevación representativa de la presencia o ausencia de una variación genética o variación del número de copias (por ejemplo, una duplicación, inserción o deleción). En algunas implementaciones, una constante de elevación esperada se predetermina para una elevación representativa de la presencia o ausencia de una variación del número de copias materno, variación del número de copias fetal, o una variación del número de copias materno y una variación del número de copias fetal. Una constante de elevación esperada para una variación del número de copias puede ser cualquier constante o conjunto de constantes adecuadas.

En algunas implementaciones, la constante de elevación esperada para una duplicación homocigota (por ejemplo, una duplicación homocigota) puede ser de aproximadamente 1,6 a aproximadamente 2,4, de aproximadamente 1,7 a aproximadamente 2,3, de aproximadamente 1,8 a aproximadamente 2,2 o de aproximadamente 1,9 a aproximadamente 2,1. Algunas veces la constante de elevación esperada para una duplicación homocigota es de aproximadamente 1,6, 1,7, 1,8, 1,9, 2,0, 2,1, 2,2, 2,3 o aproximadamente 2,4. A menudo, la constante de elevación esperada para una duplicación homocigota es de aproximadamente 1,90, 1,92, 1,94, 1,96, 1,98, 2,0, 2,02, 2,04, 2,06, 2,08 o aproximadamente 2,10. A menudo, la constante de elevación esperada para una duplicación homocigota es de aproximadamente 2.

En algunas implementaciones, la constante de elevación esperada para una duplicación heterocigota (por ejemplo, una duplicación homocigota) es de aproximadamente 1,2 a aproximadamente 1,8, de aproximadamente 1,3 a aproximadamente 1,7 o de aproximadamente 1,4 a aproximadamente 1,6. Algunas veces la constante de elevación esperada para una duplicación heterocigota es de aproximadamente 1,2, 1,3, 1,4, 1,5, 1,6, 1,7 o aproximadamente 1,8. A menudo, la constante de elevación esperada para una duplicación heterocigota es de aproximadamente 1,40, 1,42, 1,44, 1,46, 1,48, 1,5, 1,52, 1,54, 1,56, 1,58 o aproximadamente 1,60. En algunas implementaciones, la constante de elevación esperada para una duplicación heterocigota es de aproximadamente 1,5.

En algunas implementaciones, la constante de elevación esperada para la ausencia de una variación del número de copias (por ejemplo, la ausencia de una variación del número de copias materno y/o variación del número de copias fetal) es de aproximadamente 1,3 a aproximadamente 0,7, de aproximadamente 1,2 a aproximadamente 0,8 o de aproximadamente 1,1 a aproximadamente 0,9. Algunas veces la constante de elevación esperada para la ausencia de una variación del número de copias es de aproximadamente 1,3, 1,2, 1,1, 1,0, 0,9, 0,8 o aproximadamente 0,7. A menudo, la constante de elevación esperada para la ausencia de una variación del número de copias es de aproximadamente 1,09, 1,08, 1,06, 1,04, 1,02, 1,0, 0,98, 0,96, 0,94 o aproximadamente 0,92. En algunas implementaciones, la constante de elevación esperada para la ausencia de una variación del número de copias es de aproximadamente 1.

En algunas implementaciones, la constante de elevación esperada para una deleción heterocigota (por ejemplo, una deleción materna, fetal, o una deleción materna y una fetal heterocigota) es de aproximadamente 0,2 a aproximadamente 0,8, de aproximadamente 0,3 a aproximadamente 0,7 o de aproximadamente 0,4 a aproximadamente 0,6. Algunas veces, la constante de elevación esperada para una deleción heterocigota es de aproximadamente 0,2, 0,3, 0,4, 0,5, 0,6, 0,7 o aproximadamente 0,8. A menudo, la constante de elevación esperada para una deleción heterocigota es de

aproximadamente 0,40, 0,42, 0,44, 0,46, 0,48, 0,5, 0,52, 0,54, 0,56, 0,58 o aproximadamente 0,60. En algunas implementaciones, la constante de elevación esperada para una deleción heterocigota es de aproximadamente 0,5.

En algunas implementaciones, la constante de elevación esperada para una deleción homocigota (por ejemplo, una deleción homocigota) puede ser de aproximadamente -0,4 a aproximadamente 0,4, de aproximadamente -0,3 a aproximadamente 0,3, de aproximadamente -0,2 a aproximadamente 0,2 o de aproximadamente -0,1 a aproximadamente 0,1. Algunas veces, la constante de elevación esperada para una deleción homocigota es de aproximadamente -0,4, -0,3, -0,2, -0,1, 0,0, 0,1, 0,2, 0,3 o aproximadamente 0,4. A menudo, la constante de elevación esperada para una deleción homocigota es de aproximadamente -0,1, -0,08, -0,06, -0,04, -0,02, 0,0, 0,02, 0,04, 0,06, 0,08 o aproximadamente 0,10. A menudo, la constante de elevación esperada para una deleción homocigota es de aproximadamente 0.

Rango de elevación esperado

A veces la presencia o ausencia de una variación genética o variación del número de copias (por ejemplo, una variación del número de copias materno, una variación del número de copias fetal, o una variación del número de copias materno y una variación del número de copias fetal) se determina por una elevación que se encuentra dentro o fuera de un rango de elevación esperado. Un rango de elevación esperado se determina a menudo según una elevación esperada. Algunas veces se determina un rango de elevación esperado para una elevación que no comprende sustancialmente ninguna variación genética o sustancialmente ninguna variación del número de copias. Puede usarse un método adecuado para determinar un rango de elevación esperado.

A veces, un rango de elevación esperado se define según un valor de incertidumbre adecuado calculado para una elevación. Los ejemplos no limitativos de un valor de incertidumbre son una desviación estándar, error estándar, varianza calculada, valor de p y desviación media absoluta (D.M.A.). A veces, un rango de elevación esperado para una variación genética o una variación del número de copias se determina, en parte, mediante el cálculo del valor de incertidumbre para una elevación (por ejemplo, una primera elevación, una segunda elevación, una primera elevación y una segunda elevación). Algunas veces se define un rango de elevación esperado según un valor de incertidumbre calculado para un perfil (por ejemplo, un perfil de recuentos normalizados para un cromosoma o segmento del mismo). En algunas implementaciones, se calcula un valor de incertidumbre para una elevación que no comprende sustancialmente ninguna variación genética o sustancialmente ninguna variación del número de copias. En algunas implementaciones, se calcula un valor de incertidumbre para una primera elevación, una segunda elevación o una primera elevación y una segunda elevación. En algunas implementaciones se determina un valor de incertidumbre para una primera elevación, una segunda elevación o una segunda elevación que comprende una primera elevación.

Un rango de elevación esperado a veces se calcula, en parte, al multiplicar, sumar, restar o dividir un valor de incertidumbre por una constante (por ejemplo, una constante predeterminada) *n.* Puede usarse Un procedimiento matemático o combinación de procedimientos adecuados. La constante *n* (por ejemplo, constante predeterminada *n)* a veces se denomina intervalo de confianza. Un intervalo de confianza seleccionado se determina según la constante *n* que se selecciona. La constante *n* (por ejemplo, la constante predeterminada *n,* el intervalo de confianza) puede determinarse de manera adecuada. La constante *n* puede ser un número o fracción de un número mayor de cero. La constante *n* puede ser un número entero. A menudo, la constante *n* es un número menor de 10. Algunas veces, la constante *n* es un número menor de aproximadamente 10, menor de aproximadamente 9, menor de aproximadamente 8, menor de aproximadamente 7, menor de aproximadamente 6, menor de aproximadamente 5, menor de aproximadamente 4, menor de aproximadamente 3 o menor de aproximadamente 2. Algunas veces, la constante *n* es de aproximadamente 10, 9,5, 9, 8,5, 8, 7,5, 7, 6,5, 6, 5,5, 5, 4,5, 4, 3,5, 3, 2,5, 2 o 1. La constante *n* puede determinarse empíricamente a partir de datos derivados de sujetos (una mujer embarazada y/o un feto) con una disposición genética conocida.

A menudo, un valor de incertidumbre y una constante *n* define un rango (por ejemplo, un punto de corte de incertidumbre). Por ejemplo, algunas veces un valor de incertidumbre es una desviación estándar (por ejemplo, +/- 5) y se multiplica por una constante *n* (por ejemplo, un intervalo de confianza) definiendo de ese modo un rango o punto de corte de incertidumbre (por ejemplo, de 5*n* a - 5*n*).

En algunas implementaciones, un rango de elevación esperado para una variación genética (por ejemplo, una variación del número de copias materno, una variación del número de copias fetal, o una variación del número de copias materno y una variación del número de copias fetal) es la suma de una elevación esperada más una constante *n* veces la incertidumbre (por ejemplo, *n* x sigma (por ejemplo, 6 sigma)). Algunas veces el rango de elevación esperado para una variación genética o variación del número de copias designado por *k* puede definirse por la fórmula:

Fórmula R: $(\text{rango de elevación esperado})_k = (\text{elevación esperada})_k + n\sigma$

donde σ es un valor de incertidumbre, *n* es una constante (por ejemplo, una constante predeterminada) y el rango de elevación esperado y la elevación esperada son para la variación genética *k* (por ejemplo, *k* = una deleción heterocigota, por ejemplo, *k* = la ausencia de una variación genética). Por ejemplo, para una elevación esperada igual a 1 (por ejemplo, la ausencia de una variación del número de copias), un valor de incertidumbre (es decir σ) igual a +/- 0,05, y *n* = 3, el rango de elevación esperado se define como de 1,15 a 0,85. En algunas implementaciones, el rango de elevación esperado para una duplicación heterocigota se determina como de 1,65 a 1,35 cuando la elevación esperada para una

duplicación heterocigota es 1,5, *n* = 3, y el valor de incertidumbre *o* es +/- 0,05. En algunas implementaciones el rango de elevación esperado para una deleción heterocigota se determina como de 0,65 a 0,35 cuando la elevación esperada para una duplicación heterocigota es 0,5, *n* = 3, y el valor de incertidumbre *o* es +/- 0,05. En algunas implementaciones el rango de elevación esperado para una duplicación homocigota se determina como de 2,15 a 1,85 cuando la elevación esperada para una duplicación heterocigota es de 2,0, *n* = 3 y el valor de incertidumbre σ es +/- 0,05. En algunas implementaciones el rango de elevación esperado para una deleción homocigota se determina como de 0,15 a -0,15 cuando la elevación esperada para una duplicación heterocigota es de 0,0, *n* = 3 y el valor de incertidumbre *o* es +/- 0,05.

Algunas veces se determina un rango de elevación esperado para una variación del número de copias homocigotas (por ejemplo, una variación del número de copias homocigotas materno, fetal o materno y fetal), en parte, según un rango de elevación esperado para una variación del número de copias heterocigotas correspondiente. Por ejemplo, algunas veces un rango de elevación esperado para una duplicación homocigota comprende todos los valores mayores que un límite superior de un rango de elevación esperado para una duplicación heterocigota. Algunas veces, un rango de elevación esperado para una duplicación homocigota comprende todos los valores mayores que o iguales a un límite superior de un rango de elevación esperado para una duplicación heterocigota. Algunas veces, un rango de elevación esperado para una duplicación homocigota comprende todos los valores mayores que un límite superior de un rango de elevación esperado para una duplicación heterocigota y menores que el límite superior definido por la fórmula R, donde o es un valor de incertidumbre y es un valor de positivo, *n* es una constante y *k* es una duplicación homocigota. Algunas veces, un rango de elevación esperado para una duplicación homocigota comprende todos los valores mayores que o iguales a un límite superior de un rango de elevación esperado para una duplicación heterocigota y menores que o iguales al límite superior definido por la fórmula R, donde o es un valor de incertidumbre, o es un valor de positivo, *n* es una constante y *k* es una duplicación homocigota.

En algunas implementaciones, un rango de elevación esperado para una deleción homocigota comprende todos los valores menores que un límite inferior de un rango de elevación esperado para una deleción heterocigota. Algunas veces, un rango de elevación esperado para una deleción homocigota comprende todos los valores menores que o iguales a un límite inferior de un rango de elevación esperado para una deleción heterocigota. Algunas veces, un rango de elevación esperado para una deleción homocigota comprende todos los valores menores que un límite inferior de un rango de elevación esperado para una deleción heterocigota y mayores que el límite inferior definido por la fórmula R, donde o es un valor de incertidumbre, o es un valor negativo, *n* es una constante y *k* es una deleción homocigota. Algunas veces, un rango de elevación esperado para una deleción homocigota comprende todos los valores menores que o iguales a un límite inferior de un rango de elevación esperado para una deleción heterocigota y mayores que o iguales al límite inferior definido por la fórmula R, donde o es un valor de incertidumbre, o es un valor negativo, *n* es una constante y *k* es una deleción homocigota.

Puede usarse un valor de incertidumbre para determinar un valor umbral. En algunas implementaciones, un rango (por ejemplo, un rango umbral) se obtiene calculando el valor de incertidumbre determinado a partir de recuentos sin procesar, filtrados y/o normalizados. Un rango puede determinarse multiplicando el valor de incertidumbre para una elevación (por ejemplo, recuentos normalizados de una elevación) por una constante predeterminada (por ejemplo, 1, 2, 3, 4, 5, 6, etc.) que representa el múltiplo de incertidumbre (por ejemplo, número de desviaciones estándar) elegido como umbral de punto de corte (por ejemplo, multiplicar por 3 para 3 desviaciones estándar), mediante lo cual se genera un rango, en algunas implementaciones. Un rango puede determinarse sumando y/o restando un valor (por ejemplo, un valor predeterminado, un valor de incertidumbre, un valor de incertidumbre multiplicado por una constante predeterminada) a y/o de una elevación mediante lo cual se genera un rango, en algunas implementaciones. Por ejemplo, para una elevación igual a 1, una desviación estándar de +/-0,2, en la que una constante predeterminada es 3, el intervalo puede calcularse como (1 + 3(0,2)) a (1 + 3(-0,2)), o de 1,6 a 0,4. Un rango a veces puede definir un rango esperado o rango de elevación esperado para una variación del número de copias. En determinadas implementaciones, algunas o todas las secciones genómicas que superan un valor umbral, que se encuentran fuera de un rango o que se encuentran dentro de un rango de valores, se eliminan como parte de, antes de, o después de un procedimiento de normalización. En algunas implementaciones, algunas o todas las secciones genómicas que superan un valor umbral calculado, que se encuentran fuera de un rango o que se encuentran dentro de un rango se ponderan o ajustan como parte de, o antes del procedimiento de normalización o clasificación.

En el presente documento se describen ejemplos de ponderación. Las expresiones “datos redundantes”, y “lecturas mapeadas redundantes”, tal como se usan en el presente documento, se refieren a lecturas de secuencia derivadas de muestras que se identifican que ya se han asignado a una ubicación genómica (por ejemplo, posición de base) y/o contado para una sección genómica.

En algunas implementaciones, se determina un valor de incertidumbre según la siguiente fórmula:

$$Z = \frac{L_A - L_o}{\sqrt{\frac{\sigma_A^2}{N_A} + \frac{\sigma_o^2}{N_o}}}$$

donde Z representa la desviación normalizada entre dos elevaciones, L es la elevación media (o mediana) y sigma es la desviación estándar (o D.M.A). El subíndice O indica un segmento de un perfil (por ejemplo, una segunda

elevación, un cromosoma, un NRV, un “nivel euploide”, un nivel ausente de una variación del número de copias), y A indica otro segmento de un perfil (por ejemplo, una primera elevación, una elevación que representa una variación del número de copias, una elevación que representa una aneuploidía (por ejemplo, una trisomía). La variable N_o representa el número total de secciones genómicas en el segmento del perfil indicado por el subíndice O. N_A representa el número total de secciones genómicas en el segmento del perfil indicado por el subíndice a.

Categorización de una variación del número de copias

Una elevación (por ejemplo, una primera elevación) que difiere significativamente de otra elevación (por ejemplo, una segunda elevación) puede categorizarse a menudo como una variación del número de copias (por ejemplo, una variación del número de copias materno y/o fetal, una variación del número de copias fetal, una deleción, duplicación, inserción) según un rango de elevación esperado. En algunas implementaciones, la presencia de una variación del número de copias se categoriza cuando una primera elevación es significativamente diferente de una segunda elevación y la primera elevación se encuentra dentro del rango de elevación esperado para una variación del número de copias. Por ejemplo, una variación del número de copias (por ejemplo, una variación del número de copias materno y/o fetal, una variación del número de copias fetal) puede categorizarse cuando una primera elevación es significativamente diferente de una segunda elevación y la primera elevación se encuentra dentro del rango de elevación esperado para una variación del número de copias.

Algunas veces, una duplicación heterocigota (por ejemplo, una duplicación heterocigota materna o fetal, o materna y fetal, heterocigota) o deleción heterocigota (por ejemplo, una deleción materna o fetal, o materna y fetal, heterocigota) se categoriza cuando una primera elevación es significativamente diferente de una segunda elevación y la primera elevación se encuentra dentro del rango de elevación esperado para una duplicación heterocigota o deleción heterocigota, respectivamente. Algunas veces, una duplicación homocigota o deleción homocigota se categoriza cuando una primera elevación es significativamente diferente de una segunda elevación y la primera elevación se encuentra dentro del rango de elevación esperado para una duplicación homocigota o deleción homocigota, respectivamente.

Módulo de establecimiento de rango

Los rangos esperados (por ejemplo, rangos de elevación esperados) para diversas variaciones del número de copias (por ejemplo, duplicaciones, inserciones y/o deleciones) o rangos para la ausencia de una variación del número de copias pueden proporcionarse por un módulo de establecimiento de rango o por un aparato que comprende un módulo de establecimiento de rango. En algunos casos, las elevaciones esperadas son proporcionadas por un módulo de establecimiento de rango o por un aparato que comprende un módulo de establecimiento de rango. En algunas implementaciones, se requiere un módulo de establecimiento de rango o un aparato que comprende un módulo de establecimiento de rango para proporcionar elevaciones y/o rangos esperados. Algunas veces, un módulo de establecimiento de rango reúne, ensambla y/o recibe información y/o datos de otro módulo o aparato. Algunas veces, un módulo de establecimiento de rango o un aparato que comprende un módulo de establecimiento de rango proporciona y/o transfiere información y/o datos a otro módulo o aparato. Algunas veces, un módulo de establecimiento de rango acepta y recopila información y/o datos de un componente o periférico. A menudo, un módulo de establecimiento de rango reúne y ensambla elevaciones, elevaciones de referencia, valores de incertidumbre y/o constantes. Algunas veces, un módulo de establecimiento de rango acepta y recopila información y/o datos de entrada de un operador de un aparato. Por ejemplo, algunas veces un operador de un aparato proporciona una constante, un valor umbral, una fórmula o un valor predeterminado a un módulo. Un aparato que comprende un módulo de establecimiento de rango puede comprender al menos un procesador. En algunas implementaciones, las elevaciones esperadas y los rangos esperados se proporcionan por un aparato que incluye un procesador (por ejemplo, uno o más procesadores), procesador que puede realizar y/o implementar una o más instrucciones (por ejemplo, procesos, rutinas y/o subrutinas) del módulo de establecimiento de rango. En algunas implementaciones, las elevaciones y los rangos esperados se proporcionan por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunas implementaciones, un módulo de establecimiento de rango funciona con uno o más procesadores externos (por ejemplo, una red interna o externa, un servidor, dispositivo de almacenamiento y/o una red de almacenamiento (por ejemplo, una nube)). En algunas implementaciones, los rangos esperados se proporcionan por un aparato que comprende un componente o periférico adecuado. Un módulo de establecimiento de rango puede recibir datos normalizados de un módulo de normalización o datos de comparación de un módulo de comparación. La información y/o los datos derivados de o transformados por un módulo de establecimiento de rango (por ejemplo, rangos establecidos, límites de rango, rangos de elevación esperados, umbrales y/o rangos de umbral) pueden transferirse de un módulo de establecimiento de rango a un módulo de ajuste, un módulo de resultados, un módulo de categorización, módulo de representación gráfica u otro aparato y/o módulo adecuado.

Módulo de categorización

Una variación del número de copias (por ejemplo, una variación del número de copias materno y/o fetal, una variación del número de copias fetal, una duplicación, inserción, deleción) puede categorizarse por un módulo de categorización o por un aparato que comprende un módulo de categorización. A veces, una variación del número de copias (por ejemplo, una variación del número de copias materno y/o fetal) se categoriza por un módulo de categorización.

A veces, una elevación (por ejemplo, una primera elevación) que se determina que es significativamente diferente de otra elevación (por ejemplo, una segunda elevación) se identifica como representativa de una variación del número de copias por un módulo de categorización. A veces, la ausencia de una variación del número de copias se determina por un módulo de categorización. En algunas implementaciones, una determinación de una variación del número de copias puede determinarse por un aparato que comprende un módulo de categorización. Un módulo de categorización puede especializarse para categorizar una variación del número de copias materno y/o fetal, una variación del número de copias fetal, una duplicación, deleción o inserción o la falta de los mismos o una combinación de los anteriores. Por ejemplo, un módulo de categorización que identifica una deleción materna puede ser diferente y/o distinto de un módulo de categorización que identifica una duplicación fetal. En algunas implementaciones, se requiere un módulo de categorización o un aparato que comprende un módulo de categorización para identificar una variación del número de copias o un determinante del resultado de una variación del número de copias. Un aparato que comprende un módulo de categorización puede comprender al menos un procesador. En algunas implementaciones, una variación del número de copias o un determinante del resultado de una variación del número de copias se categoriza por un aparato que incluye un procesador (por ejemplo, uno o más procesadores), procesador que puede realizar y/o implementar una o más instrucciones (por ejemplo, procesos, rutinas y/o subrutinas) desde el módulo de categorización. En algunas implementaciones, una variación del número de copias o un determinante del resultado de una variación del número de copias se categoriza por un aparato que puede incluir múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunas implementaciones, un módulo de categorización funciona con uno o más procesadores externos (por ejemplo, una red interna o externa, un servidor, dispositivo de almacenamiento y/o una red de almacenamiento (por ejemplo, una nube)).

Algunas veces, un módulo de categorización transfiere o recibe y/o recopila información y/o datos a o desde un componente o periférico. A menudo, un módulo de categorización recibe, reúne y/o ensambla recuentos, elevaciones, perfiles, información y/o datos normalizados, elevaciones de referencia, elevaciones esperadas, rangos esperados, valores de incertidumbre, ajustes, elevaciones ajustadas, representaciones gráficas, comparaciones y/o constantes. Algunas veces, un módulo de categorización acepta y recopila información y/o datos de entrada de un operador de un aparato. Por ejemplo, algunas veces un operador de un aparato proporciona una constante, un valor umbral, una fórmula o un valor predeterminado a un módulo. En algunas implementaciones, se proporcionan información y/o datos por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunas implementaciones, la identificación o categorización de una variación del número de copias o un determinante del resultado de una variación del número de copias se proporciona por un aparato que comprende un componente o periférico adecuado. Algunas veces, un módulo de categorización recopila, ensambla y/o recibe información y/o datos de otro módulo o aparato. Un módulo de categorización puede recibir datos normalizados de un módulo de normalización, elevaciones y/o rangos esperados de un módulo de establecimiento de rango, datos de comparación de un módulo de comparación, representaciones gráficas de un módulo de representación gráfica y/o datos de ajuste de un módulo de ajuste. Un módulo de categorización puede transformar información y/o datos que recibe en una determinación de la presencia o ausencia de una variación del número de copias. Un módulo de categorización puede transformar información y/o datos que recibe en una determinación de que una elevación representa una sección genómica que comprende una variación del número de copias o un tipo específico de variación del número de copias (por ejemplo, una deleción homocigota materna). La información y/o los datos relacionados con una variación del número de copias o un determinante del resultado de una variación del número de copias pueden transferirse desde un módulo de categorización a un aparato y/o módulo adecuado. Una variación del número de copias o un determinante del resultado de una variación del número de copias categorizados mediante los métodos descritos en el presente documento puede verificarse independientemente mediante pruebas adicionales (por ejemplo, mediante secuenciación dirigida de ácido nucleico materno y/o fetal).

Determinación de fracción fetal basándose en la elevación

En algunas implementaciones, se determina una fracción fetal según una elevación categorizada como representativa de una variación del número de copias materno y/o fetal. Por ejemplo, determinar la fracción fetal comprende a menudo evaluar una elevación esperada para una variación del número de copias materno y/o fetal utilizada para la determinación de fracción fetal. Algunas veces se determina una fracción fetal para una elevación (por ejemplo, una primera elevación) categorizada como representativa de una variación del número de copias según un rango de elevación esperado determinado para el mismo tipo de variación del número de copias. A menudo, se determina una fracción fetal según una elevación observada que se encuentra dentro de un rango de elevación esperado y, de ese modo, se categoriza como una variación del número de copias materno y/o fetal. Algunas veces, se determina una fracción fetal cuando una elevación observada (por ejemplo, una primera elevación) categorizada como una variación del número de copias materno y/o fetal es diferente de la elevación esperada determinada para la misma variación del número de copias materno y/o fetal.

En algunas implementaciones, una elevación (por ejemplo, una primera elevación, una elevación observada), es significativamente diferente de una segunda elevación, la primera elevación se categoriza como una variación del número de copias materno y/o fetal, y se determina una fracción fetal según la primera elevación. A veces, una primera elevación es una elevación observada y/u obtenida experimentalmente que es significativamente diferente de una segunda elevación en un perfil y se determina una fracción fetal según la primera elevación. Algunas veces la primera elevación es una elevación promedio, media o sumada y se determina una fracción fetal según la primera elevación. En algunos casos, una primera elevación y una segunda elevación son elevaciones observadas y/u obtenidas experimentalmente y se determina una fracción fetal según la primera elevación. En algunos casos, una primera elevación comprende recuentos normalizados para

un primer conjunto de secciones genómicas y una segunda elevación comprende recuentos normalizados para un segundo conjunto de secciones genómicas y se determina una fracción fetal según la primera elevación.

Algunas veces, un primer conjunto de secciones genómicas de una primera elevación incluye una variación del número de copias (por ejemplo, la primera elevación es representativa de una variación del número de copias) y se determina una fracción fetal según la primera elevación. Algunas veces, el primer conjunto de secciones genómicas de una primera elevación incluye una variación del número de copias homocigotas o heterocigotas materno y se determina una fracción fetal según la primera elevación. A veces, un perfil comprende una primera elevación para un primer conjunto de secciones genómicas y una segunda elevación para un segundo conjunto de secciones genómicas, el segundo conjunto de secciones genómicas no incluye sustancialmente ninguna variación del número de copias (por ejemplo, una variación del número de copias materno, variación del número de copias fetal, o una variación del número de copias materno y una variación del número de copias fetal) y se determina una fracción fetal según la primera elevación.

En algunas implementaciones, una elevación (por ejemplo, una primera elevación, una elevación observada), es significativamente diferente de una segunda elevación, la primera elevación se categoriza en cuanto a una variación del número de copias materno y/o fetal, y se determina una fracción fetal según la primera elevación y/o una elevación esperada de la variación del número de copias. Algunas veces, una primera elevación se categoriza en cuanto a una variación del número de copias según una elevación esperada para una variación del número de copias y se determina una fracción fetal según una diferencia entre la primera elevación y la elevación esperada. En algunos casos, una elevación (por ejemplo, una primera elevación, una elevación observada) se categoriza como una variación del número de copias materno y/o fetal, y se determina una fracción fetal como el doble de la diferencia entre la primera elevación y la elevación esperada de la variación del número de copias. Algunas veces, una elevación (por ejemplo, una primera elevación, una elevación observada) se categoriza como una variación del número de copias materno y/o fetal, la primera elevación se resta de la elevación esperada, proporcionando de ese modo una diferencia, y se determina una fracción fetal como el doble de la diferencia.

Algunas veces una elevación (por ejemplo, una primera elevación, una elevación observada) se categoriza como una variación del número de copias materno y/o fetal, una elevación esperada se resta de una primera elevación proporcionando de ese modo una diferencia, y se determina la fracción fetal como el doble de la diferencia.

A menudo, se proporciona una fracción fetal como un porcentaje. Por ejemplo, una fracción fetal puede dividirse entre 100 proporcionando de ese modo un valor porcentual. Por ejemplo, para una primera elevación representativa de una duplicación homocigota materna y que tiene una elevación de 155 y una elevación esperada para una duplicación homocigota materna que tiene una elevación de 150, puede determinarse una fracción fetal como del 10 % (por ejemplo, (fracción fetal = 2 x (155 - 150)).

En algunas implementaciones, se determina una fracción fetal a partir de dos o más elevaciones dentro de un perfil que se categorizan como variaciones del número de copias. Por ejemplo, a veces dos o más elevaciones (por ejemplo, dos o más primeras elevaciones) en un perfil se identifican como significativamente diferentes de una elevación de referencia (por ejemplo, una segunda elevación, una elevación que no incluye sustancialmente ninguna variación del número de copias), las dos o más elevaciones se categorizan como representativas de una variación del número de copias materno y/o fetal y se determina una fracción fetal a partir de cada una de las dos o más elevaciones. Algunas veces, se determina una fracción fetal a partir de aproximadamente 3 o más, aproximadamente 4 o más, aproximadamente 5 o más, aproximadamente 6 o más, aproximadamente 7 o más, aproximadamente 8 o más, o aproximadamente 9 o más determinaciones de fracción fetal dentro de un perfil. Algunas veces, se determina una fracción fetal a partir de aproximadamente 10 o más, aproximadamente 20 o más, aproximadamente 30 o más, aproximadamente 40 o más, aproximadamente 50 o más, aproximadamente 60 o más, aproximadamente 70 o más, aproximadamente 80 o más, o aproximadamente 90 o más determinaciones de fracción fetal dentro de un perfil. Algunas veces, se determina una fracción fetal a partir de aproximadamente 100 o más, aproximadamente 200 o más, aproximadamente 300 o más, aproximadamente 400 o más, aproximadamente 500 o más, aproximadamente 600 o más, aproximadamente 700 o más, aproximadamente 800 o más, aproximadamente 900 o más, o aproximadamente 1000 o más determinaciones de fracción fetal dentro de un perfil. Algunas veces, se determina una fracción fetal a partir de aproximadamente 10 a aproximadamente 1000, de aproximadamente 20 a aproximadamente 900, de aproximadamente 30 a aproximadamente 700, de aproximadamente 40 a aproximadamente 600, de aproximadamente 50 a aproximadamente 500, de aproximadamente 50 a aproximadamente 400, de aproximadamente 50 a aproximadamente 300, de aproximadamente 50 a aproximadamente 200 o de aproximadamente 50 a aproximadamente 100 determinaciones de fracción fetal dentro de un perfil.

En algunas implementaciones, se determina una fracción fetal como el promedio o la media de múltiples determinaciones de fracciones fetales dentro de un perfil. En algunos casos, una fracción fetal determinada a partir de múltiples determinaciones de fracciones fetales es una media (por ejemplo, un promedio, un promedio estándar, una mediana, o similares) de múltiples determinaciones de fracción fetal. A menudo, una fracción fetal determinada a partir de múltiples determinaciones de fracción fetal es un valor medio determinado mediante un método adecuado conocido en la técnica o descrito en el presente documento. Algunas veces, un valor medio de una determinación de fracción fetal es una media ponderada. Algunas veces, un valor medio de una determinación de fracción fetal es una media no ponderada. Una determinación de fracción fetal media, en mediana o promedio (es decir, un valor medio, de mediana o promedio de determinación de fracción fetal) generada a partir de múltiples determinaciones de fracción fetal está algunas veces

asociada con un valor de incertidumbre (por ejemplo, una varianza, desviación estándar, D.M.A., o similares). Antes de determinar un valor medio, de mediana o promedio de fracción fetal a partir de múltiples determinaciones, se eliminan una o más determinaciones desviadas en algunas implementaciones (descritas con mayor detalle en el presente documento).

Algunas determinaciones de fracción fetal dentro de un perfil a veces no se incluyen en la determinación global de una fracción fetal (por ejemplo, determinación media o promedio de fracción fetal). Algunas veces una determinación de fracción fetal se deriva de una primera elevación (por ejemplo, una primera elevación que es significativamente diferente de una segunda elevación) en un perfil y la primera elevación no es indicativa de una variación genética. Por ejemplo, algunas primeras elevaciones (por ejemplo, aumentos bruscos o depresiones) en un perfil se generan a partir de anomalías o causas desconocidas. Tales valores generan a menudo determinaciones de fracción fetal que difieren significativamente de otras determinaciones de fracción fetal obtenidas a partir de variaciones verdaderas del número de copias. Algunas veces, las determinaciones de fracción fetal que difieren significativamente de otras determinaciones de fracción fetal en un perfil se identifican y eliminan de una determinación de fracción fetal. Por ejemplo, algunas determinaciones de fracción fetal obtenidas a partir de depresiones y aumentos bruscos anómalos se identifican comparándolas con otras determinaciones de fracción fetal dentro de un perfil y se excluyen de la determinación global de fracción fetal.

Algunas veces, una determinación de fracción fetal independiente que difiere significativamente de una determinación de fracción fetal media, en mediana o promedio es una diferencia identificada, reconocida y/u observable. En algunos casos, la expresión "difiere significativamente" puede significar una diferencia estadísticamente diferente y/o una diferencia estadísticamente significativa. Una determinación de fracción fetal "independiente" puede ser una fracción fetal determinada (por ejemplo, en algunos casos, una determinación individual) a partir de una elevación específica categorizada como una variación del número de copias. Puede usarse cualquier umbral o rango adecuado para determinar que una determinación de fracción fetal difiere significativamente de una determinación de fracción fetal media, en mediana o promedio. En algunos casos, una determinación de fracción fetal difiere significativamente de una determinación de fracción fetal media, en mediana o promedio y la determinación puede expresarse como una desviación porcentual del valor promedio o medio. En algunos casos, una determinación de fracción fetal que difiere significativamente de una determinación de fracción fetal media, en mediana o promedio difiere en aproximadamente el 10 por ciento o más. Algunas veces, una determinación de fracción fetal que difiere significativamente de una determinación de fracción fetal media, en mediana o promedio difiere en aproximadamente el 15 por ciento o más. Algunas veces, una determinación de fracción fetal que difiere significativamente de una determinación de fracción fetal media, en mediana o promedio difiere en aproximadamente el 15 % a aproximadamente el 100 % o más.

En algunos casos, una determinación de fracción fetal difiere significativamente de una determinación de fracción fetal media, en mediana o promedio según un múltiplo de un valor de incertidumbre asociado con la determinación media o promedio de fracción fetal. A menudo, un valor de incertidumbre y constante *n* (por ejemplo, un intervalo de confianza) define un rango (por ejemplo, un punto de corte de incertidumbre). Por ejemplo, algunas veces un valor de incertidumbre es una desviación estándar para las determinaciones de fracción fetal (por ejemplo, +/- 5) y se multiplica por una constante *n* (por ejemplo, un intervalo de confianza) definiendo de ese modo un rango o punto de corte de incertidumbre (por ejemplo, de 5*n* a -5*n*, algunas veces denominado 5 sigma). Algunas veces, una determinación de fracción fetal independiente se encuentra fuera de un rango definido por el punto de corte de incertidumbre y se considera significativamente diferente de una determinación de fracción fetal media, en mediana o promedio. Por ejemplo, para un valor medio de 10 y un punto de corte de incertidumbre de 3, una fracción fetal independiente mayor de 13 o menor de 7 es significativamente diferente. Algunas veces, una determinación de fracción fetal que difiere significativamente de una determinación de fracción fetal media, en mediana o promedio difiere en más de *n* veces el valor de incertidumbre (por ejemplo, *n* x sigma), donde *n* es aproximadamente igual a o mayor de 1, 2, 3, 4, 5, 6, 7, 8, 9 o 10. Algunas veces, una determinación de fracción fetal que difiere significativamente de una determinación de fracción fetal media, en mediana o promedio difiere en más de *n* veces el valor de incertidumbre (por ejemplo, *n* x sigma), donde *n* es aproximadamente igual a o mayor de 1,1, 1,2, 1,3, 1,4, 1,5, 1,6, 1,7, 1,8, 1,9, 2,0, 2,1, 2,2, 2,3, 2,4, 2,5, 2,6, 2,7, 2,8, 2,9, 3,0, 3,1, 3,2, 3,3, 3,4, 3,5, 3,6, 3,7, 3,8, 3,9 o 4,0.

En algunas implementaciones, una elevación es representativa de una microploidía fetal y/o materna. Algunas veces una elevación (por ejemplo, una primera elevación, una elevación observada), es significativamente diferente de una segunda elevación, la primera elevación se categoriza como una variación del número de copias materno y/o fetal, y la primera elevación y/o la segunda elevación son representativas de una microploidía fetal y/o una microploidía materna. En algunos casos una primera elevación es representativa de una microploidía fetal, Algunas veces una primera elevación es representativa de una microploidía materna. A menudo, una primera elevación es representativa de una microploidía fetal y una microploidía materna. Algunas veces una elevación (por ejemplo, una primera elevación, una elevación observada), es significativamente diferente de una segunda elevación, la primera elevación se categoriza como una variación del número de copias materno y/o fetal, la primera elevación es representativa de una microploidía fetal y/o materna y se determina una fracción fetal según la microploidía fetal y/o materna. En algunos casos, una primera elevación se categoriza como una variación del número de copias materno y/o fetal, la primera elevación es representativa de una microploidía fetal y se determina una fracción fetal según la microploidía fetal. Algunas veces, una primera elevación se categoriza como una variación del número de copias materno y/o fetal, la primera elevación es representativa de una microploidía materna y se determina una fracción fetal según la microploidía materna. Algunas veces una primera elevación se categoriza como una variación del número de copias materno y/o fetal, la primera elevación es representativa de una microploidía materna y una fetal y se determina una fracción fetal según la microploidía materna y fetal.

En algunas implementaciones, una determinación de una fracción fetal comprende determinar una microploidía fetal y/o materna. Algunas veces una elevación (por ejemplo, una primera elevación, una elevación observada), es significativamente diferente de una segunda elevación, la primera elevación se categoriza como una variación del número de copias materno y/o fetal, una microploidía fetal y/o materna se determina según la primera elevación y/o la segunda elevación y se determina una fracción fetal. Algunas veces, una primera elevación se categoriza como una variación del número de copias materno y/o fetal, una microploidía fetal se determina según la primera elevación y/o la segunda elevación y se determina una fracción fetal según la microploidía fetal. En algunos casos, una primera elevación se categoriza como una variación del número de copias materno y/o fetal, una microploidía materna se determina según la primera elevación y/o la segunda elevación y se determina una fracción fetal según la microploidía materna. Algunas veces una primera elevación se categoriza como una variación del número de copias materno y/o fetal, una microploidía materna y fetal se determina según la primera elevación y/o la segunda elevación y se determina una fracción fetal según la microploidía materna y fetal.

A menudo, se determina una fracción fetal cuando la microploidía de la madre es diferente de (por ejemplo, no igual a) la microploidía del feto para una elevación dada o para una elevación categorizada como una variación del número de copias. Algunas veces se determina una fracción fetal cuando la madre es homocigota para una duplicación (por ejemplo, una microploidía de 2) y el feto es heterocigota para la misma duplicación (por ejemplo, una microploidía de 1,5). Algunas veces se determina una fracción fetal cuando la madre es heterocigota para una duplicación (por ejemplo, una microploidía de 1,5) y el feto es homocigoto para la misma duplicación (por ejemplo, una microploidía de 2) o la duplicación está ausente en el feto (por ejemplo, una microploidía de 1). Algunas veces se determina una fracción fetal cuando la madre es homocigota para una deleción (por ejemplo, una microploidía de 0) y el feto es heterocigoto para la misma deleción (por ejemplo, una microploidía de 0,5). Algunas veces se determina una fracción fetal cuando la madre es heterocigota para una deleción (por ejemplo, una microploidía de 0,5) y el feto es homocigoto para la misma deleción (por ejemplo, una microploidía de 0) o la deleción está ausente en el feto (por ejemplo, una microploidía de 1).

En algunos casos, no puede determinarse una fracción fetal cuando la microploidía de la madre es la misma (por ejemplo, identificada como la misma) que la microploidía del feto para una elevación dada identificada como una variación del número de copias. Por ejemplo, para una elevación dada en la que tanto la madre como el feto portan el mismo número de copias de una variación del número de copias, no se determina una fracción fetal, en algunas implementaciones. Por ejemplo, no puede determinarse una fracción fetal para una elevación categorizada como una variación del número de copias cuando tanto la madre como el feto son homocigotos para la misma deleción u homocigotos para la misma duplicación. En algunos casos, no puede determinarse una fracción fetal para una elevación categorizada como una variación del número de copias cuando tanto la madre como el feto son heterocigotos para la misma deleción o heterocigotos para la misma duplicación. En implementaciones en las que se realizan múltiples determinaciones de fracción fetal para una muestra, las determinaciones que se desvían significativamente de una media, mediana o valor promedio pueden resultar de una variación del número de copias para la cual la ploidía materna es igual a la ploidía fetal, y tales determinaciones pueden eliminarse de la consideración.

En algunas implementaciones se desconoce la microploidía de una variación del número de copias materno y la variación del número de copias fetal. A veces, en casos en los que no hay determinación de microploidía fetal y/o materna para una variación del número de copias, se genera una fracción fetal y se compara con una determinación de fracción fetal media, en mediana o promedio. Una determinación de fracción fetal para una variación del número de copias que difiere significativamente de una determinación de fracción fetal media, en mediana o promedio se debe a que a veces la microploidía de la madre y el feto son las mismas para la variación del número de copias. Una determinación de fracción fetal que difiere significativamente de una determinación de fracción fetal media, en mediana o promedio se excluye a menudo de una determinación global de fracción fetal independientemente de la fuente o causa de la diferencia. En algunas implementaciones, la microploidía de la madre y/o el feto se determina y/o verifica mediante un método conocido en la técnica (por ejemplo, mediante métodos de secuenciación dirigida).

Ajustes de elevación

En algunas implementaciones, se ajustan una o más elevaciones. Un procedimiento para ajustar una elevación se denomina a menudo relleno. En algunas implementaciones, se ajustan múltiples elevaciones en un perfil (por ejemplo, un perfil de un genoma, un perfil cromosómico, un perfil de una porción o un segmento de un cromosoma). A veces, se ajustan de aproximadamente 1 a aproximadamente 10.000 o más elevaciones en un perfil. Algunas veces, se ajustan de aproximadamente 1 a aproximadamente 1000, de 1 a aproximadamente 900, de 1 a aproximadamente 800, de 1 a aproximadamente 700, de 1 a aproximadamente 600, de 1 a aproximadamente 500, de 1 a aproximadamente 400, de 1 a aproximadamente 300, de 1 a aproximadamente 200, de 1 a aproximadamente 100, de 1 a aproximadamente 50, de 1 a aproximadamente 25, de 1 a aproximadamente 20, de 1 a aproximadamente 15, de 1 a aproximadamente 10 o de 1 a aproximadamente 5 elevaciones en un perfil. A veces se ajusta una elevación. En algunas implementaciones, se ajusta una elevación (por ejemplo, una primera elevación de un perfil de recuento normalizado) que difiere significativamente de una segunda elevación. Algunas veces se ajusta una elevación clasificada como variación del número de copias.

A veces, una elevación (por ejemplo, una primera elevación de un perfil de recuento normalizado) que difiere significativamente de una segunda elevación se categoriza como una variación del número de copias (por ejemplo, una variación del número de copias, por ejemplo, una variación del número de copias materno) y se ajusta. En algunas

implementaciones, una elevación (por ejemplo, una primera elevación) está dentro de un rango de elevación esperado para una variación del número de copias materno, la variación del número de copias fetal, o una variación del número de copias materno y una variación del número de copias fetal y se ajusta la elevación. A veces, no se ajustan una o más elevaciones (por ejemplo, elevaciones en un perfil). En algunas implementaciones, una elevación (por ejemplo, una primera elevación) está fuera de un rango de elevación esperado para una variación del número de copias y la elevación no se ajusta. A menudo, no se ajusta una elevación dentro de un rango de elevación esperado para la ausencia de una variación del número de copias. Puede realizarse cualquier número adecuado de ajustes a una o más elevaciones en un perfil. En algunas implementaciones, se ajustan una o más elevaciones. Algunas veces, se ajustan 2 o más, 3 o más, 5 o más, 6 o más, 7 o más, 8 o más, 9 o más y, algunas veces, 10 o más elevaciones.

En algunas implementaciones, un valor de una primera elevación se ajusta según un valor de una segunda elevación. Algunas veces, una primera elevación, identificada como representativa de una variación del número de copias, se ajusta al valor de una segunda elevación, en la que la segunda elevación se asocia a menudo con la ausencia de variación del número de copias. En algunos casos, un valor de una primera elevación, identificado como representativo de una variación del número de copias, se ajusta de modo que el valor de la primera elevación es aproximadamente igual a un valor de una segunda elevación.

Un ajuste puede comprender una operación matemática adecuada. Algunas veces, un ajuste comprende una o más operaciones matemáticas. Algunas veces, se ajusta una elevación mediante normalización, filtrado, promediado, multiplicación, división, suma o resta o combinación de los mismos. Algunas veces, se ajusta una elevación mediante un valor predeterminado o una constante. Algunas veces, se ajusta una elevación modificando el valor de la elevación al valor de otra elevación. Por ejemplo, una primera elevación puede ajustarse modificando su valor al valor de una segunda elevación. Un valor en tales casos puede ser un valor procesado (por ejemplo, valor medio, normalizado y similares).

Algunas veces, una elevación se categoriza como una variación del número de copias (por ejemplo, una variación del número de copias materno) y se ajusta según un valor predeterminado denominado en el presente documento valor de ajuste predeterminado (PAV). A menudo, se determina un PAV para una variación específica del número de copias. A menudo, un PAV determinado para una variación del número de copias específica (por ejemplo, duplicación homocigota, deleción homocigota, duplicación heterocigota, deleción heterocigota) se usa para ajustar una elevación categorizada como una variación del número de copias específica (por ejemplo, duplicación homocigota, deleción homocigota, duplicación heterocigota, deleción heterocigota). En algunos casos, una elevación se categoriza como una variación del número de copias y, después, se ajusta según un PAV específico para el tipo de variación del número de copias categorizada. Algunas veces, una elevación (por ejemplo, una primera elevación) se categoriza como una variación del número de copias materno, variación del número de copias fetal, o una variación del número de copias materno y una variación del número de copias fetal y se ajusta sumando o restando un PAV de la elevación. A menudo, una elevación (por ejemplo, una primera elevación) se categoriza como una variación del número de copias materno y se ajusta mediante la suma de un PAV a la elevación. Por ejemplo, una elevación categorizada como una duplicación (por ejemplo, una duplicación homocigota materna, fetal o materna y fetal) puede ajustarse sumando un PAV determinado para una duplicación específica (por ejemplo, una duplicación homocigota) proporcionando de ese modo una elevación ajustada. A menudo, un PAV determinado para una duplicación del número de copias es un valor negativo. En algunas implementaciones, proporcionar un ajuste a una elevación representativa de una duplicación usando un PAV determinado para una duplicación da como resultado una reducción del valor de la elevación. En algunas implementaciones, una elevación (por ejemplo, una primera elevación) que difiere significativamente de una segunda elevación se categoriza como una deleción del número de copias (por ejemplo, una deleción homocigota, deleción heterocigota, duplicación homocigota, duplicación homocigota) y la primera elevación se ajusta mediante la suma de un PAV determinado para una deleción del número de copias. A menudo, un PAV determinado para una deleción del número de copias es un valor de positivo. En algunas implementaciones el proporcionar un ajuste a una elevación representativa de una deleción utilizando un PAV determinado para una deleción da como resultado un aumento del valor de la elevación.

Un PAV puede ser cualquier valor adecuado. A menudo, un PAV se determina según una variación del número de copias y es específico para esta (por ejemplo, una variación del número de copias categorizada). En algunos casos, un PAV se determina según una elevación esperada para una variación del número de copias (por ejemplo, una variación del número de copias categorizada) y/o un factor de PAV. A veces, un PAV se determina multiplicando una elevación esperada por un factor de PAV. Por ejemplo, un PAV para una variación del número de copias puede determinarse multiplicando una elevación esperada determinada para una variación del número de copias (por ejemplo, una deleción heterocigota) por un factor de PAV determinado para la misma variación del número de copias (por ejemplo, una deleción heterocigota). Por ejemplo, un PAV puede determinarse mediante la siguiente fórmula:

$$PAV_k = (\text{Elevación Esperada})_k \times (\text{factor de PAV})_k$$

para la variación del número de copias k (p. ej., k = una deleción heterocigota)

Un factor de PAV puede ser cualquier valor adecuado. Algunas veces, un factor de PAV para una duplicación homocigota es de entre aproximadamente -0,6 y aproximadamente -0,4. Algunas veces un factor de PAV para una duplicación homocigota es de aproximadamente -0,60, -0,59, -0,58, -0,57, -0,56, -0,55, -0,54, -0,53, -0,52, -0,51,

-0,50, -0,49, -0,48, -0,47, -0,46, -0,45, -0,44, -0,43, -0,42, -0,41 y -0,40. A menudo, un factor de PAV para una duplicación homocigota es de aproximadamente -0,5.

Por ejemplo, para un NRV de aproximadamente 1 y una elevación esperada de una duplicación homocigota igual a aproximadamente 2, el PAV para la duplicación homocigota se determina como de aproximadamente -1 según la fórmula anterior. En este caso, una primera elevación categorizada como una duplicación homocigota se ajusta sumando aproximadamente -1 al valor de la primera elevación, por ejemplo.

Algunas veces, un factor de PAV para una duplicación heterocigota es de entre aproximadamente -0,4 y aproximadamente -0,2. Algunas veces, un factor de PAV para una duplicación heterocigota es de aproximadamente -0,40, -0,39, -0,38, -0,37, -0,36, -0,35, -0,34, -0,33, -0,32, -0,31, -0,30, -0,29, -0,28, -0,27, -0,26, -0,25, -0,24, -0,23, -0,22, -0,21 y -0,20. A menudo, un factor de PAV para una duplicación heterocigota es de aproximadamente -0,33.

Por ejemplo, para un NRV de aproximadamente 1 y una elevación esperada de una duplicación heterocigota igual a aproximadamente 1,5, el PAV para la duplicación homocigota se determina como de aproximadamente -0,495 según la fórmula anterior. En este caso, una primera elevación categorizada como una duplicación heterocigota se ajusta sumando aproximadamente -0,495 al valor de la primera elevación, por ejemplo.

Algunas veces, un factor de PAV para una deleción heterocigota se encuentra entre aproximadamente 0,4 y aproximadamente 0,2. Algunas veces, un factor de PAV para una deleción heterocigota es de aproximadamente 0,40, 0,39, 0,38, 0,37, 0,36, 0,35, 0,34, 0,33, 0,32, 0,31, 0,30, 0,29, 0,28, 0,27, 0,26, 0,25, 0,24, 0,23, 0,22, 0,21 y 0,20. A menudo, un factor de PAV para una deleción heterocigota es de aproximadamente 0,33.

Por ejemplo, para un NRV de aproximadamente 1 y una elevación esperada de una deleción heterocigota igual a aproximadamente 0,5, el PAV para la deleción heterocigota se determina como de aproximadamente 0,495 según la fórmula anterior. En este caso, una primera elevación categorizada como una deleción heterocigota se ajusta sumando aproximadamente 0,495 al valor de la primera elevación, por ejemplo.

Algunas veces, un factor de PAV para una deleción homocigota se encuentra entre aproximadamente 0,6 y aproximadamente 0,4.

Algunas veces, un factor de PAV para una deleción homocigota es de aproximadamente 0,60, 0,59, 0,58, 0,57, 0,56, 0,55, 0,54, 0,53, 0,52, 0,51, 0,50, 0,49, 0,48, 0,47, 0,46, 0,45, 0,44, 0,43, 0,42, 0,41 y 0,40. A menudo, un factor de PAV para una deleción homocigota es de aproximadamente 0,5.

Por ejemplo, para un NRV de aproximadamente 1 y una elevación esperada de una deleción homocigota igual a aproximadamente 0, el PAV para la deleción homocigota se determina como de aproximadamente 1 según la fórmula anterior. En este caso, una primera elevación categorizada como una deleción homocigota se ajusta sumando aproximadamente 1 al valor de la primera elevación, por ejemplo.

En algunos casos, un PAV es aproximadamente igual o igual a una elevación esperada para una variación del número de copias (por ejemplo, la elevación esperada de una variación del número de copias).

En algunas implementaciones, los recuentos de una elevación se normalizan antes de realizar un ajuste. En algunos casos, los recuentos de algunas o todas las elevaciones en un perfil se normalizan antes de realizar un ajuste. Por ejemplo, los recuentos de una elevación pueden normalizarse según los recuentos de una elevación de referencia o un NRV. En algunos casos, los recuentos de una elevación (por ejemplo, una segunda elevación) se normalizan según los recuentos de una elevación de referencia o un NRV, y los recuentos de todas las demás elevaciones (por ejemplo, una primera elevación) en un perfil se normalizan en relación con los recuentos de la misma elevación de referencia o NRV antes de realizar un ajuste.

En algunas implementaciones, una elevación de un perfil resulta de uno o más ajustes. En algunos casos, una elevación de un perfil se determina después de ajustar una o más elevaciones en el perfil. En algunas implementaciones, una elevación de un perfil se calcula de nuevo después de realizar uno o más ajustes.

En algunas implementaciones, una variación del número de copias (por ejemplo, una variación del número de copias materno, una variación del número de copias fetal, o una variación del número de copias materno y una variación del número de copias fetal) se determina (por ejemplo, se determina directa o indirectamente) a partir de un ajuste. Por ejemplo, una elevación en un perfil que se ajustó (por ejemplo, una primera elevación ajustada) puede identificarse como una variación del número de copias materno. En algunas implementaciones, la magnitud del ajuste indica el tipo de variación del número de copias (por ejemplo, deleción heterocigota, duplicación homocigota, y similares). En algunos casos, una elevación ajustada en un perfil puede identificarse como representativa de una variación del número de copias según el valor de un PAV para la variación del número de copias. Por ejemplo, para un perfil dado, PAV es de aproximadamente -1 para una duplicación homocigota, aproximadamente -0,5 para una duplicación heterocigota, aproximadamente 0,5 para una deleción heterocigota y aproximadamente 1 para una deleción homocigota. En el ejemplo anterior, una elevación ajustada en aproximadamente -1 puede identificarse como una

duplicación homocigota, por ejemplo. En algunas implementaciones, una o más variaciones del número de copias pueden determinarse a partir de un perfil o una elevación que comprende uno o más ajustes.

En algunos casos, se comparan las elevaciones ajustadas dentro de un perfil. Algunas veces se identifican anomalías y errores comparando elevaciones ajustadas. Por ejemplo, a menudo se comparan una o más elevaciones ajustadas en un perfil y una elevación particular puede identificarse como una anomalía o un error. A veces se identifica una anomalía o un error dentro de una o más secciones genómicas que constituyen una elevación. Una anomalía o un error puede identificarse dentro de la misma elevación (por ejemplo, en un perfil) o en una o más elevaciones que representan secciones genómicas adyacentes, contiguas, limítrofes o aledañas. Algunas veces, una o más elevaciones ajustadas son elevaciones de secciones genómicas adyacentes, contiguas, limítrofes o aledañas en las que se comparan una o más elevaciones ajustadas y se identifica una anomalía o un error. Una anomalía o un error puede ser un pico o depresión en un perfil o elevación en el que se conoce o desconoce una causa del pico o la depresión. En algunos casos, se comparan elevaciones ajustadas y se identifica una anomalía o un error en el que la anomalía o el error se debe a un error estocástico, sistemático, aleatorio o de usuario. Algunas veces se comparan elevaciones ajustadas y se elimina una anomalía o un error de un perfil. En algunos casos, se comparan las elevaciones ajustadas y se ajusta una anomalía o un error.

Módulo de ajuste

En algunas implementaciones, los ajustes (por ejemplo, ajustes para elevaciones o perfiles) los realiza un módulo de ajuste o un aparato que comprende un módulo de ajuste. En algunas implementaciones, se requiere un módulo de ajuste o un aparato que comprende un módulo de ajuste para ajustar una elevación. Un aparato que comprende un módulo de ajuste puede comprender al menos un procesador. En algunas implementaciones, se proporciona una elevación ajustada por un aparato que incluye un procesador (por ejemplo, uno o más procesadores), procesador que puede realizar y/o implementar una o más instrucciones (por ejemplo, procesos, rutinas y/o subrutinas) del módulo de ajuste. En algunas implementaciones, se ajusta una elevación por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunas implementaciones, un módulo de ajuste funciona con uno o más procesadores externos (por ejemplo, una red interna o externa, un servidor, dispositivo de almacenamiento y/o una red de almacenamiento (por ejemplo, una nube)). Algunas veces, un aparato que comprende un módulo de ajuste recopila, ensambla y/o recibe información y/o datos de otro módulo o aparato. Algunas veces, un aparato que comprende un módulo de ajuste proporciona y/o transfiere información y/o datos a otro módulo o aparato.

Algunas veces, un módulo de ajuste recibe y recopila información y/o datos de un componente o periférico. A menudo, un módulo de ajuste recibe, recopila y/o ensambla recuentos, elevaciones, perfiles, elevaciones de referencia, elevaciones esperadas, rangos de elevación esperados, valores de incertidumbre, ajustes y/o constantes. A menudo, un módulo de ajuste recibe, recopila y/o ensambla elevaciones (por ejemplo, primeras elevaciones) que se categorizan o determinan como variaciones del número de copias (por ejemplo, una variación del número de copias materno, una variación del número de copias fetal, o una variación del número de copias materno y una variación del número de copias fetal). Algunas veces, un módulo de ajuste acepta y recopila información y/o datos de entrada de un operador de un aparato. Por ejemplo, algunas veces un operador de un aparato proporciona una constante, un valor umbral, una fórmula o un valor predeterminado a un módulo. En algunas implementaciones, se proporcionan información y/o datos por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunas implementaciones, se ajusta una elevación por un aparato que comprende un componente o periférico adecuado. Un aparato que comprende un módulo de ajuste puede recibir datos normalizados de un módulo de normalización, rangos de un módulo de establecimiento de rango, datos de comparación de un módulo de comparación, elevaciones identificadas (por ejemplo, identificadas como una variación del número de copias) de un módulo de categorización, y/o datos de ajuste de otro módulo de ajuste. Un módulo de ajuste puede recibir información y/o datos, transformar la información y/o los datos recibidos y proporcionar ajustes. La información y/o los datos derivados de, o transformados por, un módulo de ajuste puede transferirse de un módulo de ajuste a un módulo de categorización o a un aparato y/o módulo adecuado. Una elevación ajustada mediante los métodos descritos en el presente documento puede verificarse y/o ajustarse independientemente mediante pruebas adicionales (por ejemplo, mediante secuenciación dirigida de ácido nucleico materno y/o fetal).

Módulo de representación gráfica

En algunas implementaciones, se representa gráficamente un recuento, una elevación y/o un perfil (por ejemplo, se grafica). Algunas veces, una representación gráfica (por ejemplo, un gráfico) comprende un ajuste. Algunas veces, una representación gráfica comprende un ajuste de un recuento, una elevación y/o un perfil. A veces, se representa gráficamente un recuento, una elevación y/o un perfil y un recuento, una elevación y/o un perfil comprenden un ajuste. A menudo, se representa gráficamente un recuento, una elevación y/o un perfil y se comparan un recuento, una elevación y/o un perfil. Algunas veces, se identifica y/o categoriza una variación del número de copias (por ejemplo, una aneuploidía, una variación del número de copias) a partir de una representación gráfica de un recuento, una elevación y/o un perfil. Algunas veces se determina un resultado a partir de una representación gráfica de un recuento, una elevación y/o un perfil. En algunas implementaciones, una representación gráfica (por ejemplo, un gráfico) se realiza (por ejemplo, se genera) por un módulo de representación gráfica o un aparato que comprende un módulo de representación gráfica. En algunas implementaciones, se requiere un módulo de representación gráfica o un aparato que comprende un módulo de

representación gráfica para representar gráficamente un recuento, una elevación o un perfil. Un módulo de representación gráfica puede visualizar una representación gráfica o enviar una representación gráfica a una pantalla de visualización (por ejemplo, un módulo de visualización). Un aparato que comprende un módulo de representación gráfica puede comprender al menos un procesador. En algunas implementaciones, se proporciona una representación gráfica por un aparato que incluye un procesador (por ejemplo, uno o más procesadores), procesador que puede realizar y/o implementar una o más instrucciones (por ejemplo, procesos, rutinas y/o subrutinas) del módulo de representación gráfica. En algunas implementaciones, se realiza una representación gráfica por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunas implementaciones, un módulo de representación gráfica funciona con uno o más procesadores externos (por ejemplo, una red interna o externa, un servidor, dispositivo de almacenamiento y/o una red de almacenamiento (por ejemplo, una nube)). Algunas veces, un aparato que comprende un módulo de representación gráfica recopila, ensambla y/o recibe información y/o datos de otro módulo o aparato. Algunas veces, un módulo de representación gráfica recibe y recopila información y/o datos de un componente o periférico. A menudo, un módulo de representación gráfica recibe, recopila, ensambla y/o representa gráficamente lecturas de secuencia, secciones genómicas, lecturas mapeadas, recuentos, elevaciones, perfiles, elevaciones de referencia, elevaciones esperadas, rangos de elevación esperados, valores de incertidumbre, comparaciones, elevaciones categorizadas (por ejemplo, elevaciones identificadas como variaciones del número de copias) y/o resultados, ajustes y/o constantes. Algunas veces, un módulo de representación gráfica acepta y recopila información y/o datos de entrada de un operador de un aparato. Por ejemplo, algunas veces un operador de un aparato proporciona una constante, un valor umbral, una fórmula o un valor predeterminado a un módulo de representación gráfica. En algunas implementaciones, se proporcionan información y/o datos por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunas implementaciones, un recuento, una elevación y/o un perfil se representan gráficamente por un aparato que comprende un componente o periférico adecuado. Un aparato que comprende un módulo de representación gráfica puede recibir datos normalizados de un módulo de normalización, rangos de un módulo de establecimiento de rango, datos de comparación de un módulo de comparación, datos de categorización de un módulo de categorización y/o datos de ajuste de un módulo de ajuste. Un módulo de representación gráfica puede recibir información y/o datos, transformar la información y/o los datos y proporcionar los datos representados gráficamente. Algunas veces, un aparato que comprende un módulo de representación gráfica proporciona y/o transfiere información y/o datos a otro módulo o aparato. Un aparato que comprende un módulo de representación gráfica puede representar gráficamente un recuento, una elevación y/o un perfil y proporcionar o transferir información y/o datos relacionados con la representación gráfica a un aparato y/o módulo adecuado. A menudo, un módulo de representación gráfica recibe, recopila, ensambla y/o representa gráficamente elevaciones (por ejemplo, perfiles, primeras elevaciones) y transfiere información y/o datos representados gráficamente a y desde un módulo de ajuste y/o módulo de comparación. La información y/o los datos representados gráficamente a veces se transfieren de un módulo de representación gráfica a un módulo de categorización y/o un periférico (por ejemplo, una pantalla de visualización o impresora). En algunas implementaciones, las representaciones gráficas se categorizan y/o se determina que comprenden una variación genética (por ejemplo, una aneuploidía) o una variación del número de copias (por ejemplo, una variación del número de copias materno y/o fetal). Un recuento, una elevación y/o un perfil representados gráficamente mediante los métodos descritos en el presente documento pueden verificarse y/o ajustarse independientemente mediante pruebas adicionales (por ejemplo, por secuenciación dirigida de ácido nucleico materno y/o fetal).

Algunas veces se determina un resultado según una o más elevaciones. En algunas implementaciones, una determinación de la presencia o ausencia de una variación genética (por ejemplo, una aneuploidía cromosómica) se determina según una o más elevaciones ajustadas. Algunas veces, una determinación de la presencia o ausencia de una variación genética (por ejemplo, una aneuploidía cromosómica) se determina según un perfil que comprende de 1 a aproximadamente 10.000 elevaciones ajustadas. A menudo, una determinación de la presencia o ausencia de una variación genética (por ejemplo, una aneuploidía cromosómica) se determina según un perfil que comprende de aproximadamente 1 a aproximadamente 1000, de 1 a aproximadamente 900, de 1 a aproximadamente 800, de 1 a aproximadamente 700, de 1 a aproximadamente 600, de 1 a aproximadamente 500, de 1 a aproximadamente 400, de 1 a aproximadamente 300, de 1 a aproximadamente 200, de 1 a aproximadamente 100, de 1 a aproximadamente 50, de 1 a aproximadamente 25, de 1 a aproximadamente 20, de 1 a aproximadamente 15, de 1 a aproximadamente 10 o de 1 a aproximadamente 5 ajustes. Algunas veces se determina la presencia o ausencia de una variación genética (por ejemplo, una aneuploidía cromosómica) según un perfil que comprende aproximadamente 1 ajuste (por ejemplo, una elevación ajustada). A veces se determina un resultado según uno o más perfiles (por ejemplo, un perfil de un cromosoma o segmento del mismo) que comprende uno o más, 2 o más, 3 o más, 5 o más, 6 o más, 7 o más, 8 o más, 9 o más o a veces 10 o más ajustes. A veces, una determinación de la presencia o ausencia de una variación genética (por ejemplo, una aneuploidía cromosómica) se determina según un perfil en el que no se ajustan algunas elevaciones en un perfil. A veces, una determinación de la presencia o ausencia de una variación genética (por ejemplo, una aneuploidía cromosómica) se determina según un perfil en el que no se realizan ajustes.

En algunas implementaciones, un ajuste de una elevación (por ejemplo, una primera elevación) en un perfil reduce una determinación falsa o un resultado falso. En algunas implementaciones, un ajuste de una elevación (por ejemplo, una primera elevación) en un perfil reduce la frecuencia y/o probabilidad (por ejemplo, probabilidad estadística, posibilidad) de una determinación falsa o un resultado falso. Una determinación o un resultado falso puede ser una determinación o un resultado que no sea exacto. Una determinación o un resultado falso puede ser una determinación o un resultado que no refleja la composición genética real o verdadera o la disposición genética real o verdadera (por ejemplo, la presencia o ausencia de una variación genética) de un sujeto (por ejemplo, una mujer embarazada, un feto y/o una combinación de los mismos). Algunas veces, una determinación o un resultado falso es una determinación de falso

negativo. En algunas implementaciones, una determinación negativa o un resultado negativo es la ausencia de una variación genética (por ejemplo, aneuploidía, variación del número de copias). Algunas veces, una determinación falsa o un resultado falso es una determinación de falso positivo o un resultado de falso positivo. En algunas implementaciones, una determinación positiva o un resultado positivo es la presencia de una variación genética (por ejemplo, aneuploidía, variación del número de copias). En algunas implementaciones, se utiliza una determinación o un resultado en un diagnóstico. En algunas implementaciones, una determinación o un resultado es para un feto.

Resultado

Los métodos descritos en el presente documento pueden proporcionar una determinación de la presencia o ausencia de una variación genética (por ejemplo, aneuploidía fetal) para una muestra, proporcionando de ese modo un resultado (por ejemplo, proporcionando de ese modo un resultado determinante de la presencia o ausencia de una variación genética (por ejemplo, aneuploidía fetal)). Una variación genética incluye a menudo una ganancia, una pérdida y/o alteración (por ejemplo, duplicación, deleción, fusión, inserción, mutación, reorganización, sustitución o metilación aberrante) de información genética (por ejemplo, cromosomas, segmentos de cromosomas, regiones polimórficas, regiones translocadas, secuencias de nucleótidos alteradas, similares o combinaciones de los anteriores) que dan como resultado un cambio detectable en el genoma o la información genética de un sujeto de prueba con respecto a una referencia. La presencia o ausencia de una variación genética puede determinarse mediante la transformación, el análisis y/o la manipulación de lecturas de secuencia que se han mapeado en secciones genómicas (por ejemplo, bins genómicos).

Los métodos descritos en el presente documento determinan, algunas veces, la presencia o ausencia de una aneuploidía fetal (por ejemplo, aneuploidía cromosómica completa, aneuploidía cromosómica parcial o aberración cromosómica segmentaria (por ejemplo, mosaicismo, deleción y/o inserción)) para una muestra de prueba de una mujer embarazada que porta un feto. A veces los métodos descritos en el presente documento detectan euploidía o falta de euploidía (no euploidía) para una muestra de una mujer embarazada que porta un feto. Los métodos descritos en el presente documento detectan, algunas veces, trisomía para uno o más cromosomas (por ejemplo, el cromosoma 13, el cromosoma 18, el cromosoma 21 o combinación de los mismos) o segmento de los mismos.

En algunas implementaciones, la presencia o ausencia de una variación genética (por ejemplo, una aneuploidía fetal) se determina mediante un método descrito en el presente documento, mediante un método conocido en la técnica o mediante una combinación de los mismos. La presencia o ausencia de una variación genética se determina generalmente a partir de recuentos de lecturas de secuencia mapeadas en secciones genómicas de un genoma de referencia. Los recuentos de lecturas de secuencia usados para determinar la presencia o ausencia de una variación genética son, algunas veces, recuentos sin procesar y/o recuentos filtrados y a menudo recuentos normalizados. Pueden usarse uno o varios procedimientos de normalización adecuados para generar recuentos normalizados, los ejemplos no limitativos de los cuales incluyen normalización basada en bins, normalización por contenido de GC, regresión lineal y no lineal por mínimos cuadrados, LOESS, LOESS de GC, LOWESS, PERUN, RM, GCRM y combinaciones de los mismos. Los recuentos normalizados, algunas veces, se expresan como uno o más niveles o elevaciones en un perfil para un conjunto o conjuntos particulares de secciones genómicas. Algunas veces, los recuentos normalizados se ajustan o rellenan antes de determinar la presencia o ausencia de una variación genética.

Algunas veces se determina la presencia o ausencia de una variación genética (por ejemplo, aneuploidía fetal) sin comparar recuentos de un conjunto de secciones genómicas con una referencia. Los recuentos medidos para una muestra de prueba y que se encuentran en una región de prueba (por ejemplo, un conjunto de secciones genómicas de interés) se denominan en el presente documento "recuentos de prueba". Los recuentos de prueba son, algunas veces, recuentos procesados, recuentos promediados o sumados, una representación, recuentos normalizados o uno o más niveles o elevaciones, tal como se describe en el presente documento.

Algunas veces, los recuentos de prueba se promedian o suman (por ejemplo, se calcula un promedio, media, mediana, moda o suma) para un conjunto de secciones genómicas, y los recuentos promediados o sumados se comparan con un umbral o rango. Algunas veces, los recuentos de prueba se expresan como una representación, que puede expresarse como una razón o un porcentaje de recuentos para un primer conjunto de secciones genómicas con respecto a recuentos para un segundo conjunto de secciones genómicas. Algunas veces, el primer conjunto de secciones genómicas es para uno o más cromosomas de prueba (por ejemplo, el cromosoma 13, el cromosoma 18, el cromosoma 21, o combinación de los mismos) y algunas veces el segundo conjunto de secciones genómicas es para el genoma o una parte del genoma (por ejemplo, autosomas o autosomas y cromosomas sexuales). Algunas veces, una representación se compara con un umbral o rango. Algunas veces, los recuentos de prueba se expresan como uno o más niveles o elevaciones para recuentos normalizados en un conjunto de secciones genómicas, y el o los niveles o elevaciones se comparan con un umbral o rango. Los recuentos de prueba (por ejemplo, recuentos promediados o sumados, representación, recuentos normalizados, uno o más niveles o elevaciones) por encima o por debajo de un umbral particular, en un rango particular o fuera de un rango particular, a veces determinan la presencia de una variación genética o falta de euploidía (por ejemplo, no euploidía). Los recuentos de prueba (por ejemplo, recuentos promediados o sumados, representación, recuentos normalizados, uno o más niveles o elevaciones) por debajo o por encima de un umbral particular, en un rango particular o fuera de un rango particular, a veces son determinantes de la ausencia de una variación genética o euploidía.

La presencia o ausencia de una variación genética (por ejemplo, aneuploidía fetal) se determina, algunas veces, comparando los recuentos de prueba (por ejemplo, recuentos sin procesar, recuentos filtrados, recuentos promediados o sumados, representación, recuentos normalizados, uno o más niveles o elevaciones, para un conjunto de secciones genómicas) con una referencia. Una referencia puede ser una determinación adecuada de recuentos. Los recuentos para una referencia a veces son recuentos sin procesar, recuentos filtrados, recuentos promediados o sumados, representación, recuentos normalizados, uno o más niveles o elevaciones, para un conjunto de secciones genómicas. Los recuentos de referencia son a menudo recuentos para una región de prueba euploide.

En determinadas implementaciones, los recuentos de prueba a veces son para un primer conjunto de secciones genómicas y una referencia incluye los recuentos para un segundo conjunto de secciones genómicas diferentes del primer conjunto de secciones genómicas. Algunas veces, los recuentos de referencia son para una muestra de ácido nucleico de la misma mujer embarazada de la cual se obtiene la muestra de prueba. Algunas veces, los recuentos de referencia son para una muestra de ácido nucleico de una o más mujeres embarazadas diferentes a las mujeres de las que se obtuvo la muestra de prueba. En algunas implementaciones, un primer conjunto de secciones genómicas está en el cromosoma 13, el cromosoma 18, el cromosoma 21, el segmento de los mismos o una combinación de los anteriores, y el segundo conjunto de secciones genómicas está en otro cromosoma o cromosomas o segmento de los mismos. En un ejemplo no limitativo, en el que un primer conjunto de secciones genómicas está en el cromosoma 21 o segmento del mismo, un segundo conjunto de secciones genómicas está a menudo en otro cromosoma (por ejemplo, el cromosoma 1, el cromosoma 13, el cromosoma 14, el cromosoma 18, el cromosoma 19, segmento de los mismos o combinación de los anteriores). Una referencia está ubicada a menudo en un cromosoma o segmento del mismo que es normalmente euploide. Por ejemplo, el cromosoma 1 y el cromosoma 19 son a menudo euploides en fetos debido a una alta tasa de mortalidad fetal precoz asociada con aneuploidías del cromosoma 1 y del cromosoma 19. Puede generarse una medida de desviación entre los recuentos de prueba y los recuentos de referencia.

A veces, una referencia comprende recuentos para el mismo conjunto de secciones genómicas que para los recuentos de prueba, en la que los recuentos para la referencia son de una o más muestras de referencia (por ejemplo, a menudo múltiples muestras de referencia de múltiples sujetos de referencia). Una muestra de referencia es a menudo de una o más mujeres embarazadas diferentes a la mujer de la cual se obtiene una muestra de prueba. Puede generarse una medida de desviación entre los recuentos de prueba y los recuentos de referencia.

Puede seleccionarse una medida de desviación adecuada entre los recuentos de prueba y los recuentos de referencia; los ejemplos no limitativos de los mismos incluyen desviación estándar, desviación absoluta promedio, mediana de desviación absoluta, desviación absoluta máxima, puntuación estándar (por ejemplo, valor de z, puntuación z, puntuación normal, variable normalizada) y similares. En algunas implementaciones, las muestras de referencia son euploides para una región de prueba y se evalúa la desviación entre los recuentos de prueba y los recuentos de referencia. Una desviación menor de tres entre los recuentos de prueba y los recuentos de referencia (por ejemplo, 3-sigma para la desviación estándar) es a menudo indicativa de una región de prueba euploide (por ejemplo, ausencia de una variación genética). A menudo, una desviación mayor de tres entre los recuentos de prueba y los recuentos de referencia es indicativa de una región de prueba no euploide (por ejemplo, presencia de una variación genética). Los recuentos de prueba significativamente por debajo de los recuentos de referencia, recuentos de referencia que son indicativos de euploidía, algunas veces son determinantes de una monosomía. Los recuentos de prueba significativamente por encima de los recuentos de referencia, recuentos de referencia que son indicativos de euploidía, algunas veces son determinantes de una trisomía. Una medida de desviación entre los recuentos de prueba para una muestra de prueba y los recuentos de referencia para múltiples sujetos de referencia pueden representarse gráficamente y visualizarse (por ejemplo, gráfico de puntuación z).

Cualquier otra referencia adecuada puede factorizarse con recuentos de prueba para determinar la presencia o ausencia de una variación genética (o determinación de euploides o no euploides) para una región de prueba de una muestra de prueba. Por ejemplo, una determinación de fracción fetal puede factorizarse con recuentos de pruebas para determinar la presencia o ausencia de una variación genética. Puede usarse un procedimiento adecuado para cuantificar la fracción fetal, los ejemplos no limitativos de los mismos incluyen un procedimiento de espectrometría de masas, procedimiento de secuenciación o combinación de los mismos.

El personal de laboratorio (por ejemplo, un gerente de laboratorio) puede analizar valores (por ejemplo, recuentos de prueba, recuentos de referencia, nivel de desviación) subyacentes a una determinación de la presencia o ausencia de una variación genética (o determinación de euploides o no euploides para una región de prueba). Para las identificaciones referidas a la presencia o ausencia de una variación genética cercana o cuestionable, el personal de laboratorio puede volver a ordenar la misma prueba y/u ordenar una prueba diferente (por ejemplo, cariotipado y/o amniocentesis en el caso de determinaciones de aneuploidía fetal), que usan el mismo ácido nucleico de muestra o diferente de un sujeto de prueba.

Algunas veces, una variación genética se asocia con una afección médica. Una determinación del resultado de una variación genética es, algunas veces, una determinación del resultado de la presencia o ausencia de una afección (por ejemplo, una afección médica), enfermedad, un síndrome o una anomalía, o incluye la detección de una afección, enfermedad, un síndrome o una anomalía (por ejemplo, los ejemplos no limitativos enumerados en la tabla 1). En algunos casos, un diagnóstico comprende la evaluación de un resultado. Un resultado determinante de la presencia o ausencia de una afección (por ejemplo, una afección médica), enfermedad, un síndrome o una

anomalía mediante los métodos descritos en el presente documento a veces pueden verificarse independientemente mediante pruebas adicionales (por ejemplo, mediante cariotipado y/o amniocentesis).

El análisis y procesamiento de datos puede proporcionar uno o más resultados. El término "resultado", tal como se usa en el presente documento, puede referirse a un resultado del procesamiento de datos que facilita la determinación de la presencia o ausencia de una variación genética (por ejemplo, una aneuploidía, una variación del número de copias). Algunas veces, el término "resultado", tal como se usa en el presente documento, se refiere a una conclusión que predice y/o determina la presencia o ausencia de una variación genética (por ejemplo, una aneuploidía, una variación del número de copias). Algunas veces el término "resultado", tal como se usa en el presente documento, se refiere a una conclusión que predice y/o determina un riesgo o probabilidad de la presencia o ausencia de una variación genética (por ejemplo, una aneuploidía, una variación del número de copias) en un sujeto (por ejemplo, un feto). Algunas veces, un diagnóstico comprende el uso de un resultado. Por ejemplo, un profesional sanitario puede analizar un resultado y proporcionar bases de diagnóstico en, o basándose en parte en, el resultado. En algunas implementaciones, la determinación, detección o el diagnóstico de una afección, un síndrome o una anomalía (por ejemplo, enumerado en la tabla 1) comprende el uso de un determinante del resultado de la presencia o ausencia de una variación genética. En algunas implementaciones, un resultado basado en lecturas de secuencia mapeadas contadas o transformaciones de las mismas es determinante de la presencia o ausencia de una variación genética. En algunas implementaciones, un resultado generado utilizando uno o más métodos (por ejemplo, métodos de procesamiento de datos) descritos en el presente documento es determinante de la presencia o ausencia de una o más afecciones, síndromes o anomalías enumerados en la tabla 1. A veces un diagnóstico comprende una determinación de una presencia o ausencia de una afección, un síndrome o una anomalía. A menudo, un diagnóstico comprende una determinación de una variación genética como la naturaleza y/o la causa de una afección, un síndrome o una anomalía. Algunas veces, un resultado no es un diagnóstico. Un resultado comprende a menudo uno o más valores numéricos generados usando un método de procesamiento descrito en el presente documento en el contexto de una o más consideraciones de probabilidad. Una consideración de riesgo o probabilidad puede incluir, pero no se limita a: un valor de incertidumbre, una medida de variabilidad, nivel de confianza, sensibilidad, especificidad, desviación estándar, coeficiente de variación (CV) y/o nivel de confianza, puntuaciones Z, valores de chi, valores de phi, valores de ploidía, fracción fetal ajustada, razones de área, mediana de elevación, similares o combinaciones de los mismos. Una consideración de probabilidad puede facilitar la determinación de si un sujeto está en riesgo de tener, o tiene, una variación genética, y un determinante del resultado de una presencia o ausencia de un trastorno genético incluye a menudo tal consideración.

Un resultado algunas veces es un fenotipo. Un resultado a veces es un fenotipo con un nivel asociado de confianza (por ejemplo, un valor de incertidumbre, por ejemplo, un feto es positivo para la trisomía 21 con un nivel de confianza del 99 %, un sujeto de prueba es negativo para un cáncer asociado con una variación genética a un nivel de confianza del 95 %). Diferentes métodos para generar valores de resultado a veces pueden producir diferentes tipos de resultados. Generalmente, existen cuatro tipos de puntuaciones o identificaciones posibles que pueden realizarse basadas en los valores de resultados generados con el uso de los métodos descritos en el presente documento: verdadero positivo, falso positivo, verdadero negativo y falso negativo. Los términos "puntuación", "puntuaciones", "identificación" e "identificaciones", tal como se usan en el presente documento, se refieren al cálculo de la probabilidad de que una variación genética particular esté presente o ausente en un sujeto/muestra. El valor de una puntuación puede usarse para determinar, por ejemplo, una variación, diferencia o razón de lecturas de secuencia mapeadas que pueden corresponder a una variación genética. Por ejemplo, calcular una puntuación positiva para una variación genética o sección genómica seleccionada a partir de un conjunto de datos, con respecto a un genoma de referencia puede conducir a una identificación de la presencia o ausencia de una variación genética, variación genética que a veces se asocia con una afección médica (por ejemplo, cáncer, preeclampsia, trisomía, monosomía, y similares). En algunas implementaciones, un resultado comprende una elevación, un perfil y/o una representación gráfica (por ejemplo, un gráfico de perfiles). En aquellas implementaciones en las que un resultado comprende un perfil, puede usarse un perfil adecuado o combinación de perfiles para un resultado. Los ejemplos no limitativos de perfiles que pueden usarse para un resultado incluyen perfiles de puntuación z, perfiles de valor de p, perfiles de valor de chi, perfiles de valor de phi, similares, y combinaciones de los mismos.

Un resultado generado para determinar la presencia o ausencia de una variación genética incluye, algunas veces, un resultado nulo (por ejemplo, un punto de datos entre dos agrupaciones, un valor numérico con una desviación estándar que abarca valores tanto para la presencia como para la ausencia de una variación genética, un conjunto de datos con un gráfico de perfiles que no es similar a los gráficos de perfiles para sujetos que tienen o están libres de la variación genética que se investiga). En algunas implementaciones, un resultado indicativo de un resultado nulo todavía es un resultado determinante, y la determinación puede incluir la necesidad de información adicional y/o una repetición de la generación y/o análisis de datos para determinar la presencia o ausencia de una variación genética.

Puede generarse un resultado después de realizar una o más etapas de procesamiento descritas en el presente documento, en algunas implementaciones. En determinadas implementaciones, se genera un resultado que resulta de una de las etapas de procesamiento descritas en el presente documento, y en algunas implementaciones, puede generarse un resultado después de realizar cada manipulación estadística y/o matemática de un conjunto de datos. Un resultado que corresponde a la determinación de la presencia o ausencia de una variación genética puede expresarse en una forma adecuada, forma que comprende, sin limitarse a, una probabilidad (por ejemplo, razón de probabilidades, valor de p), probabilidad, valor dentro o fuera de una agrupación, valor de por encima o por debajo de un valor umbral, valor dentro de un rango (por ejemplo, un rango umbral), valor con una medida de varianza o confianza, o factor de riesgo, asociado con la

presencia o ausencia de una variación genética para un sujeto o muestra. En determinadas implementaciones, la comparación entre las muestras permite la confirmación de la identidad de la muestra (por ejemplo, permite la identificación de muestras repetidas y/o muestras que se han mezclado (por ejemplo, mal etiquetadas, combinadas, y similares)).

En algunas implementaciones, un resultado comprende un valor por encima o por debajo de un umbral predeterminado o valor de punto de corte (por ejemplo, mayor de 1, menor de 1), y un nivel de incertidumbre o confianza asociado con el valor. Algunas veces, un valor umbral o de punto de corte predeterminado es una elevación esperada o un rango de elevación esperado. Un resultado puede describir además una suposición usada en el procesamiento de datos. En determinadas implementaciones, un resultado comprende un valor que se encuentra dentro o fuera de un rango predeterminado de valores (por ejemplo, un rango umbral) y el nivel de incertidumbre o de confianza asociado para ese valor que está dentro o fuera del rango. En algunas implementaciones, un resultado comprende un valor que es igual a un valor predeterminado (por ejemplo, igual a 1, igual a cero), o es igual a un valor dentro de un rango de valores de predeterminado, y su nivel de incertidumbre o de confianza asociado para ese valor es igual o está dentro o fuera de un rango. Algunas veces, un resultado se representa gráficamente como una representación gráfica (por ejemplo, gráfico de perfiles).

Tal como se mencionó anteriormente, un resultado puede caracterizarse como un verdadero positivo, verdadero negativo, falso positivo o falso negativo. La expresión "verdadero positivo", tal como se usa en el presente documento, se refiere a un sujeto diagnosticado correctamente de una variación genética. La expresión "falso positivo", tal como se usa en el presente documento, se refiere a un sujeto mal identificado como que tiene una variación genética. La expresión "verdadero negativo", tal como se usa en el presente documento, se refiere a un sujeto identificado correctamente como que no tiene una variación genética. La expresión "falso negativo", tal como se usa en el presente documento, se refiere a un sujeto mal identificado como que no tiene una variación genética. Dos medidas de rendimiento para cualquier método dado pueden calcularse basándose en las razones de estas apariciones: (i) un valor de sensibilidad, que es generalmente la fracción de positivos predichos que se identifican correctamente como positivos; y (ii) un valor de especificidad, que es generalmente la fracción de negativos predichos identificados correctamente como negativos. El término "sensibilidad", tal como se usa en el presente documento, se refiere al número de verdaderos positivos dividido entre el número de verdaderos positivos más el número de falsos negativos, en el que la sensibilidad (sens) puede estar dentro del rango de $0 < sens < 1$. Idealmente, el número de falsos negativos es igual a cero o próximo a cero, de modo que ningún sujeto se identifica incorrectamente como que no tiene al menos una variación genética cuando realmente tiene al menos una variación genética. Por el contrario, a menudo se realiza una evaluación de la capacidad de un algoritmo de predicción para clasificar correctamente los negativos, una medición complementaria a la sensibilidad. El término "especificidad", tal como se usa en el presente documento, se refiere al número de verdaderos negativos dividido entre el número de verdaderos negativos más el número de falsos positivos, en el que la sensibilidad (espec) puede estar dentro del rango de $0 < espec < 1$. Idealmente, el número de falsos positivos es igual a cero o próximo a cero, de modo que ningún sujeto se identifica incorrectamente como que tiene al menos una variación genética cuando no tiene la variación genética que se evalúa.

En determinadas implementaciones, uno o más de sensibilidad, especificidad y/o nivel de confianza se expresan como un porcentaje. En algunas implementaciones, el porcentaje, independientemente para cada variable, es mayor de aproximadamente el 90 % (por ejemplo, aproximadamente el 90, 91, 92, 93, 94, 95, 96, 97, 98 o el 99 %, o mayor del 99 % (por ejemplo, aproximadamente el 99,5 %, o mayor, aproximadamente el 99,9 % o mayor, aproximadamente el 99,95 % o mayor, aproximadamente el 99,99 % o mayor)). El coeficiente de variación (CV) en algunas implementaciones se expresa como un porcentaje, y algunas veces el porcentaje es de aproximadamente el 10 % o menos (por ejemplo, aproximadamente el 10, 9, 8, 7, 6, 5, 4, 3, 2 o el 1 %, o menos del 1 % (por ejemplo, aproximadamente el 0,5 % o menos, aproximadamente el 0,1 % o menos, aproximadamente el 0,05 % o menos, aproximadamente el 0,01 % o menos)). Una probabilidad (por ejemplo, que un resultado particular no se deba al azar) en determinadas implementaciones se expresa como una puntuación Z, un valor de p o los resultados de una prueba de la t. En algunas implementaciones, una varianza medida, intervalo de confianza, sensibilidad, especificidad y similares (por ejemplo, denominados colectivamente parámetros de confianza) para un resultado pueden generarse usando una o más manipulaciones de procesamiento de datos descritas en el presente documento. Los ejemplos específicos para generar resultados y niveles de confianza asociados se describen en la sección Ejemplos.

Algunas veces se selecciona un método que tiene sensibilidad y especificidad iguales a uno, o el 100 %, o cerca de uno (por ejemplo, de aproximadamente el 90 % a aproximadamente el 99 %). En algunas implementaciones, se selecciona un método que tiene una sensibilidad igual a 1 o el 100 % y, en determinadas implementaciones, se selecciona un método que tiene una sensibilidad próxima a 1 (por ejemplo, una sensibilidad de aproximadamente el 90 %, una sensibilidad de aproximadamente el 91 %, una sensibilidad de aproximadamente el 92 %, una sensibilidad de aproximadamente el 93 %, una sensibilidad de aproximadamente el 94 %, una sensibilidad de aproximadamente el 95 %, una sensibilidad de aproximadamente el 96 %, una sensibilidad de aproximadamente el 97 %, una sensibilidad de aproximadamente el 98 % o una sensibilidad de aproximadamente el 99 %). En algunas implementaciones se selecciona un método que tiene una especificidad equivalente a 1 o el 100 % y, en determinadas implementaciones, se selecciona un método que tiene una especificidad próxima a 1 (por ejemplo, una especificidad de aproximadamente el 90 %, una especificidad de aproximadamente el 91 %, una especificidad de aproximadamente el 92 %, una especificidad de aproximadamente el 93 %, una especificidad de aproximadamente el 94 %, una especificidad de aproximadamente el 95 %, una especificidad de aproximadamente el 96 %, una especificidad de aproximadamente el 97 %, una especificidad de aproximadamente el 98 % o una especificidad de aproximadamente el 99 %).

Módulo de resultados

La presencia o ausencia de una variación genética (una aneuploidía, una aneuploidía fetal, una variación del número de copias) puede identificarse por un módulo de resultados o por un aparato que comprende un módulo de resultados. Algunas veces se identifica una variación genética por un módulo de resultados. A menudo, un módulo de resultados identifica una determinación de la presencia o ausencia de una aneuploidía. En algunas implementaciones, un determinante del resultado de una variación genética (una aneuploidía, una variación del número de copias) puede identificarse por un módulo de resultados o por un aparato que comprende un módulo de resultados. Un módulo de resultados puede especializarse para determinar una variación genética específica (por ejemplo, una trisomía, una trisomía 21, una trisomía 18). Por ejemplo, un módulo de resultados que identifica una trisomía 21 puede ser diferente y/o diferente de un módulo de resultados que identifica una trisomía 18. En algunas implementaciones, se requiere un módulo de resultados o un aparato que comprende un módulo de resultados para identificar una variación genética o un determinante del resultado de una variación genética (por ejemplo, una aneuploidía, una variación del número de copias). Un aparato que comprende un módulo de resultados puede comprender al menos un procesador. En algunas implementaciones, se proporciona una variación genética o un determinante del resultado de una variación genética por un aparato que incluye un procesador (por ejemplo, uno o más procesadores), procesador que puede realizar y/o implementar una o más instrucciones (por ejemplo, procesos, rutinas y/o subrutinas) del módulo de resultados. En algunas implementaciones, una variación genética o un determinante del resultado de una variación genética se identifica por un aparato que puede incluir múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunas implementaciones, un módulo de resultados funciona con uno o más procesadores externos (por ejemplo, una red interna o externa, un servidor, dispositivo de almacenamiento y/o una red de almacenamiento (por ejemplo, una nube)). Algunas veces, un aparato que comprende un módulo de resultados recopila, ensambla y/o recibe información y/o datos de otro módulo o aparato. Algunas veces, un aparato que comprende un módulo de resultados proporciona y/o transfiere información y/o datos a otro módulo o aparato. Algunas veces, un módulo de resultados transfiere, recibe o recopila información y/o datos a o desde un componente o periférico. A menudo, un módulo de resultados recibe, recopila y/o ensambla recuentos, elevaciones, perfiles, información y/o datos normalizados, elevaciones de referencia, elevaciones esperadas, rangos esperados, valores de incertidumbre, ajustes, elevaciones ajustadas, representaciones gráficas, elevaciones categorizadas, comparaciones y/o constantes. Algunas veces, un módulo de resultados acepta y recopila información y/o datos de entrada de un operador de un aparato. Por ejemplo, algunas veces un operador de un aparato proporciona una constante, un valor umbral, una fórmula o un valor predeterminado a un módulo de resultados. En algunas implementaciones, se proporcionan información y/o datos por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunas implementaciones, la identificación de una variación genética o un determinante del resultado de una variación genética se proporciona por un aparato que comprende un componente o periférico adecuado. Un aparato que comprende un módulo de resultados puede recibir datos normalizados de un módulo de normalización, elevaciones y/o rangos esperados de un módulo de establecimiento de rango, datos de comparación de un módulo de comparación, elevaciones categorizadas de un módulo de categorización, representaciones gráficas de un módulo de representación gráfica y/o datos de ajuste de un módulo de ajuste. Un módulo de resultados puede recibir información y/o datos, transformar la información y/o los datos y proporcionar un resultado. Un módulo de resultados puede proporcionar o transferir información y/o datos relacionados con una variación genética o un determinante del resultado de una variación genética a un aparato y/o módulo adecuado. Una variación genética o un determinante del resultado de una variación genética identificada mediante los métodos descritos en el presente documento pueden verificarse independientemente mediante pruebas adicionales (por ejemplo, por secuenciación dirigida de ácido nucleico materno y/o fetal).

Después de generar uno o más resultados, se usa a menudo un resultado para proporcionar una determinación de la presencia o ausencia de una variación genética y/o afección médica asociada. Normalmente, se proporciona un resultado a un profesional de atención sanitaria (por ejemplo, técnico o gerente de laboratorio; médico o asistente). A menudo, un módulo de resultados proporciona un resultado. Algunas veces, un módulo de representación gráfica proporciona un resultado. Algunas veces se proporciona un resultado en un componente o periférico de un aparato. Por ejemplo, algunas veces una impresora o pantalla de visualización proporciona un resultado. En algunas implementaciones, un determinante del resultado de la presencia o ausencia de una variación genética se proporciona a un profesional sanitario en forma de un informe, y en determinadas implementaciones el informe comprende una visualización de un valor de resultado y un parámetro de confianza asociado. Generalmente, un resultado puede mostrarse en un formato adecuado que facilita la determinación de la presencia o ausencia de una variación genética y/o afección médica. Los ejemplos no limitativos de formatos adecuados para usar para informar y/o visualizar conjuntos de datos o para informar sobre un resultado incluyen datos digitales, un gráfico, un gráfico 2D, un gráfico 3D y un gráfico 4D, una imagen, un pictograma, una tabla, un gráfico de barras, un gráfico circular, un diagrama de flujo, un diagrama de dispersión, un mapa, un histograma, un gráfico de densidad, un gráfico de funciones, un diagrama de circuitos, un diagrama de bloques, un mapa de burbujas, un diagrama de constelaciones, un diagrama de contorno, un cartograma, un diagrama de araña, un diagrama de Venn, un nomograma, y similares, y una combinación de los anteriores. Varios ejemplos de representaciones de resultados se muestran en los dibujos y se describen en los ejemplos.

Generar un resultado puede considerarse una transformación de datos leídos de secuencia de ácido nucleico, o similares, en una representación de un ácido nucleico celular de un sujeto, en determinadas implementaciones. Por ejemplo, el análisis de lecturas de secuencia de ácido nucleico de un sujeto y la generación de un resultado y/o perfil

cromosómico puede considerarse una transformación de fragmentos de lectura de secuencia relativamente pequeños en una representación de una estructura cromosómica relativamente grande. En algunas implementaciones, un resultado procede de una transformación de lecturas de secuencia de un sujeto (por ejemplo, una mujer embarazada), en una representación de una estructura existente (por ejemplo, un genoma, un cromosoma o segmento del mismo) presente en el sujeto (por ejemplo, un ácido nucleico materno y/o fetal). En algunas implementaciones, un resultado comprende una transformación de lecturas de secuencia de un primer sujeto (por ejemplo, una mujer embarazada), en una representación compuesta de estructuras (por ejemplo, un genoma, un cromosoma o segmento del mismo), y una segunda transformación de la representación compuesta que produce una representación de una estructura presente en un primer sujeto (por ejemplo, una mujer embarazada) y/o un segundo sujeto (por ejemplo, un feto).

Uso de resultados

Un profesional sanitario, u otro individuo cualificado, que recibe un informe que comprende uno o más resultados determinantes de la presencia o ausencia de una variación genética, puede usar los datos visualizados en el informe para realizar una identificación en relación con el estado del paciente o sujeto de prueba. El profesional sanitario puede realizar una recomendación basada en el resultado proporcionado, en algunas implementaciones. Un profesional sanitario o individuo cualificado puede proporcionar a un paciente o sujeto de prueba una identificación o puntuación con respecto a la presencia o ausencia de la variación genética basándose en el valor o valores de resultado y parámetros de confianza asociados proporcionados en un informe, en algunas implementaciones. En determinadas implementaciones, un profesional sanitario o un individuo cualificado realiza una puntuación o identificación manualmente, usando la observación visual del informe proporcionado. En determinadas implementaciones, una puntuación o identificación se realiza mediante una rutina automatizada, algunas veces integrada en software, y revisada por un profesional sanitario o individuo cualificado para obtener precisión antes de proporcionar información a un paciente o sujeto de prueba. La expresión "recibir un informe", tal como se usa en el presente documento, se refiere a obtener, mediante un medio de comunicación, una representación escrita y/o gráfica que comprende un resultado, que después de la revisión permite a un profesional sanitario u otro individuo cualificado determinar la presencia o ausencia de una variación genética en un paciente o sujeto de prueba. El informe puede generarse mediante un ordenador o mediante la introducción de datos por seres humanos, y puede comunicarse usando medios electrónicos (por ejemplo, a través de Internet, a través de ordenador, a través de fax, desde una ubicación de red a otra ubicación en el mismo sitio físico o en sitios físicos diferentes), o mediante otro método para enviar o recibir datos (por ejemplo, servicio de correo, servicio de mensajería y similares). En algunas implementaciones, el resultado se transmite a un profesional de atención sanitaria en un medio adecuado incluyendo, sin limitación, de modo verbal, en forma de documento o archivo. El archivo puede ser, por ejemplo, pero sin limitarse a, un archivo acústico, un archivo legible por ordenador, un archivo en papel, un archivo de laboratorio o un archivo de historial clínico.

La expresión "proporcionar un resultado" y equivalentes gramaticales de la misma, tal como se usa en el presente documento puede referirse además a un método para obtener tal información incluyendo, sin limitación, obtener la información de un laboratorio (por ejemplo, un archivo de laboratorio). Un archivo de laboratorio puede generarse por un laboratorio que lleva a cabo uno o más ensayos o una o más etapas de procesamiento de datos para determinar la presencia o ausencia de la afección médica. El laboratorio puede estar en la misma ubicación o en una ubicación diferente (por ejemplo, en otro país) que el personal que identifica la presencia o ausencia de la afección medica del archivo de laboratorio. Por ejemplo, el archivo de laboratorio puede generarse en una ubicación y transmitirse a otra ubicación en la cual la información en la misma se transmitirá al sujeto femenino gestante. El archivo de laboratorio puede estar en forma tangible o en forma electrónica (por ejemplo, forma legible por ordenador), en determinadas implementaciones.

En algunas implementaciones, puede proporcionarse un resultado a un profesional de atención sanitaria, médico o individuo cualificado de un laboratorio y el profesional de atención sanitaria, médico o individuo cualificado puede realizar un diagnóstico basándose en el resultado. En algunas implementaciones, puede proporcionarse un resultado a un profesional de atención sanitaria, médico o individuo cualificado de un laboratorio y el profesional de atención sanitaria, médico o individuo cualificado puede realizar un diagnóstico basado, en parte, en el resultado junto con información y/o datos adicionales y otros resultados

Un profesional sanitario o individuo cualificado puede proporcionar una recomendación adecuada basándose en los resultados o resultados proporcionados en el informe. Los ejemplos no limitativos de recomendaciones que pueden proporcionarse basándose en el informe de resultados proporcionado incluyen cirugía, radioterapia, quimioterapia, asesoramiento genético, soluciones de tratamiento después del nacimiento (por ejemplo, planificación vital, cuidado asistido a largo plazo, medicamentos, tratamientos sintomáticos), interrupción del embarazo, trasplante de órganos, transfusión de sangre, similares o combinaciones de los anteriores. En algunas implementaciones, la recomendación depende de la clasificación basada en los resultados proporcionada (por ejemplo, síndrome de Down, síndrome de Turner, afecciones médicas asociadas con variaciones genéticas en T13, afecciones médicas asociadas con variaciones genéticas en T18).

El software puede usarse para realizar una o más etapas en los procedimientos descritos en el presente documento, incluyendo, pero sin limitación, contar, procesar datos, generar un resultado y/o proporcionar una o más recomendaciones basándose en los resultados generados, tal como se describe con mayor detalle más adelante.

Transformaciones

Tal como se mencionó anteriormente, los datos a veces se transforman de una forma a otra. Los términos "transformado/a", "transformación" y derivaciones gramaticales o equivalentes de los mismos, tal como se usan en el presente documento, se refieren a una alteración de los datos de un material de partida físico (por ejemplo, ácido nucleico de muestra de sujeto de referencia y/o sujeto de prueba) en una representación digital del material de partida físico (por ejemplo, datos de lectura de secuencia), y en algunas implementaciones incluye una transformación adicional en uno o más valores numéricos o representaciones gráficas de la representación digital que pueden usarse para proporcionar un resultado. En determinadas implementaciones, el uno o más valores numéricos y/o representaciones gráficas de datos representados digitalmente pueden usarse para representar el aspecto del genoma físico de un sujeto de prueba (por ejemplo, representar virtualmente o representar visualmente la presencia o ausencia de una inserción genómica, duplicación o deleción; representar la presencia o ausencia de una variación en la cantidad física de una secuencia asociada con afecciones médicas). A veces, una representación virtual se transforma además en uno o más valores numéricos o representaciones gráficas de la representación digital del material de partida. Estos procedimientos pueden transformar material de partida físico en un valor numérico o representación gráfica, o una representación del aspecto físico del genoma de un sujeto de prueba.

En algunas implementaciones, la transformación de un conjunto de datos facilita proporcionar un resultado al reducir la complejidad de los datos y/o la dimensionalidad de los datos. La complejidad del conjunto de datos a veces se reduce durante el procedimiento de transformación de un material de partida físico en una representación virtual del material de partida (por ejemplo, lecturas de secuencia representativas del material de partida físico). Una característica o variable adecuada puede usarse para reducir la complejidad y/o dimensionalidad del conjunto de datos. Los ejemplos no limitativos de características que pueden seleccionarse para su uso como característica objetivo para el procesamiento de datos incluyen contenido de GC, predicción del sexo del feto, identificación de aneuploidía cromosómica, identificación de genes o proteínas particulares, identificación de cáncer, enfermedades, genes/rasgos heredados, anomalías cromosómicas, una categoría biológica, una categoría química, una categoría bioquímica, una categoría de genes o proteínas, una ontología génica, una ontología de proteínas, genes corregulados, genes de señalización celular, genes del ciclo celular, proteínas que pertenecen a los genes anteriores, variantes génicas, variantes de proteínas, genes corregulados, proteínas correguladas, secuencia de aminoácidos, secuencia de nucleótidos, datos de estructura proteica y similares, y combinaciones de los anteriores. Los ejemplos no limitativos de la reducción de la complejidad y/o la dimensionalidad del conjunto de datos incluyen la reducción de una pluralidad de lecturas de secuencia a gráficos de perfiles, la reducción de una pluralidad de lecturas de secuencia a valores numéricos (por ejemplo, valores normalizados, puntuaciones Z, valores de p); reducción de métodos de análisis múltiple a gráficos de probabilidad o puntos únicos; análisis de componentes principales de cantidades derivadas; y similares o combinaciones de estos.

Sistemas, aparatos y productos de programa informático de normalización de sección genómica

En determinados aspectos, se proporciona un sistema que comprende uno o más procesadores y memoria, memoria que comprende instrucciones ejecutables por el uno o más procesadores y memoria que comprende recuentos de lecturas de secuencia de ácido nucleico de muestra circulante, libre de células de un sujeto de prueba mapeado en secciones genómicas de un genoma de referencia; e instrucciones ejecutables por uno o más procesadores que están configuradas para: (a) generar un perfil de recuento normalizado de muestra normalizando los recuentos de las lecturas de secuencia para cada una de las secciones genómicas; y (b) determinar la presencia o ausencia de una aberración cromosómica segmentaria o una aneuploidía fetal o ambos a partir del perfil de recuento normalizado de muestra en (a).

También en determinados aspectos se proporciona un aparato que comprende uno o más procesadores y memoria, memoria que comprende instrucciones ejecutables por uno o más procesadores y memoria que comprende recuentos de lecturas de secuencia de ácido nucleico de muestra circulante, libre de células de un sujeto de prueba mapeado en secciones genómicas de un genoma de referencia; e instrucciones ejecutables por uno o más procesadores que están configuradas para: (a) generar un perfil de recuento normalizado de muestra normalizando los recuentos de las lecturas de secuencia para cada una de las secciones genómicas; y (b) determinar la presencia o ausencia de una aberración cromosómica segmentaria o una aneuploidía fetal o ambos a partir del perfil de recuento normalizado de muestra en (a).

También se proporciona en determinados aspectos un producto de programa informático incorporado de manera tangible en un medio legible por ordenador, que comprende instrucciones que cuando se ejecutan por uno o más procesadores están configuradas para: (a) acceder a recuentos de lecturas de secuencia de ácido nucleico de muestra circulante, libre de células de un sujeto de prueba mapeado en secciones genómicas de un genoma de referencia; (b) generar un perfil de recuento normalizado de muestra normalizando recuentos de las lecturas de secuencia para cada una de las secciones genómicas; y (c) determinar la presencia o ausencia de una aberración cromosómica segmentaria o una aneuploidía fetal o ambas a partir del perfil de recuento normalizado de muestra en (b).

En algunas implementaciones, los recuentos de las lecturas de secuencia para cada una de las secciones genómicas en un segmento del genoma de referencia (por ejemplo, el segmento es un cromosoma) se normalizan individualmente según los recuentos totales de lecturas de secuencia en las secciones genómicas en el segmento. Algunas secciones genómicas en el segmento algunas veces se eliminan (por ejemplo, se filtran) y las secciones genómicas restantes en el segmento se normalizan.

En determinadas realizaciones, el sistema, aparato y/o producto de programa informático comprende: (i) un módulo de secuenciación configurado para obtener lecturas de secuencia de ácido nucleico; (ii) un módulo de mapeo configurado para mapear lecturas de secuencia de ácido nucleico en porciones de un genoma de referencia; (iii) un módulo de ponderación configurado para ponderar secciones genómicas, (iv) un módulo de filtrado configurado para filtrar secciones genómicas o recuentos mapeados en una sección genómica, (v) un módulo de recuento configurado para proporcionar recuentos de lecturas de secuencia de ácido nucleico mapeadas en porciones de un genoma de referencia; (vi) un módulo de normalización configurado para proporcionar recuentos normalizados; (vii) un módulo de comparación configurado para proporcionar una identificación de una primera elevación que es significativamente diferente de una segunda elevación; (viii) un módulo de establecimiento de rango configurado para proporcionar uno o más rangos de nivel esperados; (ix) un módulo de categorización configurado para identificar una elevación representativa de una variación del número de copias; (x) un módulo de ajuste configurado para ajustar un nivel identificado como una variación del número de copias; (xi) un módulo de representación gráfica configurado para graficar y mostrar un nivel y/o un perfil; (xii) un módulo de resultados configurado para determinar un resultado (por ejemplo, resultado determinante de la presencia o ausencia de una aneuploidía fetal); (xiii) un módulo de organización de visualización de datos configurado para indicar la presencia o ausencia de una aberración cromosómica segmentaria o una aneuploidía fetal o ambas; (xiv) un módulo de procesamiento lógico configurado para realizar una o más lecturas de secuencia de mapa, contar lecturas de secuencia mapeadas, normalizar recuentos y generar un resultado; o (xv) combinación de dos o más de los anteriores.

En algunas implementaciones, el módulo de secuenciación y el módulo de mapeo están configurados para transferir lecturas de secuencia desde el módulo de secuenciación al módulo de mapeo. El módulo de mapeo y el módulo de recuento a veces están configurados para transferir lecturas de secuencia mapeadas desde el módulo de mapeo al módulo de recuento. A veces, el módulo de recuento y el módulo de filtrado están configurados para transferir recuentos desde el módulo de recuento al módulo de filtrado. A veces, el módulo de recuento y el módulo de ponderación están configurados para transferir recuentos desde el módulo de recuento al módulo de ponderación. El módulo de mapeo y el módulo de filtrado a veces están configurados para transferir lecturas de secuencia mapeadas desde el módulo de mapeo al módulo de filtrado. El módulo de mapeo y el módulo de ponderación a veces están configurados para transferir lecturas de secuencia mapeadas desde el módulo de mapeo al módulo de ponderación. A veces, el módulo de ponderación, el módulo de filtrado y el módulo de recuento están configurados para transferir secciones genómicas filtradas y/o ponderadas desde el módulo de ponderación y el módulo de filtrado al módulo de recuento. A veces, el módulo de ponderación y el módulo de normalización están configurados para transferir secciones genómicas ponderadas desde el módulo de ponderación al módulo de normalización. A veces, el módulo de filtrado y el módulo de normalización están configurados para transferir secciones genómicas filtradas desde el módulo de filtrado al módulo de normalización. En algunas implementaciones, el módulo de normalización y/o módulo de comparación están configurados para transferir recuentos normalizados al módulo de comparación y/o módulo de establecimiento de rango. El módulo de comparación, el módulo de establecimiento de rango y/o el módulo de categorización están configurados independientemente para transferir (i) una identificación de una primera elevación que es significativamente diferente de una segunda elevación y/o (ii) un rango de nivel esperado desde el módulo de comparación y/o el módulo de establecimiento de rango al módulo de categorización, en algunas implementaciones. En determinadas implementaciones, el módulo de categorización y el módulo de ajuste están configurados para transferir una elevación categorizada como una variación del número de copias desde el módulo de categorización hasta el módulo de ajuste. En algunas implementaciones, el módulo de ajuste, el módulo de representación gráfica y el módulo de resultados están configurados para transferir uno o más niveles ajustados desde el módulo de ajuste al módulo de representación gráfica o el módulo de resultados. El módulo de normalización a veces está configurado para transferir recuentos de lectura de secuencia normalizada mapeados en uno o más del módulo de comparación, el módulo de establecimiento de rango, el módulo de categorización, el módulo de ajuste, el módulo de resultados o el módulo de representación gráfica.

Sistemas parametrizados de eliminación de errores y de normalización no sesgada, aparatos y productos de programas informatico

Se proporcionan en determinados aspectos un sistema que comprende uno o más procesadores y memoria, memoria que comprende instrucciones ejecutables por el uno o más microprocesadores y memoria que comprende recuentos de lecturas de secuencia mapeadas en porciones de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante, libre de células de una muestra de prueba; e instrucciones ejecutables por el uno o más procesadores que están configuradas para: (a) determinar un sesgo de guanina y citosina (GC) para cada una de las porciones del genoma de referencia para múltiples muestras de una relación ajustada para cada muestra entre (i) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, y (ii) el contenido de GC para cada una de las porciones; y (b) calcular un nivel de sección genómica para cada una de las porciones del genoma de referencia a partir de una relación ajustada entre (i) el sesgo de GC y (ii) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, proporcionándose de ese modo niveles de sección genómica calculados, mediante lo cual el sesgo en los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia se reduce en los niveles de sección genómica calculados.

También se proporciona en determinados aspectos un sistema que comprende uno o más procesadores y memoria, memoria que comprende instrucciones ejecutables por el uno o más microprocesadores y memoria que comprende recuentos de lecturas de secuencia mapeadas en porciones de un genoma de referencia, lecturas de secuencia que son

lecturas de ácido nucleico circulante, libre de células de una muestra de prueba; e instrucciones ejecutables por el uno o más procesadores que están configuradas para: (a) determinar un sesgo de guanina y citosina (GC) para cada una de las porciones del genoma de referencia para múltiples muestras de una relación ajustada para cada muestra entre (i) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, y (ii) el contenido de GC para cada una de las porciones; y (b) calcular un nivel de sección genómica para cada una de las porciones del genoma de referencia a partir de una relación ajustada entre (i) el sesgo de GC y (ii) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, proporcionándose de ese modo niveles de sección genómica calculados, mediante lo cual el sesgo en los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia se reduce en los niveles de sección genómica calculados.

También se proporciona en determinados aspectos un producto de programa informático incorporado de manera tangible en un medio legible por ordenador, que comprende instrucciones que cuando se ejecutan por uno o más procesadores están configuradas para: (a) acceder a recuentos de lecturas de secuencia mapeadas en porciones de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante, libre de células de una muestra de prueba; (b) determinar un sesgo de guanina y citosina (GC) para cada una de las porciones del genoma de referencia para múltiples muestras a partir de una relación ajustada para cada muestra entre (i) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, y (ii) el contenido de GC para cada una de las porciones; y (c) calcular un nivel de sección genómica para cada una de las porciones del genoma de referencia a partir de una relación ajustada entre (i) el sesgo de GC y (ii) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, proporcionándose de ese modo niveles de sección genómica calculados, mediante lo cual el sesgo en los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia se reduce en los niveles de sección genómica calculados.

Se proporciona en determinados aspectos un sistema que comprende uno o más procesadores y memoria, memoria que comprende instrucciones ejecutables por el uno o más microprocesadores y memoria que comprende recuentos de lecturas de secuencia mapeadas en porciones de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante, libre de células de una mujer embarazada que porta un feto; e instrucciones ejecutables por el uno o más procesadores que están configuradas para: (a) determinar un sesgo de guanina y citosina (GC) para cada una de las porciones del genoma de referencia para múltiples muestras de una relación ajustada para cada muestra entre (i) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, y (ii) el contenido de GC para cada una de las porciones; (b) calcular un nivel de sección genómica para cada una de las porciones del genoma de referencia a partir de una relación ajustada entre el sesgo de GC y los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, proporcionando de ese modo niveles de sección genómica calculados; y (c) identificar la presencia o ausencia de una aneuploidía para el feto según los niveles de la sección genómica calculados con una sensibilidad del 95 % o mayor y una especificidad del 95 % o mayor.

También se proporciona en determinados aspectos un aparato que comprende uno o más procesadores y memoria, memoria que comprende instrucciones ejecutables por el uno o más microprocesadores y memoria que comprende recuentos de lecturas de secuencia mapeadas en porciones de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante, libre de células de una mujer embarazada que porta un feto; e instrucciones ejecutables por el uno o más procesadores que están configuradas para: (a) determinar un sesgo de guanina y citosina (GC) para cada una de las porciones del genoma de referencia para múltiples muestras de una relación ajustada para cada muestra entre (i) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, y (ii) el contenido de GC para cada una de las porciones; (b) calcular un nivel de sección genómica para cada una de las porciones del genoma de referencia a partir de una relación ajustada entre el sesgo de GC y los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, proporcionando de ese modo niveles de sección genómica calculados; y (c) identificar la presencia o ausencia de una aneuploidía para el feto según los niveles de la sección genómica calculados con una sensibilidad del 95 % o mayor y una especificidad del 95 % o mayor.

También en determinados aspectos se proporciona un producto de programa informático incorporado de manera tangible en un medio legible por ordenador, que comprende instrucciones que cuando se ejecutan por uno o más procesadores están configuradas para: (a) acceder a recuentos de lecturas de secuencia mapeadas en porciones de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante, libre de células de una mujer embarazada que porta un feto; (b) determinar un sesgo de guanina y citosina (GC) para cada una de las porciones del genoma de referencia a través de múltiples muestras a partir de una relación ajustada para cada muestra entre (i) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, y (ii) el contenido de GC para cada una de las porciones; (c) calcular un nivel de sección genómica para cada una de las porciones del genoma de referencia a partir de una relación ajustada entre el sesgo de GC y los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, proporcionándose de ese modo niveles de sección genómica calculados; y (d) identificar la presencia o ausencia de una aneuploidía para el feto según los niveles de sección genómica calculados con una sensibilidad del 95 % o más y una especificidad del 95 % o más.

También se proporciona en determinados aspectos un sistema que comprende uno o más procesadores y memoria, memoria que comprende instrucciones ejecutables por el uno o más microprocesadores y memoria que comprende recuentos de lecturas de secuencia mapeadas en porciones de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante, libre de células de una mujer embarazada que porta un feto; e instrucciones

ejecutables por el uno o más procesadores que están configuradas para: (a) determinar el sesgo experimental para cada una de las porciones del genoma de referencia para múltiples muestras a partir de una relación ajustada entre (i) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, y (ii) una característica de mapeo para cada una de las porciones; y (b) calcular un nivel de sección genómica para cada una de las porciones del genoma de referencia a partir de una relación ajustada entre el sesgo experimental y los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, proporcionándose de ese modo niveles de sección genómica calculados, mediante lo cual el sesgo en los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia se reduce en los niveles de sección genómica calculados.

Se proporciona en determinados aspectos un aparato que comprende uno o más procesadores y memoria, memoria que comprende instrucciones ejecutables por el uno o más microprocesadores y memoria que comprende recuentos de lecturas de secuencia mapeadas en porciones de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante, libre de células de una mujer embarazada que porta un feto; e instrucciones ejecutables por el uno o más procesadores que están configuradas para: (a) determinar el sesgo experimental para cada una de las porciones del genoma de referencia para múltiples muestras a partir de una relación ajustada entre (i) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, y (ii) una característica de mapeo para cada una de las porciones; y (b) calcular un nivel de sección genómica para cada una de las porciones del genoma de referencia a partir de una relación ajustada entre el sesgo experimental y los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, proporcionándose de ese modo niveles de sección genómica calculados, mediante lo cual el sesgo en los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia se reduce en los niveles de sección genómica calculados.

También se proporciona en determinados aspectos un producto de programa informático incorporado de manera tangible en un medio legible por ordenador, que comprende instrucciones que cuando se ejecutan por uno o más procesadores están configuradas para: (a) acceder a recuentos de lecturas de secuencia mapeadas en porciones de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante, libre de células de una muestra de prueba; (b) determinar el sesgo experimental para cada una de las porciones del genoma de referencia a través de múltiples muestras a partir de una relación ajustada entre (i) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, y (ii) una característica de mapeo para cada una de las porciones; y (c) calcular un nivel de sección genómica para cada una de las porciones del genoma de referencia a partir de una relación ajustada entre el sesgo experimental y los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, proporcionándose de ese modo niveles de sección genómica calculados, mediante lo cual el sesgo en los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia se reduce en los niveles de sección genómica calculados.

En determinadas realizaciones, el sistema, aparato y/o producto de programa informático comprende: (i) un módulo de secuenciación configurado para obtener lecturas de secuencia de ácido nucleico; (ii) un módulo de mapeo configurado para mapear lecturas de secuencia de ácido nucleico en porciones de un genoma de referencia; (iii) un módulo de ponderación configurado para ponderar secciones genómicas; (iv) un módulo de filtrado configurado para filtrar secciones genómicas o recuentos mapeados en una sección genómica; (v) un módulo de recuento configurado para proporcionar recuentos de lecturas de secuencia de ácido nucleico mapeadas en porciones de un genoma de referencia; (vi) un módulo de normalización configurado para proporcionar recuentos normalizados; (vii) un módulo de comparación configurado para proporcionar una identificación de una primera elevación que es significativamente diferente de una segunda elevación; (viii) un módulo de establecimiento de rango configurado para proporcionar uno o más rangos de nivel esperados; (ix) un módulo de categorización configurado para identificar una elevación representativa de una variación del número de copias; (x) un módulo de ajuste configurado para ajustar un nivel identificado como una variación del número de copias; (xi) un módulo de representación gráfica configurado para graficar y visualizar un nivel y/o un perfil; (xii) un módulo de resultados configurado para determinar un resultado (por ejemplo, determinante del resultado de la presencia o ausencia de una aneuploidía fetal); (xiii) un módulo de organización de visualización de datos configurado para indicar la presencia o ausencia de una aberración cromosómica segmentaria o una aneuploidía fetal o ambas; (xiv) un módulo de procesamiento lógico configurado para realizar una o más lecturas de secuencia de mapa, contar lecturas de secuencia mapeadas, normalizar recuentos y generar un resultado; o (xv) combinación de dos o más de los anteriores.

En algunas implementaciones, el módulo de secuenciación y el módulo de mapeo están configurados para transferir lecturas de secuencia desde el módulo de secuenciación al módulo de mapeo. El módulo de mapeo y el módulo de recuento a veces están configurados para transferir lecturas de secuencia mapeadas desde el módulo de mapeo al módulo de recuento. A veces, el módulo de recuento y el módulo de filtrado están configurados para transferir recuentos desde el módulo de recuento al módulo de filtrado. A veces, el módulo de recuento y el módulo de ponderación están configurados para transferir recuentos desde el módulo de recuento al módulo de ponderación. El módulo de mapeo y el módulo de filtrado a veces están configurados para transferir lecturas de secuencia mapeadas desde el módulo de mapeo al módulo de filtrado. El módulo de mapeo y el módulo de ponderación a veces están configurados para transferir lecturas de secuencia mapeadas desde el módulo de mapeo al módulo de ponderación. A veces, el módulo de ponderación, el módulo de filtrado y el módulo de recuento están configurados para transferir secciones genómicas filtradas y/o ponderadas desde el módulo de ponderación y el módulo de filtrado al módulo de recuento. A veces, el módulo de ponderación y el módulo de normalización están configurados para transferir secciones genómicas ponderadas desde el módulo de ponderación al

módulo de normalización. A veces, el módulo de filtrado y el módulo de normalización están configurados para transferir secciones genómicas filtradas desde el módulo de filtrado al módulo de normalización. En algunas implementaciones, el módulo de normalización y/o módulo de comparación están configurados para transferir recuentos normalizados al módulo de comparación y/o módulo de establecimiento de rango. El módulo de comparación, el módulo de establecimiento de rango y/o el módulo de categorización están configurados independientemente para transferir (i) una identificación de una primera elevación que es significativamente diferente de una segunda elevación y/o (ii) un rango de nivel esperado desde el módulo de comparación y/o el módulo de establecimiento de rango al módulo de categorización, en algunas implementaciones. En determinadas implementaciones, el módulo de categorización y el módulo de ajuste están configurados para transferir una elevación categorizada como una variación del número de copias desde el módulo de categorización hasta el módulo de ajuste. En algunas implementaciones, el módulo de ajuste, el módulo de representación gráfica y el módulo de resultados están configurados para transferir uno o más niveles ajustados desde el módulo de ajuste al módulo de representación gráfica o el módulo de resultados. El módulo de normalización a veces está configurado para transferir recuentos de lectura de secuencia normalizada mapeados en uno o más del módulo de comparación, el módulo de establecimiento de rango, el módulo de categorización, el módulo de ajuste, el módulo de resultados o el módulo de representación gráfica.

Sistemas, aparatos y productos de programa informático de ajuste

También se proporciona en determinados aspectos un sistema que comprende uno o más procesadores y memoria, memoria que comprende instrucciones ejecutables por el uno o más microprocesadores y memoria que comprende recuentos de lecturas de secuencia mapeadas en secciones genómicas de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante, libre de células de una mujer embarazada; e instrucciones ejecutables por el uno o más procesadores que están configuradas para: (a) normalizar los recuentos mapeados en las secciones genómicas del genoma de referencia, proporcionándose de ese modo un perfil de recuentos normalizados para las secciones genómicas; (b) identificar una primera elevación de los recuentos normalizados significativamente diferente de una segunda elevación de los recuentos normalizados en el perfil, primera elevación que es para un primer conjunto de secciones genómicas, y segunda elevación que es para un segundo conjunto de secciones genómicas; (c) determinar un rango de elevación esperado para una variación del número de copias homocigotas y heterocigotas según un valor de incertidumbre para un segmento del genoma; (d) ajustar la primera elevación en un valor predeterminado cuando la primera elevación está dentro de uno de los rangos de elevación esperados, proporcionándose de ese modo un ajuste de la primera elevación; y (e) determinar la presencia o ausencia de una aneuploidía cromosómica en el feto según las elevaciones de secciones genómicas que comprenden el ajuste de (d), mediante lo cual el resultado determinante de la presencia o ausencia de la aneuploidía cromosómica se genera a partir de las lecturas de secuencia de ácido nucleico.

También se proporciona en determinados aspectos un aparato que comprende uno o más procesadores y memoria, memoria que comprende instrucciones ejecutables por el uno o más microprocesadores y memoria que comprende recuentos de lecturas de secuencia mapeadas en secciones genómicas de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante, libre de células de una mujer embarazada; e instrucciones ejecutables por el uno o más procesadores que están configuradas para: (a) normalizar los recuentos mapeados en las secciones genómicas del genoma de referencia, proporcionándose de ese modo un perfil de recuentos normalizados para las secciones genómicas; (b) identificar una primera elevación de los recuentos normalizados significativamente diferente de una segunda elevación de los recuentos normalizados en el perfil, primera elevación que es para un primer conjunto de secciones genómicas, y segunda elevación que es para un segundo conjunto de secciones genómicas; (c) determinar un rango de elevación esperado para una variación del número de copias homocigotas y heterocigotas según un valor de incertidumbre para un segmento del genoma; (d) ajustar la primera elevación en un valor predeterminado cuando la primera elevación está dentro de uno de los rangos de elevación esperados, proporcionándose de ese modo un ajuste de la primera elevación; y (e) determinar la presencia o ausencia de una aneuploidía cromosómica en el feto según las elevaciones de secciones genómicas que comprenden el ajuste de (d), mediante lo cual el resultado determinante de la presencia o ausencia de la aneuploidía cromosómica se genera a partir de las lecturas de secuencia de ácido nucleico.

También en determinados aspectos se proporciona un producto de programa informático incorporado de manera tangible en un medio legible por ordenador, que comprende instrucciones que cuando se ejecutan por uno o más procesadores están configuradas para: (a) acceder a recuentos de lecturas de secuencia de ácidos nucleicos mapeadas en secciones genómicas de un genoma de referencia, lecturas de secuencia que son lecturas de ácidos nucleicos circulantes, libres de células de una mujer embarazada; (b) normalizar los recuentos mapeados en las secciones genómicas del genoma de referencia, proporcionándose de ese modo un perfil de recuentos normalizados para las secciones genómicas; (c) identificar una primera elevación de los recuentos normalizados significativamente diferente de una segunda elevación de los recuentos normalizados en el perfil, primera elevación que es para un primer conjunto de secciones genómicas, y segunda elevación que es para un segundo conjunto de secciones genómicas; (d) determinar un rango de elevación esperado para una variación del número de copias homocigotas y heterocigotas según un valor de incertidumbre para un segmento del genoma; (e) ajustar la primera elevación en un valor predeterminado cuando la primera elevación está dentro de uno de los rangos de elevación esperados, proporcionándose de ese modo un ajuste de la primera elevación; y (f) determinar la presencia o ausencia de una aneuploidía cromosómica en el feto según las elevaciones de secciones genómicas que comprenden el ajuste de (e), mediante lo cual el resultado determinante de la presencia o ausencia de la aneuploidía cromosómica se genera a partir de las lecturas de secuencia de ácido nucleico.

También se proporciona en determinados aspectos un sistema que comprende uno o más procesadores y memoria, memoria que comprende instrucciones ejecutables por el uno o más microprocesadores y memoria que comprende recuentos de lecturas de secuencia de ácido nucleico mapeadas en secciones genómicas de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante, libre de células de una mujer embarazada; e instrucciones ejecutables por el uno o más procesadores que están configuradas para: (a) normalizar los recuentos mapeados en las secciones genómicas del genoma de referencia, proporcionándose de ese modo un perfil de recuentos normalizados para las secciones genómicas; (b) identificar una primera elevación de los recuentos normalizados significativamente diferente de una segunda elevación de los recuentos normalizados en el perfil, primera elevación que es para un primer conjunto de secciones genómicas, y segunda elevación que es para un segundo conjunto de secciones genómicas; (c) determinar un rango de elevación esperado para una variación del número de copias homocigotas y heterocigotas según un valor de incertidumbre para un segmento del genoma; y (d) identificar una variación del número de copias materno y/o fetal dentro de la sección genómica basándose en uno de los rangos de elevación esperados, mediante lo cual la variación del número de copias materno y/o fetal se identifica a partir de las lecturas de secuencia de ácido nucleico.

También se proporciona en determinados aspectos un aparato que comprende uno o más procesadores y memoria, memoria que comprende instrucciones ejecutables por el uno o más microprocesadores y memoria que comprende recuentos de lecturas de secuencia de ácido nucleico mapeadas en secciones genómicas de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante, libre de células de una mujer embarazada; e instrucciones ejecutables por el uno o más procesadores que están configuradas para: (a) normalizar los recuentos mapeados en las secciones genómicas del genoma de referencia, proporcionándose de ese modo un perfil de recuentos normalizados para las secciones genómicas; (b) identificar una primera elevación de los recuentos normalizados significativamente diferente de una segunda elevación de los recuentos normalizados en el perfil, primera elevación que es para un primer conjunto de secciones genómicas, y segunda elevación que es para un segundo conjunto de secciones genómicas; (c) determinar un rango de elevación esperado para una variación del número de copias homocigotas y heterocigotas según un valor de incertidumbre para un segmento del genoma; y (d) identificar una variación del número de copias materno y/o fetal dentro de la sección genómica basándose en uno de los rangos de elevación esperados, mediante lo cual la variación del número de copias materno y/o fetal se identifica a partir de las lecturas de secuencia de ácido nucleico.

También se proporciona en determinados aspectos un producto de programa informático incorporado de manera tangible en un medio legible por ordenador, que comprende instrucciones que cuando se ejecutan por uno o más procesadores están configuradas para: (a) acceder a recuentos de lecturas de secuencia de ácidos nucleicos mapeadas en secciones genómicas de un genoma de referencia, lecturas de secuencia que son lecturas de ácidos nucleicos circulantes, libres de células de una mujer embarazada; (b) normalizar los recuentos mapeados en las secciones genómicas del genoma de referencia, proporcionándose de ese modo un perfil de recuentos normalizados para las secciones genómicas; (c) identificar una primera elevación de los recuentos normalizados significativamente diferente de una segunda elevación de los recuentos normalizados en el perfil, primera elevación que es para un primer conjunto de secciones genómicas, y segunda elevación que es para un segundo conjunto de secciones genómicas; (d) determinar un rango de elevación esperado para una variación del número de copias homocigotas y heterocigotas según un valor de incertidumbre para un segmento del genoma; y (e) identificar una variación del número de copias materno y/o fetal dentro de la sección genómica basándose en uno de los rangos de elevación esperados, mediante lo cual la variación del número de copias materno y/o fetal se identifica a partir de las lecturas de secuencia de ácido nucleico.

Se proporciona en algunos aspectos un sistema que comprende uno o más procesadores y memoria, memoria que comprende instrucciones ejecutables por el uno o más microprocesadores y memoria que comprende recuentos de lecturas de secuencia de ácido nucleico mapeadas en secciones genómicas de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante, libre de células de una mujer embarazada; e instrucciones ejecutables por el uno o más procesadores que están configuradas para: (a) normalizar los recuentos mapeados en las secciones genómicas del genoma de referencia, proporcionándose de ese modo un perfil de recuentos normalizados para las secciones genómicas; (b) identificar una primera elevación de los recuentos normalizados significativamente diferente de una segunda elevación de los recuentos normalizados en el perfil, primera elevación que es para un primer conjunto de secciones genómicas, y segunda elevación que es para un segundo conjunto de secciones genómicas; (c) determinar un rango de elevación esperado para una variación del número de copias homocigotas y heterocigotas según un valor de incertidumbre para un segmento del genoma; (d) ajustar la primera elevación según la segunda elevación, proporcionándose de ese modo un ajuste de la primera elevación; y (e) determinar la presencia o ausencia de una aneuploidía cromosómica en el feto según las elevaciones de secciones genómicas que comprenden el ajuste de (d), mediante lo cual el resultado determinante de la presencia o ausencia de la aneuploidía cromosómica se genera a partir de las lecturas de secuencia de ácido nucleico.

Se proporciona en determinados aspectos un aparato que comprende uno o más procesadores y memoria, memoria que comprende instrucciones ejecutables por el uno o más microprocesadores y memoria que comprende recuentos de lecturas de secuencia mapeadas en secciones genómicas de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante, libre de células de una mujer embarazada; e instrucciones ejecutables por el uno o más procesadores que están configuradas para: (a) normalizar los recuentos mapeados en las secciones genómicas del genoma de referencia, proporcionándose de ese modo un perfil de recuentos normalizados para las secciones genómicas; (b) identificar una primera elevación de los recuentos normalizados significativamente diferente de una segunda elevación

de los recuentos normalizados en el perfil, primera elevación que es para un primer conjunto de secciones genómicas, y segunda elevación que es para un segundo conjunto de secciones genómicas; (c) determinar un rango de elevación esperado para una variación del número de copias homocigotas y heterocigotas según un valor de incertidumbre para un segmento del genoma; (d) ajustar la primera elevación según la segunda elevación, proporcionándose de ese modo un ajuste de la primera elevación; y (e) determinar la presencia o ausencia de una aneuploidía cromosómica en el feto según las elevaciones de secciones genómicas que comprenden el ajuste de (d), mediante lo cual el resultado determinante de la presencia o ausencia de la aneuploidía cromosómica se genera a partir de las lecturas de secuencia de ácido nucleico.

En algunos aspectos se proporciona un producto de programa informático incorporado de manera tangible en un medio legible por ordenador, que comprende instrucciones que cuando se ejecutan por uno o más procesadores están configuradas para: (a) acceder a recuentos de lecturas de secuencia de ácidos nucleicos mapeadas en secciones genómicas de un genoma de referencia, lecturas de secuencia que son lecturas de ácidos nucleicos circulantes, libres de células de una mujer embarazada; (b) normalizar los recuentos mapeados en las secciones genómicas del genoma de referencia, proporcionándose de ese modo un perfil de recuentos normalizados para las secciones genómicas; (c) identificar una primera elevación de los recuentos normalizados significativamente diferente de una segunda elevación de los recuentos normalizados en el perfil, primera elevación que es para un primer conjunto de secciones genómicas, y segunda elevación que es para un segundo conjunto de secciones genómicas; (d) determinar un rango de elevación esperado para una variación del número de copias homocigotas y heterocigotas según un valor de incertidumbre para un segmento del genoma; (e) ajustar la primera elevación según una segunda elevación, proporcionándose de ese modo un ajuste de la primera elevación; y (f) determinar la presencia o ausencia de una aneuploidía cromosómica en el feto según las elevaciones de secciones genómicas que comprenden el ajuste de (e), mediante lo cual el resultado determinante de la presencia o ausencia de la aneuploidía cromosómica se genera a partir de las lecturas de secuencia de ácido nucleico.

En determinadas realizaciones, el sistema, aparato y/o producto de programa informático comprende: (i) un módulo de secuenciación configurado para obtener lecturas de secuencia de ácido nucleico; (ii) un módulo de mapeo configurado para mapear lecturas de secuencia de ácido nucleico en porciones de un genoma de referencia; (iii) un módulo de ponderación configurado para ponderar secciones genómicas; (iv) un módulo de filtrado configurado para filtrar secciones genómicas o recuentos mapeados en una sección genómica; (v) un módulo de recuento configurado para proporcionar recuentos de lecturas de secuencia de ácido nucleico mapeadas en porciones de un genoma de referencia; (vi) un módulo de normalización configurado para proporcionar recuentos normalizados; (vii) un módulo de comparación configurado para proporcionar una identificación de una primera elevación que es significativamente diferente de una segunda elevación; (viii) un módulo de establecimiento de rango configurado para proporcionar uno o más rangos de nivel esperados; (ix) un módulo de categorización configurado para identificar una elevación representativa de una variación del número de copias; (x) un módulo de ajuste configurado para ajustar un nivel identificado como una variación del número de copias; (xi) un módulo de representación gráfica configurado para graficar y visualizar un nivel y/o un perfil; (xii) un módulo de resultados configurado para determinar un resultado (por ejemplo, determinante del resultado de la presencia o ausencia de una aneuploidía fetal); (xiii) un módulo de organización de visualización de datos configurado para indicar la presencia o ausencia de una aberración cromosómica segmentaria o una aneuploidía fetal o ambas; (xiv) un módulo de procesamiento lógico configurado para realizar una o más lecturas de secuencia de mapa, contar lecturas de secuencia mapeadas, normalizar recuentos y generar un resultado; o (xv) combinación de dos o más de los anteriores.

En algunas implementaciones, el módulo de secuenciación y el módulo de mapeo están configurados para transferir lecturas de secuencia desde el módulo de secuenciación al módulo de mapeo. El módulo de mapeo y el módulo de recuento a veces están configurados para transferir lecturas de secuencia mapeadas desde el módulo de mapeo al módulo de recuento. A veces, el módulo de recuento y el módulo de filtrado están configurados para transferir recuentos desde el módulo de recuento al módulo de filtrado. A veces, el módulo de recuento y el módulo de ponderación están configurados para transferir recuentos desde el módulo de recuento al módulo de ponderación. El módulo de mapeo y el módulo de filtrado a veces están configurados para transferir lecturas de secuencia mapeadas desde el módulo de mapeo al módulo de filtrado. El módulo de mapeo y el módulo de ponderación a veces están configurados para transferir lecturas de secuencia mapeadas desde el módulo de mapeo al módulo de ponderación. A veces, el módulo de ponderación, el módulo de filtrado y el módulo de recuento están configurados para transferir secciones genómicas filtradas y/o ponderadas desde el módulo de ponderación y el módulo de filtrado al módulo de recuento. A veces, el módulo de ponderación y el módulo de normalización están configurados para transferir secciones genómicas ponderadas desde el módulo de ponderación al módulo de normalización. A veces, el módulo de filtrado y el módulo de normalización están configurados para transferir secciones genómicas filtradas desde el módulo de filtrado al módulo de normalización. En algunas implementaciones, el módulo de normalización y/o módulo de comparación están configurados para transferir recuentos normalizados al módulo de comparación y/o módulo de establecimiento de rango. El módulo de comparación, el módulo de establecimiento de rango y/o el módulo de categorización están configurados independientemente para transferir (i) una identificación de una primera elevación que es significativamente diferente de una segunda elevación y/o (ii) un rango de nivel esperado desde el módulo de comparación y/o el módulo de establecimiento de rango al módulo de categorización, en algunas implementaciones. En determinadas implementaciones, el módulo de categorización y el módulo de ajuste están configurados para transferir una elevación categorizada como una variación del número de copias desde el módulo de categorización hasta el módulo de ajuste. En algunas implementaciones, el módulo de ajuste, el módulo de representación gráfica y el módulo de resultados están configurados para transferir uno o más niveles ajustados desde el módulo de ajuste al módulo de representación gráfica o el módulo de resultados. El módulo de normalización a veces está configurado para

transferir recuentos de lectura de secuencia normalizada mapeados en uno o más del módulo de comparación, el módulo de establecimiento de rango, el módulo de categorización, el módulo de ajuste, el módulo de resultados o el módulo de representación gráfica.

Máquinas, software e interfaces

Determinados procedimientos y métodos descritos en el presente documento (por ejemplo, cuantificación, mapeo, normalización, establecimiento de rango, ajuste, categorización, recuento y/o determinación de lecturas de secuencia, recuentos, elevaciones (por ejemplo, elevaciones) y/o perfiles) no pueden realizarse a menudo sin un ordenador, procesador, software, módulo u otro aparato. Los métodos descritos en el presente documento normalmente son métodos implementados por ordenador, y una o más porciones de un método a veces las realizan uno o más procesadores. Las implementaciones relacionadas con los métodos descritos en este documento son aplicables generalmente al mismo procedimiento o a procedimientos relacionados implementados por instrucciones en los sistemas, aparatos y productos de programa informático descritos en el presente documento. En algunas implementaciones, los procedimientos y métodos descritos en el presente documento (por ejemplo, cuantificación, recuento y/o determinación de lecturas de secuencia, recuentos, elevaciones y/o perfiles de secuencia) se realizan mediante métodos automatizados. En algunas implementaciones, un método automatizado se incorpora en software, módulos, procesadores, periféricos y/o un aparato que comprende similares, que determinan lecturas de secuencia, recuentos, mapeo, etiquetas de secuencia mapeadas, elevaciones, perfiles, normalizaciones, comparaciones, establecimiento de rango, categorización, ajustes, representación gráfica, resultados, transformaciones e identificaciones. Tal como se usa en el presente documento, software se refiere a instrucciones de programas legibles por ordenador que, cuando se ejecutan por un procesador, realizan operaciones informáticas, tal como se describe en el presente documento.

Las lecturas, recuentos, elevaciones y perfiles de secuencias derivados de un sujeto de prueba (por ejemplo, un paciente, una mujer embarazada) y/o de un sujeto de referencia pueden analizarse y procesarse adicionalmente para determinar la presencia o ausencia de una variación genética. Las lecturas, recuentos, elevaciones y/o perfiles de secuencias se denominan, algunas veces, “datos” o “conjuntos de datos”. En algunas implementaciones, los datos o conjuntos de datos pueden caracterizarse por una o más características o variables (por ejemplo, basadas en secuencia [por ejemplo, contenido de GC, secuencia de nucleótidos específica, similares], específicas de función [por ejemplo, genes expresados, genes de cáncer, similares], basadas en la ubicación [específicas de genoma, específicas de cromosoma, específicas de sección genómica o bin], similares y combinaciones de los mismos). En determinadas implementaciones, los datos o conjuntos de datos pueden organizarse en una matriz que tiene dos o más dimensiones basándose en una o más características o variables. Los datos organizados en matrices pueden organizarse usando cualquier característica o variable adecuada.

Un ejemplo no limitativo de datos en una matriz incluye datos organizados por edad materna, ploidía materna y contribución fetal. En determinadas implementaciones, los conjuntos de datos caracterizados por una o más características o variables a veces se procesan después del recuento.

Pueden usarse aparatos, software e interfaces para llevar a cabo los métodos descritos en el presente documento. Con el uso de aparatos, programas e interfaces, un usuario puede introducir, solicitar, consultar o determinar opciones para usar información, programas o procedimientos particulares (por ejemplo, mapear lecturas de secuencia, procesar datos mapeados y/o proporcionar un resultado), que puede implicar implementar algoritmos de análisis estadístico, algoritmos de significación estadística, algoritmos estadísticos, etapas iterativas, algoritmos de validación y representaciones gráficas, por ejemplo. En algunas implementaciones, un usuario puede introducir un conjunto de datos como información de entrada, un usuario puede descargar uno o más conjuntos de datos mediante un medio de hardware adecuado (por ejemplo, unidad flash) y/o un usuario puede enviar un conjunto de datos de un sistema a otro para el procesamiento posterior y/o proporcionar un resultado (por ejemplo, enviar datos de lectura de secuencia de un secuenciador a un sistema informático para el mapeo de lecturas de secuencia; enviar datos de secuencia mapeados en un sistema informático para procesar y producir un resultado y/o informe).

Un sistema comprende normalmente uno o más aparatos. Cada aparato comprende uno o más de memoria, uno o más procesadores e instrucciones. Cuando un sistema incluye dos o más aparatos, algunos o todos los aparatos pueden estar ubicados en la misma ubicación, algunos o todos los aparatos pueden estar ubicados en ubicaciones diferentes, todos los aparatos pueden estar ubicados en una ubicación y/o todos los aparatos pueden estar ubicados en ubicaciones diferentes. Cuando un sistema incluye dos o más aparatos, algunos o todos los aparatos pueden estar ubicados en la misma ubicación que un usuario, algunos o todos los aparatos pueden estar ubicados en una ubicación distinta a un usuario, todos los aparatos pueden estar ubicados en la misma ubicación que el usuario y/o todos los aparatos pueden estar ubicados en una o más ubicaciones diferentes al usuario.

Algunas veces, un sistema comprende un aparato computarizado y un aparato de secuenciación, en el que el aparato de secuenciación está configurado para recibir ácido nucleico físico y generar lecturas de secuencia, y el aparato computarizado está configurado para procesar las lecturas del aparato de secuenciación. Algunas veces, el aparato computarizado está configurado para determinar la presencia o ausencia de una variación genética (por ejemplo, variación del número de copias; aneuploidía de cromosomas fetales) a partir de las lecturas de secuencia.

Por ejemplo, un usuario puede colocar una consulta en un software que, después, puede adquirir un conjunto de datos por medio de acceso a Internet y, en determinadas implementaciones, puede solicitarse a un procesador programable que adquiera un conjunto de datos adecuado basándose en parámetros dados. Un procesador programable también puede solicitar a un usuario que seleccione una o más opciones de conjuntos de datos seleccionadas por el procesador basándose en parámetros dados. Un procesador programable puede solicitar a un usuario que seleccione una o más opciones de conjuntos de datos seleccionadas por el procesador basándose en la información que se encuentra a través de Internet, otra información interna o externa o similares. Las opciones pueden elegirse para seleccionar una o más selecciones de características de datos, uno o más algoritmos estadísticos, uno o más algoritmos de análisis estadístico, uno o más algoritmos de significación estadística, etapas iterativas, uno o más algoritmos de validación y una o más representaciones gráficas de métodos, aparatos o programas informáticos.

Los sistemas abordados en el presente documento pueden comprender componentes generales de sistemas informáticos tales como, por ejemplo, servidores de red, sistemas portátiles, sistemas de escritorio, sistemas de mano, asistentes digitales personales, quioscos informáticos y similares. Un sistema informático puede comprender uno o más medios de entrada tales como un teclado, una pantalla táctil, un ratón, reconocimiento de voz u otros medios para permitir que el usuario introduzca datos en el sistema. Un sistema puede comprender además una o más salidas, incluyendo, pero sin limitación, una pantalla de visualización (por ejemplo, CRT o LCD), un altavoz, una máquina de fax, impresora (por ejemplo, impresora láser, de chorro de tinta, impacto, en blanco y negro o a color) u otra salida útil para proporcionar una salida visual, auditiva y/o impresa de información (por ejemplo, resultado y/o informe).

En un sistema, los medios de entrada y salida pueden conectarse a una unidad central de procesamiento que puede comprender entre otros componentes, un microprocesador para ejecutar instrucciones de programa y memoria para almacenar código de programa y datos. En algunas implementaciones, los procedimientos pueden implementarse como un solo sistema de usuario ubicado en un solo lugar geográfico. En determinadas implementaciones, los procedimientos pueden implementarse como un sistema multiusuario. En el caso de una implementación multiusuario, múltiples unidades centrales de procesamiento pueden conectarse por medio de una red. La red puede ser local, que abarca un único departamento en una parte de un edificio, todo un edificio, abarcar múltiples edificios, abarcar una región, abarcar todo un país o ser mundial. La red puede ser privada, ser propiedad de y estar controlada por un proveedor, o puede implementarse como un servicio basado en Internet en el que el usuario accede a una página web para introducir y recuperar información. En consecuencia, en determinadas implementaciones, un sistema incluye una o más máquinas, que pueden ser locales o remotas con respecto a un usuario. Un usuario puede acceder a más de una máquina en una ubicación o en múltiples ubicaciones, y los datos pueden mapearse y/o procesarse en serie y/o en paralelo. Por tanto, puede usarse una configuración y un control adecuados para mapear y/o procesar datos usando múltiples máquinas, tales como en redes locales, redes remotas y/o plataformas informáticas de tipo "nube".

Un sistema puede incluir una interfaz de comunicaciones en algunas implementaciones. Una interfaz de comunicaciones permite la transferencia de software y datos entre un sistema informático y uno o más dispositivos externos. Los ejemplos no limitativos de interfaces de comunicaciones incluyen un módem, una interfaz de red (tal como una tarjeta Ethernet), un puerto de comunicaciones, una ranura y tarjeta PCMCIA, y similares. El software y los datos transferidos por medio de una interfaz de comunicaciones generalmente están en forma de señales, que pueden ser señales electrónicas, electromagnéticas, ópticas y/u otras señales capaces de recibirse por una interfaz de comunicaciones. A menudo, se proporcionan señales a una interfaz de comunicaciones mediante un canal. Un canal transporta a menudo señales y puede implementarse usando hilo o cable, fibra óptica, una línea telefónica, un enlace de teléfono celular, un enlace de RF y/u otros canales de comunicaciones. Por tanto, en un ejemplo, puede usarse una interfaz de comunicaciones para recibir información de señal que puede detectarse por un módulo de detección de señal.

Los datos pueden introducirse mediante un dispositivo y/o método adecuado incluyendo, pero sin limitarse a, dispositivos de entrada manual o dispositivos de entrada de datos directa (DDE). Los ejemplos no limitativos de dispositivos manuales incluyen teclados, teclados de concepto, pantallas táctiles, lápices ópticos, ratón, bolas de rastreo, palancas de mando, tabletas gráficas, escáneres, cámaras digitales, digitalizadores de vídeo y dispositivos de reconocimiento de voz. Los ejemplos no limitativos de DDE incluyen lectores de código de barras, códigos de tira magnética, tarjetas inteligentes, reconocimiento de caracteres de tinta magnética, reconocimiento de caracteres ópticos, reconocimiento de marcas ópticas y documentos de respuesta.

En algunas implementaciones, la salida de un aparato de secuenciación puede servir como datos que pueden introducirse a través de un dispositivo de entrada. En determinadas implementaciones, las lecturas de secuencia mapeadas pueden servir como datos que pueden introducirse a través de un dispositivo de entrada. En determinadas implementaciones, se generan datos simulados mediante un procedimiento *in silico* y los datos simulados sirven como datos que pueden introducirse a través de un dispositivo de entrada. La expresión "*in silico*" se refiere a la investigación y los experimentos realizados con el uso de un ordenador. Los procedimientos *in silico* incluyen, pero no se limitan a, mapeo de lecturas de secuencia y procesamiento de lecturas de secuencia mapeadas según los procedimientos descritos en la presente invención.

Un sistema puede incluir un software útil para realizar un procedimiento descrito en el presente documento, y el software puede incluir uno o más módulos para realizar tales procedimientos (por ejemplo, módulo de secuenciación, módulo de procesamiento lógico, módulo de organización de visualización de datos). El término "software" se refiere a instrucciones

de programas legibles por ordenador que, cuando se ejecutan por un ordenador, realizan operaciones informáticas. Las instrucciones ejecutables por el uno o más procesadores a veces se proporcionan como código ejecutable, que cuando se ejecutan, pueden hacer que uno o más procesadores implementen un método descrito en el presente documento. Un módulo descrito en el presente documento puede existir como software, e instrucciones (por ejemplo, procesos, rutinas, subrutinas) incorporadas en el software pueden implementarse o realizarse por un procesador. Por ejemplo, un módulo (por ejemplo, un módulo de software) puede formar parte de un programa que realiza un procedimiento o tarea particular. El término "módulo" se refiere a una unidad funcional autónoma que puede usarse en un aparato o sistema de software más grande. Un módulo puede comprender un conjunto de instrucciones para llevar a cabo una función del módulo. Un módulo puede transformar información y/o datos. La información y/o los datos pueden estar en una forma adecuada. Por ejemplo, la información y/o los datos pueden ser digitales o analógicos. En algunos casos, la información y/o los datos pueden ser paquetes, bytes, caracteres o bits. En algunas implementaciones, la información y/o los datos pueden ser cualquier información o dato recopilado, ensamblado o utilizable. Los ejemplos no limitativos de información y/o datos incluyen un medio adecuado, imágenes, vídeo, sonido (por ejemplo, frecuencias, audibles o no audibles), números, constantes, un valor, objetos, tiempo, funciones, instrucciones, mapas, referencias, secuencias, lecturas, lecturas mapeadas, elevaciones, rangos, umbrales, señales, visualizaciones, representaciones o transformaciones de los mismos. Un módulo puede aceptar o recibir información y/o datos, transformar la información y/o los datos en una segunda forma y proporcionar o transferir la segunda forma a un aparato, periférico, componente u otro módulo. Un módulo puede realizar una o más de las siguientes funciones no limitativas: mapear lecturas de secuencia, proporcionar recuentos, ensamblar secciones genómicas, proporcionar o determinar una elevación, proporcionar un perfil de recuento, normalizar (por ejemplo, normalizar lecturas, normalizar recuentos, y similares), proporcionar un perfil de recuento normalizado o elevaciones de recuentos normalizados, comparar dos o más elevaciones, proporcionar valores de incertidumbre, proporcionar o determinar elevaciones esperadas y rangos esperados (por ejemplo, rangos de elevación esperados, rangos umbral y elevaciones umbral), proporcionar ajustes a las elevaciones (por ejemplo, ajustar una primera elevación, ajustar una segunda elevación, ajustar un perfil de un cromosoma o un segmento del mismo y/o relleno), proporcionar identificación (por ejemplo, identificar una variación del número de copias, variación genética o aneuploidía), categorizar, representar gráficamente y/o determinar un resultado, por ejemplo. Un procesador puede, en algunos casos, llevar a cabo las instrucciones en un módulo. En algunas implementaciones, se requiere que uno o más procesadores lleven a cabo instrucciones en un módulo o grupo de módulos. Un módulo puede proporcionar información y/o datos a otro módulo, aparato o fuente y puede recibir información y/o datos de otro módulo, aparato o fuente.

Algunas veces, un producto de programa informático se incorpora en un medio legible por ordenador tangible y, algunas veces, se incorpora en un medio legible por ordenador no transitorio. Algunas veces, un módulo se almacena en un medio legible por ordenador (por ejemplo, disco, unidad) o en memoria (por ejemplo, memoria de acceso aleatorio). Un módulo y procesador capaces de implementar instrucciones de un módulo pueden estar ubicados en un aparato o en diferentes aparatos. Un módulo y/o procesador capaces de implementar una instrucción para un módulo pueden estar ubicados en la misma ubicación que un usuario (por ejemplo, red local) o en una ubicación diferente de un usuario (por ejemplo, red remota, sistema de tipo nube). En implementaciones en las que se lleva a cabo un método junto con dos o más módulos, los módulos pueden estar ubicados en el mismo aparato, uno o más módulos pueden estar ubicados en diferentes aparatos en la misma ubicación física, y uno o más módulos pueden estar ubicados en diferentes aparatos en ubicaciones físicas diferentes.

Un aparato, en algunas implementaciones, comprende al menos un procesador para llevar a cabo las instrucciones en un módulo. A veces, se accede a los recuentos de lecturas de secuencia mapeadas en secciones genómicas de un genoma de referencia por un procesador que ejecuta instrucciones configuradas para llevar a cabo un método descrito en el presente documento. Los recuentos a los que se accede por un procesador pueden estar dentro de la memoria de un sistema, y puede accederse a los recuentos y colocarse en la memoria del sistema después de que se obtienen. En algunas implementaciones, un aparato incluye un procesador (por ejemplo, uno o más procesadores), procesador que puede realizar y/o implementar una o más instrucciones (por ejemplo, procesos, rutinas y/o subrutinas) de un módulo. En algunas implementaciones, un aparato incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunas implementaciones, un aparato funciona con uno o más procesadores externos (por ejemplo, una red interna o externa, un servidor, dispositivo de almacenamiento y/o una red de almacenamiento (por ejemplo, una nube)). En algunas implementaciones, un aparato comprende un módulo. Algunas veces, un aparato comprende uno o más módulos. Un aparato que comprende un módulo puede recibir y transferir a menudo uno o más de información y/o datos a y desde otros módulos. En algunos casos, un aparato comprende periféricos y/o componentes. Algunas veces, un aparato puede comprender uno o más periféricos o componentes que pueden transferir información y/o datos a y desde otros módulos, periféricos y/o componentes. Algunas veces, un aparato interacciona con un periférico y/o componente que proporciona información y/o datos. Algunas veces, los periféricos y los componentes ayudan a un aparato a realizar una función o interaccionar directamente con un módulo. Los ejemplos no limitativos de periféricos y/o componentes incluyen un periférico de ordenador, método o dispositivo de almacenamiento o E/S adecuados incluyendo, pero sin limitarse a, escáneres, impresoras, pantallas de visualización (por ejemplo, monitores, LED, LOT o CRT), cámaras, micrófonos, paneles (por ejemplo, ipad, tabletas), pantallas táctiles, teléfonos inteligentes, teléfonos móviles, dispositivos de E/S USB, dispositivos de almacenamiento masivo USB, teclados, un ratón para ordenador, lápices digitales, módems, discos duros, memorias USB, unidades flash, un procesador, un servidor, CD, DVD, tarjetas gráficas, dispositivos de E/S especializados (por ejemplo, secuenciadores, fotoceldas, tubos fotomultiplicadores, lectores ópticos, sensores, etc.), una o más celdas de flujo, componentes de manipulación de fluidos, controladores de interfaz de red, ROM, RAM, métodos y dispositivos de transferencia inalámbrica (Bluetooth, WiFi, y similares), la red mundial (www), Internet, un ordenador y/u otro módulo.

Uno o más de un módulo de secuenciación, un módulo de procesamiento lógico y un módulo de organización de visualización de datos pueden usarse en un método descrito en el presente documento. Algunas veces, un módulo de procesamiento lógico, módulo de secuenciación o módulo de organización de visualización de datos, o un aparato que comprende uno o más de tales módulos, recopilan, ensamblan, reciben, proporcionan y/o transfieren información y/o datos a o desde otro módulo, aparato, componente, periférico u operador de un aparato. Por ejemplo, algunas veces un operador de un aparato proporciona una constante, un valor umbral, una fórmula o un valor predeterminado a un módulo de procesamiento lógico, módulo de secuenciación o módulo de organización de visualización de datos. Un módulo de procesamiento lógico, módulo de secuenciación o módulo de organización de visualización de datos pueden recibir información y/o datos de otro módulo, los ejemplos no limitativos de los mismos incluyen un módulo de procesamiento lógico, módulo de secuenciación, módulo de organización de visualización de datos, módulo de secuenciación, módulo de secuenciación, módulo de mapeo, módulo de recuento, módulo de normalización, módulo de comparación, módulo de establecimiento de rango, módulo de categorización, módulo de ajuste, módulo de representación gráfica, módulo de resultados, módulo de organización de visualización de datos y/o módulo de procesamiento lógico, similares o combinaciones de los mismos. La información y/o los datos derivados de o transformados por un módulo de procesamiento lógico, módulo de secuenciación o módulo de organización de visualización de datos pueden transferirse de un módulo de procesamiento lógico, módulo de secuenciación o módulo de organización de visualización de datos a un módulo de secuenciación, módulo de secuenciación, módulo de mapeo, módulo de recuento, módulo de normalización, módulo de comparación, módulo de establecimiento de rango, módulo de categorización, módulo de ajuste, módulo de representación gráfica, módulo de resultados, módulo de organización de visualización de datos, módulo de procesamiento lógico u otro aparato y/o módulo adecuado. Un módulo de secuenciación puede recibir información y/o datos forma un módulo de procesamiento lógico y/o módulo de secuenciación y transferir información y/o datos a un módulo de procesamiento lógico y/o un módulo de mapeo, por ejemplo. Algunas veces, un módulo de procesamiento lógico orquesta, controla, limita, organiza, ordena, distribuye, divide, transforma y/o regula información y/o datos o la transferencia de información y/o datos a y desde uno o más de otros módulos, periféricos o dispositivos. Un módulo de organización de visualización de datos puede recibir información y/o datos de un módulo de procesamiento lógico y/o módulo de representación gráfica y transferir información y/o datos a un módulo de procesamiento lógico, módulo de representación gráfica, pantalla de visualización, periférico o dispositivo. Un aparato que comprende un módulo de procesamiento lógico, módulo de secuenciación o módulo de organización de visualización de datos puede comprender al menos un procesador. En algunas implementaciones, se proporcionan información y/o datos por un aparato que incluye un procesador (por ejemplo, uno o más procesadores), procesador que puede realizar y/o implementar una o más instrucciones (por ejemplo, procesos, rutinas y/o subrutinas) desde el módulo de procesamiento lógico, módulo de secuenciación y/o módulo de organización de visualización de datos. En algunas implementaciones, un módulo de procesamiento lógico, módulo de secuenciación o módulo de organización de visualización de datos funciona con uno o más procesadores externos (por ejemplo, una red interna o externa, un servidor, dispositivo de almacenamiento y/o una red de almacenamiento (por ejemplo, una nube)).

El software se proporciona a menudo en un producto de programa que contiene instrucciones de programa registradas en un medio legible por ordenador incluyendo, pero sin limitarse a, medios magnéticos que incluyen disquetes, discos duros y cinta magnética; y medios ópticos que incluyen discos CD-ROM, discos DVD, discos magneto-ópticos, unidades flash, RAM, disquetes, similares, y otros medios similares en los que pueden registrarse las instrucciones del programa. En la implementación en línea, un servidor y un sitio web mantenidos por una organización pueden estar configurados para proporcionar descargas de software a usuarios remotos, o los usuarios remotos pueden acceder a un sistema remoto mantenido por una organización para acceder de manera remota al software. El software puede obtener o recibir información de entrada. El software puede incluir un módulo que obtiene o recibe de manera específica datos (por ejemplo, un módulo receptor de datos que recibe datos leídos de secuencias y/o datos leídos mapeados) y puede incluir un módulo que procesa de manera específica los datos (por ejemplo, un módulo de procesamiento que procesa los datos recibidos (por ejemplo, filtra, normaliza, proporciona un resultado y/o informe). Los términos "obtener" y "recibir" información de entrada se refieren a recibir datos (por ejemplo, lecturas de secuencia, lecturas mapeadas) mediante medios de comunicación por ordenador desde un sitio local, o remoto, entrada de datos por seres humanos, o cualquier otro método para recibir datos. La información de entrada puede generarse en la misma ubicación en la que se recibe, o puede generarse en una ubicación diferente y transmitirse a la ubicación de recepción. En algunas implementaciones, la información de entrada se modifica antes de procesarse (por ejemplo, se coloca en un formato susceptible al procesamiento (por ejemplo, tabulado)).

En algunas implementaciones, se proporcionan productos de programa informático, tales como, por ejemplo, un producto de programa informático que comprende un medio utilizable por ordenador que tiene un código de programa legible por ordenador incorporado en el mismo, estando el código de programa legible por ordenador adaptado para ejecutarse para implementar un método que comprende: (a) obtener lecturas de secuencia de ácido nucleico de muestra de un sujeto de prueba; (b) mapear las lecturas de secuencia obtenidas en (a) en un genoma conocido, genoma conocido que se ha dividido en secciones genómicas; (c) contar las lecturas de secuencia mapeadas dentro de las secciones genómicas; (d) generar un perfil de recuento normalizado de muestra normalizando los recuentos para las secciones genómicas obtenidas en (c); y (e) determinar la presencia o ausencia de una variación genética del perfil de recuento normalizado de muestra en (d).

El software puede incluir uno o más algoritmos en determinadas implementaciones. Un algoritmo puede usarse para procesar datos y/o proporcionar un resultado o informe según una secuencia finita de instrucciones. Un algoritmo es a

menudo una lista de instrucciones definidas para completar una tarea. Partiendo de un estado inicial, las instrucciones pueden describir un cálculo que avanza a través de una serie definida de estados sucesivos, terminando eventualmente en un estado final. La transición de un estado al siguiente no es necesariamente determinista (por ejemplo, algunos algoritmos incorporan aleatoriedad). A modo de ejemplo, y sin limitación, un algoritmo puede ser un algoritmo de búsqueda, algoritmo de clasificación, algoritmo de fusión, algoritmo numérico, algoritmo gráfico, algoritmo de cadena, algoritmo de modelado, algoritmo de genometría computacional, algoritmo combinatorio, algoritmo de aprendizaje automático, algoritmo de criptografía, algoritmo de compresión de datos, algoritmo de análisis y similares. Un algoritmo puede incluir un algoritmo o dos o más algoritmos que funcionan en combinación. Un algoritmo puede ser de cualquier clase de complejidad adecuada y/o complejidad parametrizada. Puede usarse un algoritmo para el cálculo y/o procesamiento de datos y, en algunas implementaciones, puede usarse en un enfoque determinista o probabilístico/predictivo. Puede implementarse un algoritmo en un entorno informático usando un lenguaje de programación adecuado, son ejemplos no limitativos de los mismos C, C++, Java, Perl, Python, Fortran, y similares. En algunas implementaciones, un algoritmo puede configurarse o modificarse para incluir margen de errores, análisis estadístico, significación estadística y/o comparación con otra información o conjuntos de datos (por ejemplo, aplicable cuando se usa una red neural o algoritmo de agrupamiento).

En determinadas implementaciones, pueden implementarse varios algoritmos para su uso en software. Estos algoritmos pueden entrenarse con datos sin procesar en algunas implementaciones. Para cada nueva muestra de datos sin procesar, los algoritmos entrenados pueden producir un conjunto o resultado de datos procesados representativos. Un conjunto de datos procesados algunas veces tiene una complejidad reducida en comparación con el conjunto de datos original que se procesó. Basándose en un conjunto procesado, el rendimiento de un algoritmo entrenado puede evaluarse basándose en la sensibilidad y especificidad, en algunas implementaciones. Un algoritmo con la mayor sensibilidad y/o especificidad puede identificarse y usarse, en determinadas implementaciones.

En determinadas implementaciones, los datos simulados (o simulación) pueden ayudar al procesamiento de datos, por ejemplo, entrenando un algoritmo o someter a prueba un algoritmo. En algunas implementaciones, los datos simulados incluyen varios muestreos hipotéticos de agrupamientos diferentes de lecturas de secuencia. Los datos simulados pueden basarse en lo que podría esperarse de una población real o pueden sesgarse para someter a prueba un algoritmo y/o asignar una clasificación correcta. Los datos simulados se denominan además en el presente documento datos “virtuales”. Las simulaciones pueden realizarse mediante un programa informático en determinadas implementaciones. Una etapa posible en el uso de un conjunto de datos simulados es evaluar la confianza de un resultado identificado, por ejemplo, cuán bien coincide un muestreo aleatorio o representa mejor los datos originales. Un enfoque consiste en calcular un valor de probabilidad (valor de p) que estima la probabilidad de una muestra aleatoria con mejor puntuación que las muestras seleccionadas. En algunas implementaciones, puede evaluarse un modelo empírico, en el cual se supone que al menos una muestra coincide con una muestra de referencia (con o sin variaciones resueltas). En algunas implementaciones, otra distribución, tal como una distribución de Poisson, por ejemplo, puede usarse para definir la distribución de probabilidad.

Un sistema puede incluir uno o más procesadores en determinadas implementaciones. Un procesador puede conectarse a un bus de comunicaciones. Un sistema informático puede incluir una memoria principal, a menudo memoria de acceso aleatorio (RAM), y puede incluir además una memoria secundaria. La memoria en algunas implementaciones comprende un medio de almacenamiento no transitorio legible por ordenador. La memoria secundaria puede incluir, por ejemplo, una unidad de disco duro y/o una unidad de almacenamiento extraíble, que representa una unidad de disquete, una unidad de cinta magnética, una unidad de disco óptico, tarjeta de memoria y similares. Una unidad de almacenamiento extraíble a menudo lee y/o escribe en una unidad de almacenamiento extraíble. Los ejemplos no limitativos de unidades de almacenamiento extraíbles incluyen un disquete, una cinta magnética, un disco óptico y similares, que pueden leerse y escribirse, por ejemplo, en una unidad de almacenamiento extraíble. Una unidad de almacenamiento extraíble puede incluir un medio de almacenamiento utilizable por ordenador que tiene almacenados un software y/o datos informáticos.

Un procesador puede implementar software en un sistema. En algunas implementaciones, un procesador puede programarse para realizar automáticamente una tarea descrita en el presente documento que un usuario podría realizar. En consecuencia, un procesador, o algoritmo conducido por tal procesador, puede requerir poca o ninguna supervisión o entrada de un usuario (por ejemplo, el software puede programarse para implementar una función automáticamente). En algunas implementaciones, la complejidad de un procedimiento es tan grande que una sola persona o grupo de personas no podría realizar el procedimiento en un marco de tiempo lo suficientemente corto para determinar la presencia o ausencia de una variación genética.

En algunas implementaciones, la memoria secundaria puede incluir otros medios similares para permitir que los programas informáticos u otras instrucciones se carguen en un sistema informático. Por ejemplo, un sistema puede incluir una unidad de almacenamiento extraíble y un dispositivo de interfaz. Los ejemplos no limitativos de tales sistemas incluyen un cartucho de programa e interfaz de cartucho (tales como los que se encuentran en dispositivos de videojuegos), un chip de memoria extraíble (tal como una EPROM o PROM) y un enchufe asociado, y otras unidades de almacenamiento extraíbles e interfaces que permiten que el software y los datos se transfieran desde la unidad de almacenamiento extraíble a un sistema informático.

Una entidad puede generar recuentos de lecturas de secuencia, mapear las lecturas de secuencia en secciones genómicas, contar las lecturas mapeadas y utilizar las lecturas mapeadas contadas en un método, sistema, aparato o producto de programa informático descrito en el presente documento, en algunas implementaciones. Los

recuentos de lecturas de secuencia mapeadas en secciones genómicas a veces se transfieren por una entidad a una segunda entidad para su uso por la segunda entidad en un método, sistema, aparato o producto de programa informático descrito en el presente documento, en determinadas implementaciones.

En algunas implementaciones, una entidad genera lecturas de secuencia y una segunda entidad mapea esas lecturas de secuencia en secciones genómicas en un genoma de referencia en algunas implementaciones. La segunda entidad cuenta, a veces, las lecturas mapeadas y utiliza las lecturas mapeadas contadas en un método, sistema, aparato o producto de programa informático descrito en el presente documento. A veces la segunda entidad transfiere las lecturas mapeadas a una tercera entidad, y la tercera entidad cuenta las lecturas mapeadas y utiliza las lecturas mapeadas en un método, sistema, aparato o producto de programa informático descrito en el presente documento. Algunas veces, la segunda entidad cuenta las lecturas mapeadas y transfiere las lecturas mapeadas contadas a una tercera entidad, y la tercera entidad utiliza las lecturas mapeadas contadas en un método, sistema, aparato o producto de programa informático descrito en el presente documento. En implementaciones que involucran una tercera entidad, la tercera entidad a veces es la misma que la primera entidad. Es decir, la primera entidad transfiere, a veces, lecturas de secuencia a una segunda entidad, segunda entidad que puede mapear lecturas de secuencia en secciones genómicas en un genoma de referencia y/o contar las lecturas mapeadas, y la segunda entidad puede transferir las lecturas mapeadas y/o contadas a una tercera entidad. Una tercera entidad a veces puede utilizar las lecturas mapeadas y/o contadas en un método, sistema, aparato o producto informático descrito en el presente documento, en el que la tercera entidad a veces es la misma que la primera entidad y, a veces, la tercera entidad es diferente de la primera o segunda entidad.

En algunas implementaciones, una entidad obtiene sangre de una mujer embarazada, opcionalmente, aísla ácido nucleico de la sangre (por ejemplo, del plasma o suero) y transfiere la sangre o ácido nucleico a una segunda entidad que genera lecturas de secuencia del ácido nucleico.

Variaciones genéticas y afecciones médicas

La presencia o ausencia de una varianza genética puede determinarse usando un método o aparato descrito en el presente documento. En determinadas implementaciones, la presencia o ausencia de una o más variaciones genéticas se determina según un resultado proporcionado mediante los métodos y aparatos descritos en el presente documento. Una variación genética es generalmente un fenotipo genético particular presente en determinados individuos y a menudo una variación genética está presente en una subpoblación estadísticamente significativa de individuos. En algunas implementaciones, una variación genética es una anomalía cromosómica (por ejemplo, aneuploidía), anomalía cromosómica parcial o mosaicismo, cada uno de los cuales se describe con mayor detalle en el presente documento. Los ejemplos no limitativos de variaciones genéticas incluyen una o más deleciones (por ejemplo, microdeleciones), duplicaciones (por ejemplo, microduplicaciones), inserciones, mutaciones, polimorfismos (por ejemplo, polimorfismos de un solo nucleótido), fusiones, repeticiones (por ejemplo, repeticiones cortas en tándem), distintos sitios de metilación, distintos patrones de metilación, similares y combinaciones de los mismos. Una inserción, repetición, deleción, duplicación, mutación o polimorfismo puede ser de cualquier longitud, y en algunas implementaciones, tiene aproximadamente 1 base o par de bases (pb) a aproximadamente 250 megabases (Mb) de longitud. En algunas implementaciones, una inserción, repetición, deleción, duplicación, mutación o polimorfismo tiene de aproximadamente 1 base o par de bases (pb) a aproximadamente 1000 kilobases (kb) de longitud (por ejemplo, aproximadamente 10 pb, 50 pb, 100 pb, 500 pb, 1 kb, 5 kb, 10 kb, 50 kb, 100 kb, 500 kb o 1000 kb de longitud).

Una variación genética es algunas veces una deleción. Algunas veces, una deleción es una mutación (por ejemplo, una aberración genética) en la que falta una parte de un cromosoma o una secuencia de ADN. Una deleción es a menudo la pérdida de material genético. Puede eliminarse cualquier número de nucleótidos. Una deleción puede comprender la deleción de uno o más cromosomas completos, un segmento de un cromosoma, un alelo, un gen, un intrón, un exón, cualquier región no codificante, cualquier región codificante, un segmento de los mismos o combinación de los mismos. Una deleción puede comprender una microdeleción. Una deleción puede comprender la deleción de una sola base.

Algunas veces, una variación genética es una duplicación genética. Algunas veces, una duplicación es una mutación (por ejemplo, una aberración genética) en la que una parte de un cromosoma o una secuencia de ADN se copia y se inserta de nuevo en el genoma. Algunas veces, una duplicación genética (es decir, duplicación) es cualquier duplicación de una región de ADN. En algunas implementaciones, una duplicación es una secuencia de ácido nucleico que se repite a menudo en tándem, dentro de un genoma o cromosoma. En algunas implementaciones, una duplicación puede comprender una copia de uno o más cromosomas enteros, un segmento de un cromosoma, un alelo, un gen, un intrón, un exón, cualquier región no codificante, cualquier región codificante, segmento de los mismos o combinación de los mismos. Una duplicación puede comprender una microduplicación. Una duplicación comprende, algunas veces, una o más copias de un ácido nucleico duplicado. Una duplicación a veces se caracteriza como una región genética repetida una o más veces (por ejemplo, repetida 1, 2, 3, 4, 5, 6, 7, 8, 9 o 10 veces). Las duplicaciones pueden variar desde regiones pequeñas (miles de pares de bases) hasta cromosomas enteros en algunos casos. Las duplicaciones se producen a menudo como resultado de un error en la recombinación homóloga o debido a un evento de retrotransposón. Las duplicaciones se han asociado con determinados tipos de enfermedades proliferativas. Las duplicaciones pueden caracterizarse con el uso de microalineamientos genómicos o hibridación genética comparativa (HGC).

Una variación genética es, algunas veces, una inserción. Una inserción es, algunas veces, la adición de uno o más pares de bases de nucleótidos en una secuencia de ácido nucleico. Una inserción es, algunas veces, una

microinserción. Algunas veces, una inserción comprende la adición de un segmento de un cromosoma en un genoma, cromosoma o segmento de los mismos. Algunas veces, una inserción comprende la adición de un alelo, un gen, un intrón, un exón, cualquier región no codificante, cualquier región codificante, segmento de los mismos o combinación de los mismos en un genoma o segmento de los mismos. Algunas veces, una inserción comprende la adición (es decir, inserción) de ácido nucleico de origen desconocido en un genoma, cromosoma o segmento de los mismos. Algunas veces, una inserción comprende la adición (es decir, inserción) de una sola base.

Tal como se usa en el presente documento, una "variación del número de copias" es generalmente una clase o un tipo de variación genética o aberración cromosómica. Una variación del número de copias puede ser una deleción (por ejemplo, microdeleción), duplicación (por ejemplo, una microduplicación) o inserción (por ejemplo, una microinserción). A menudo, el prefijo "micro", tal como se usa en el presente documento, algunas veces es un segmento de ácido nucleico con una longitud menor de 5 Mb. Una variación del número de copias puede incluir una o más deleciones (por ejemplo, microdeleción), duplicaciones y/o inserciones (por ejemplo, una microduplicación, microinserción) de un segmento de un cromosoma. En algunos casos, una duplicación comprende una inserción. Algunas veces, una inserción es una duplicación. Algunas veces, una inserción no es una duplicación. Por ejemplo, a menudo una duplicación de una secuencia en una sección genómica aumenta los recuentos para una sección genómica en la que se encuentra la duplicación. A menudo, una duplicación de una secuencia en una sección genómica aumenta la elevación. A veces, una duplicación presente en secciones genómicas que componen una primera elevación aumenta la elevación con respecto a una segunda elevación en la que una duplicación está ausente. Algunas veces, una inserción aumenta los recuentos de una sección genómica y una secuencia que representa la inserción está presente (es decir, duplicada) en otra ubicación dentro de la misma sección genómica. A veces, una inserción no aumenta significativamente los recuentos de una sección o elevación genómica y la secuencia que se inserta no es una duplicación de una secuencia dentro de la misma sección genómica. A veces, una inserción no se detecta o representa como una duplicación y una secuencia duplicada que representa la inserción no está presente en la misma sección genómica.

En algunas implementaciones, una variación del número de copias es una variación del número de copias fetal. A menudo, una variación del número de copias fetal es una variación del número de copias en el genoma de un feto. En algunas implementaciones, una variación del número de copias es una variación del número de copias materno. Algunas veces, una variación del número de copias materno y/o fetal es una variación del número de copias dentro del genoma de una mujer embarazada (por ejemplo, una mujer que porta un feto), una mujer que da a luz o una mujer capaz de portar un feto. Una variación del número de copias puede ser una variación del número de copias heterocigotas en la que la variación (por ejemplo, una duplicación o deleción) está presente en un alelo de un genoma. Una variación del número de copias puede ser una variación del número de copias homocigotas en la que la variación está presente en ambos alelos de un genoma. En algunas implementaciones, una variación del número de copias es una variación del número de copias heterocigotas u homocigotas fetal. En algunas implementaciones, una variación del número de copias es una variación del número de copias heterocigotas u homocigotas materno y/o fetal. Algunas veces, una variación del número de copias está presente en un genoma materno y un genoma fetal, un genoma materno y no en un genoma fetal, o un genoma fetal y no en un genoma materno.

"Ploidía" se refiere al número de cromosomas presentes en un feto o su madre. Algunas veces, "ploidía" es lo mismo que "ploidía cromosómica". En seres humanos, por ejemplo, los cromosomas autosómicos están presentes a menudo en pares. Por ejemplo, en ausencia de una variación genética, la mayoría de los seres humanos tienen dos de cada cromosoma autosómico (por ejemplo, cromosomas 1-22). La presencia del complemento normal de 2 cromosomas autosómicos en un ser humano se denomina a menudo euploide. "Microploidía" es similar en significado a ploidía. "Microploidía" se refiere a menudo a la ploidía de un segmento de un cromosoma. El término "microploidía" a veces se refiere a la presencia o ausencia de una variación del número de copias (por ejemplo, una deleción, duplicación y/o una inserción) dentro de un cromosoma (por ejemplo, una deleción, duplicación o inserción homocigota o heterocigota, similares o ausencia de las mismas). "Ploidía" y "microploidía" a veces se determinan después de la normalización de los recuentos de una elevación en un perfil (por ejemplo, después de normalizar los recuentos de una elevación a un NRV de 1). Por tanto, una elevación que representa un par cromosómico autosómico (por ejemplo, un euploide) se normaliza a menudo con respecto a un NRV de 1 y se denomina ploidía de 1. De manera similar, una elevación dentro de un segmento de un cromosoma que representa la ausencia de una duplicación, deleción o inserción se normaliza a menudo a un NRV de 1 y se denomina microploidía de 1. A menudo, la ploidía y la microploidía son específicas de bins (por ejemplo, específicas de sección genómica) y específicas de muestra. La ploidía se define a menudo como múltiplos enteros de 14, representando los valores de 1,14, 0, 3/2 y 2 euploidía (por ejemplo, 2 cromosomas), 1 cromosoma presente (por ejemplo, una deleción cromosómica), ningún cromosoma presente, 3 cromosomas (por ejemplo, una trisomía) y 4 cromosomas, respectivamente. De manera similar, la microploidía se define a menudo como múltiplos enteros de 14, representando los valores de 1, ½, 0, 3/2 y 2 euploidía (por ejemplo, sin variación en el número de copias), una deleción heterocigota, deleción homocigota, duplicación heterocigota y duplicación homocigota, respectivamente. Algunos ejemplos de valores de ploidía para un feto se proporcionan en la tabla 2 para un NRV de 1.

Algunas veces, la microploidía de un feto coincide con la microploidía de la madre del feto (es decir, el sujeto femenino gestante). A veces la microploidía de un feto coincide con la microploidía de la madre del feto y tanto la madre como el feto portan la misma variación del número de copias heterocigotas, la variación del número de copias

homocigotas o ambos son euploides. Algunas veces la microploidía de un feto es diferente de la microploidía de la madre del feto. Por ejemplo, a veces la microploidía de un feto es heterocigota para una variación del número de copias, la madre es homocigota para una variación del número de copias y la microploidía del feto no coincide (por ejemplo, no es igual) con la microploidía de la madre para la variación del número de copias especificada.

Una microploidía se asocia a menudo con una elevación esperada. Por ejemplo, algunas veces una elevación (por ejemplo, una elevación en un perfil, algunas veces una elevación que no incluye sustancialmente ninguna variación del número de copias) se normaliza con respecto a un NRV de 1 y la microploidía de una duplicación homocigota es de 2, una duplicación heterocigota es de 1,5, una deleción heterocigota es de 0,5 y una deleción homocigota es de cero.

Una variación genética para la cual se identifica la presencia o ausencia para un sujeto se asocia con una afección médica en determinadas implementaciones. Por tanto, la tecnología descrita en el presente documento puede usarse para identificar la presencia o ausencia de una o más variaciones genéticas asociadas con una afección médica o un estado médico. Los ejemplos no limitativos de afecciones médicas incluyen las asociadas con discapacidad intelectual (por ejemplo, síndrome de Down), proliferación celular aberrante (por ejemplo, cáncer), presencia de un ácido nucleico de microorganismos (por ejemplo, virus, bacteria, hongo, levadura) y preeclampsia.

Los ejemplos no limitativos de variaciones genéticas, afecciones y estados médicos se describen más adelante.

Sexo del feto

En algunas implementaciones, la predicción de un sexo del feto o trastorno relacionado con el sexo (por ejemplo, aneuploidía de cromosoma sexual) puede determinarse mediante un método o aparato descrito en el presente documento. La determinación del sexo se basa generalmente en un cromosoma sexual. En seres humanos, hay dos cromosomas sexuales, los cromosomas X e Y. El cromosoma Y contiene un gen, SRY, que desencadena el desarrollo embrionario como hombre. Los cromosomas Y de seres humanos y otros mamíferos contienen además otros genes necesarios para la producción normal de esperma. Los individuos con XX son de sexo femenino y XY son de sexo masculino y variaciones no limitativas, a menudo denominadas aneuploidias en cromosomas sexuales, incluyen X0, XYY, XXX y XXY. En algunos casos, los hombres tienen dos cromosomas X y un cromosoma Y (XXY; síndrome de Klinefelter), o un cromosoma X y dos cromosomas Y (síndrome XYY; síndrome de Jacobs), y algunas mujeres tienen tres cromosomas X (XXX; síndrome triple X) o un solo cromosoma X en lugar de dos (X0; síndrome de Turner). En algunos casos, solo una porción de las células en un individuo se ve afectada por una aneuploidía en cromosomas sexuales, que puede denominarse mosaicismo (por ejemplo, mosaicismo de Turner). Otros casos incluyen aquellos en los que SRY se daña (lo que conduce a una mujer XY) o se copia en la X (lo que conduce a un hombre XX).

En determinados casos, puede ser beneficioso determinar el sexo de un feto en el útero. Por ejemplo, un paciente (por ejemplo, mujer embarazada) con antecedentes familiares de uno o más trastornos ligados al sexo puede desear determinar el sexo del feto que porta para ayudar a evaluar el riesgo de que el feto herede tal trastorno. Los trastornos ligados al sexo incluyen, sin limitación, trastornos ligados al cromosoma X e Y. Los trastornos ligados al cromosoma X incluyen trastornos recesivos ligados al cromosoma X y dominantes ligados al cromosoma X. Los ejemplos de trastornos recesivos ligados al cromosoma X incluyen, sin limitación, trastornos inmunitarios (por ejemplo, enfermedad granulomatosa crónica (CYBB), síndrome de Wiskott-Aldrich, inmunodeficiencia combinada severa ligada al cromosoma X, agammaglobulinemia ligada al cromosoma X, síndrome de hiper-IgM tipo 1, IPEX, enfermedad linfoproliferativa ligada al cromosoma X, deficiencia de properdina), trastornos hematológicos (por ejemplo, hemofilia A, hemofilia B, anemia sideroblástica ligada al cromosoma X), trastornos endocrinos (por ejemplo, síndrome de insensibilidad a andrógenos/enfermedad de Kennedy, síndrome de Kallmann KAL1, hipoplasia suprarrenal congénita ligada al cromosoma X), trastornos metabólicos (por ejemplo, deficiencia de ornitina transcarbamilasa, síndrome oculocerebrorrenal, adrenoleucodistrofia, deficiencia de glucosa-6-fosfato deshidrogenasa, deficiencia de piruvato deshidrogenasa, enfermedad de Danon/glucogenosis tipo lib, enfermedad de Fabry, síndrome de Hunter, síndrome de Lesch-Nyhan, enfermedad de Menkes/síndrome de asta occipital), trastornos del sistema nervioso (por ejemplo, síndrome de Coffin-Lowry, síndrome MASA, síndrome de alfa talasemia con retraso mental ligado al cromosoma X, síndrome de Siderius con retraso mental ligado al cromosoma X, daltonismo, albinismo ocular, enfermedad de Norrie, coroideremia, enfermedad de Charcot-Marie-Tooth (CMTX2-3), enfermedad de Pelizaeus-Merzbacher, SMAX2), trastornos de la piel y tejidos relacionados (por ejemplo, disqueratosis congénita, displasia ectodérmica hipohidrótica (DEH), ictiosis ligada al cromosoma X, distrofia corneal endotelial ligada al cromosoma X), trastornos neuromusculares (por ejemplo, distrofia muscular de Becker/Duchenne, miopatía centronuclear (MTM1), síndrome de Conradi-Hunermann, distrofia muscular de Emery-Dreifuss 1), trastornos urológicos (por ejemplo, síndrome de Alport, enfermedad de Dent, diabetes insípida nefrógena ligada al cromosoma X), trastornos óseos/dentales (por ejemplo, AMELX amelogénesis imperfecta), y otros trastornos (por ejemplo, síndrome de Barth, síndrome de McLeod, síndrome de Smith-Fineman-Myers, síndrome de Simpson-Golabi-Behmel, síndrome de Mohr-Tranebjaerg, síndrome nasodigitoacústico). Los ejemplos de trastornos dominantes ligados al cromosoma X incluyen, sin limitación, hipofosfatemia ligada al cromosoma X, hipoplasia dérmica focal, síndrome del cromosoma X frágil, síndrome de Aicardi, incontinencia pigmentaria, síndrome de Rett, síndrome CHILD, síndrome de Lujan-Fryns y síndrome bucofaciodigital 1. Los ejemplos de trastornos ligados al cromosoma Y incluyen, sin limitación, esterilidad masculina, retinitis pigmentosa y azoospermia.

Anomalías cromosómicas

En algunas implementaciones, la presencia o ausencia de una anomalía cromosómica fetal puede determinarse usando un método o aparato descrito en el presente documento. Las anomalías cromosómicas incluyen, sin limitación, una ganancia o pérdida de un cromosoma completo o una región de un cromosoma que comprende uno o más genes. Las anomalías cromosómicas incluyen monosomías, trisomías, polisomías, pérdida de heterocigosidad, deleciones y/o duplicaciones de una o más secuencias de nucleótidos (por ejemplo, uno o más genes), incluyendo deleciones y duplicaciones provocadas por translocaciones desequilibradas. Los términos "aneuploidía" y "aneuploide", tal como se usan en el presente documento, se refieren a un número anómalo de cromosomas en células de un organismo. Dado que los diferentes organismos tienen complementos cromosómicos ampliamente variables, el término "aneuploidía" no se refiere a un número particular de cromosomas, sino más bien a la situación en la que el contenido cromosómico dentro de una célula o células dadas de un organismo es anómalo. En algunas implementaciones, el término "aneuploidía" en el presente documento se refiere a un desequilibrio de material genético provocado por una pérdida o ganancia de un cromosoma completo o parte de un cromosoma. Un "aneuploidía" puede referirse a una o más deleciones y/o inserciones de un segmento de un cromosoma.

El término "monosomía", tal como se usa en el presente documento, se refiere a la falta de un cromosoma del complemento normal. La monosomía parcial puede producirse en translocaciones o deleciones desequilibradas, en las cuales solo un segmento del cromosoma está presente en una sola copia. La monosomía de cromosomas sexuales (45, X) provoca el síndrome de Turner, por ejemplo.

El término "disomía" se refiere a la presencia de dos copias de un cromosoma. Para organismos tales como seres humanos que tienen dos copias de cada cromosoma (aquellos que son diploides o "euploides"), la disomía es la condición normal. Para organismos que normalmente tienen tres o más copias de cada cromosoma (aquellos que son triploides o superiores), la disomía es un estado cromosómico aneuploide. En la disomía uniparental, ambas copias de un cromosoma provienen del mismo progenitor (sin contribución del otro progenitor).

El término "euploide", en algunas implementaciones, se refiere a un complemento normal de cromosomas.

El término "trisomía", tal como se usa en el presente documento, se refiere a la presencia de tres copias, en lugar de dos copias, de un cromosoma particular. La presencia de un cromosoma 21 extra, que se encuentra en el síndrome de Down humano, se denomina "trisomía 21". La trisomía 18 y la trisomía 13 son otras dos trisomías autosómicas humanas. La trisomía de cromosomas sexuales puede observarse en mujeres (por ejemplo, 47, XXX en el síndrome triple X) u hombres (por ejemplo, 47, XXY en el síndrome de Klinefelter; o 47, XYY en el síndrome de Jacobs).

Los términos "tetrasomía" y "pentasomía", tal como se usan en el presente documento, se refieren a la presencia de cuatro o cinco copias de un cromosoma, respectivamente. Aunque rara vez se ven con autosomas, se ha notificado tetrasomía y pentasomía en cromosomas sexuales en seres humanos, incluyendo XXXX, XXXY, XXYY, XYYY, XXXXX, XXXXY, XXXYY, XXYYY y XYYYY.

Las anomalías cromosómicas pueden estar provocadas por una variedad de mecanismos. Los mecanismos incluyen, pero no se limitan a, (i) no disyunción que se produce como resultado de un punto de control mitótico debilitado, (ii) puntos de control mitóticos inactivos que provocan ausencia de disyunción en múltiples cromosomas, (iii) unión merotélica que se produce cuando un cinetocoro se une a ambos polos del huso mitótico, (iv) una formación de huso multipolar cuando se forman más de dos polos de huso, (v) una formación de huso monopolar cuando se forma solamente un único polo de huso, y (vi) un producto intermedio tetraploide que se produce como resultado final del mecanismo de huso monopolar.

Las expresiones "monosomía parcial" y "trisomía parcial", tal como se usan en el presente documento, se refieren a un desequilibrio de material genético provocado por la pérdida o ganancia de parte de un cromosoma. Una monosomía parcial o trisomía parcial puede resultar de una translocación desequilibrada, en la que un individuo porta un cromosoma derivado formado a través de la rotura y fusión de dos cromosomas diferentes. En esta situación, el individuo tendría tres copias de parte de un cromosoma (dos copias normales y el segmento que existe en el cromosoma derivado) y solo una copia de parte del otro cromosoma implicado en el cromosoma derivado.

El término "mosaicismo", tal como se usa en el presente documento, se refiere a aneuploidía en algunas células, pero no en todas las células, de un organismo. Determinadas anomalías cromosómicas pueden existir como anomalías cromosómicas con mosaicismo y sin mosaicismo. Por ejemplo, determinados individuos con trisomía 21 tienen síndrome de Down y algunos tienen síndrome de Down sin mosaicismo. Diferentes mecanismos pueden conducir al mosaicismo. Por ejemplo, (i) un cigoto inicial puede tener tres cromosomas 21, lo que normalmente daría como resultado trisomía 21 simple, pero durante el transcurso de la división celular una o más líneas celulares perdieron uno de los cromosomas 21; y (ii) un cigoto inicial puede tener dos cromosomas 21, pero durante el transcurso de la división celular se duplicó uno de los cromosomas 21. El mosaicismo somático se produce probablemente a través de mecanismos distintos de aquellos normalmente asociados con síndromes genéticos que involucran aneuploidía completa o con mosaicismo. El mosaicismo somático se ha identificado en determinados tipos de cánceres y en neuronas, por ejemplo. En determinados casos, la trisomía 12 se ha identificado en la leucemia linfocítica crónica (LLC) y la trisomía 8 se ha identificado en la leucemia mieloide aguda (LMA). Además, los síndromes genéticos en los que un individuo está predispuesto a rotura de cromosomas (síndromes de

inestabilidad cromosómica) se asocian frecuentemente con un mayor riesgo de diversos tipos de cáncer, destacando así el papel de la aneuploidía somática en la carcinogénesis. Los métodos y protocolos descritos en el presente documento pueden identificar la presencia o ausencia de anomalías cromosómicas sin mosaicismo y con mosaicismo.

Las tablas 1A y 1B presentan una lista no limitativa de afecciones, síndromes y/o anomalías cromosómicos que pueden identificarse potencialmente mediante los métodos y aparatos descritos en el presente documento. La tabla 1B proviene de la base de datos DECIPHER a 6 de octubre de 2011 (por ejemplo, versión 5.1, basándose en las posiciones mapeadas en GRCh37; disponible en el localizador uniforme de recursos (URL) dechipher.sanger.ac.uk).

Tabla 1A

Cromosoma	Anomalía	Asociación con la enfermedad
X	XO	Síndrome de Turner
Y	XXY	Síndrome de Klinefelter
Y	XYY	Síndrome doble Y
Y	XXX	Síndrome de trisomía X
Y	XXXX	Síndrome tetra-X
Y	Deleción de Xp21	Síndrome de Duchenne/Becker, hipoplasia suprarrenal congénita, enfermedad granulomatosa crónica
Y	Deleción de Xp22	deficiencia de esteroide sulfatasa
Y	Deleción de Xq26	Enfermedad linfproliferativa ligada al cromosoma X
1	monosomía-trisomía (somática) 1 p	neuroblastoma
2	monosomía-trisomía 2q	retraso del crecimiento, retraso del desarrollo y mental, y anomalías físicas menores
3	monosomía-trisomía (somática)	Linfoma no hodgkiniano
4	monosomía-trisomía (somática)	Leucemia no linfocítica aguda (LNLA)
5	5p	síndrome del maullido de gato; síndrome de Lejeune
5	monosomía-trisomía (somática) 5q	síndrome mielodisplásico
6	monosomía-trisomía (somática)	sarcoma de células claras
7	deleción de 7q11.23	Síndrome de William
7	monosomía-trisomía	síndrome de monosomía de la infancia 7; somático: adenomas renales corticales; síndrome mielodisplásico
8	deleción de 8q24.1	síndrome de Langer-Giedon
8	monosomía-trisomía	síndrome mielodisplásico; síndrome de Warkany; somático: leucemia mielógena crónica
9	monosomía 9p	Síndrome de Alfi
9	monosomía-trisomía parcial 9p	Síndrome de Rethore

Cromosoma	Anomalía	Asociación con la enfermedad
9	trisomía	síndrome de trisomía 9 completa; síndrome de trisomía 9 con mosaicismo
10	Monosomía-trisomía (somática)	LLA o LNLA
11	11 p-	Aniridia; tumor de Wilms
11	11q-	Síndrome de Jacobson
11	monosomía trisomía (somática)	linajes mieloides afectados (LNLA, SMD)
12	monosomía-trisomía (somática)	LLC, tumor de células de la granulosa juvenil (TCGJ)
13	13q-	Síndrome 13q; síndrome de Orbeli
13	deleción de 13q14	retinoblastoma
13	monosomía-trisomía	Síndrome de Patau
14	monosomía-trisomía (somática)	trastornos mieloides (SMD, LNLA, LMC atípica)
15	deleción de 15q11-q13 monosomía	Prader-Willi, síndrome de Angelman
15	trisomía (somática)	linajes mieloides y linfoides afectados, por ejemplo, SMD, LNLA, LLA, LLC)
16	deleción de 16q13.3	Rubenstein-Taybi
3	monosomía-trisomía (somática)	carcinomas de células renales papilares (malignos)
17	17p-(somático)	Síndrome p17 en neoplasias malignas mieloides
17	deleción de 17q11.2	Smith-Magenis

17	17q13.3	Miller-Dieker
17	monosomía-trisomía (somática)	adenomas corticales renales
17	trisomía 17p11.2-12	Síndrome de Charcot-Marie-Tooth tipo 1; NHPP
18	18p-	Síndrome de monosomía parcial 18p o síndrome de Grouchy-Lamy-Thieffry
18	18q-	Síndrome de Grouchy-Lamy-Salmon-Landry
18	monosomía-trisomía	Síndrome de Edwards

Cromosoma	**Anomalía**	**Asociación con la enfermedad**
19	monosomía-trisomía	
20	20p-	síndrome de trisomía 20p
20	deleción de 20p11.2-12	Alagille
20	20q-	somático: SMD, LNLA, policitemia vera, leucemia neutrofílica crónica
20	monosomía-trisomía (somática)	carcinomas de células renales papilares (malignos)
21	monosomía-trisomía	Síndrome de Down
22	deleción de 22q11.2	Síndrome de DiGeorge, síndrome velocardiofacial, síndrome de anomalías conotruncales y de la cara, síndrome de Opitz G/BBB autosómico dominante, síndrome cardiofacial de Caylor
22	monosomía-trisomía	síndrome de trisomía 22 completa

Tabla 1B

Síndrome	**Cromosoma**	**Inicio**	**Fin**	**Intervalo (Mb)**	**Grado**
Síndrome de microdeleción de 12q14	12	65.071.919	68.645.525	3,57	
15q13.3 síndrome de microdeleción	15	30.769.995	32.701.482	1,93	
síndrome de microdeleción recurrente de 15q24	15	74.377.174	76.162.277	1,79	
síndrome de sobrecrecimiento de 15q26	15	99.357.970	102.521.392	3,16	
síndrome de microduplicación de 16p11.2	16	29.501.198	30.202.572	0,70	
síndrome de microdeleción de 16p11.2-p12.2	16	21.613.956	29.042.192	7,43	
microdeleción recurrente de 16p13.11 (locus de susceptibilidad a trastornos neurocognitivos)	16	15.504.454	16.284.248	0,78	

Síndrome	**Cromosoma**	**Inicio**	**Fin**	**Intervalo (Mb)**	**Grado**
microduplicación recurrente de 16p13.11 (locus de susceptibilidad a trastornos neurocognitivos)	16	15.504.454	16.284.248	0,78	
síndrome de microdeleción recurrente de 17q21.3	17	43.632.466	44.210.205	0,58	1
Síndrome de microdeleción de 1p36	1	10.001	5.408.761	5,40	1
microdeleción recurrente de 1q21.1 (locus de susceptibilidad a trastornos del desarrollo neurológico)	1	146.512.930	147.737.500	1,22	3
microduplicación recurrente de 1q21.1 (posible locus de susceptibilidad a trastornos del desarrollo neurológico)	1	146.512.930	147.737.500	1,22	3

1q21.1 locus de susceptibilidad trombocitopenia-síndrome de radio ausente (TAR)a	1	145.401.253	145.928.123	0,53	3
síndrome de deleción de 22q11 (síndrome velocardiofacial / DiGeorge)	22	18.546.349	22.336.469	3,79	1
síndrome de duplicación de 22q11	22	18.546.349	22.336.469	3,79	3
síndrome de deleción distal de 22q11.2	22	22.115.848	23.696.229	1,58	
síndrome de deleción de 22q13 (síndrome de Phelan-Mcdermid)	22	51.045.516	51.187.844	0,14	1
síndrome de microdeleción de 2p15-16.1	2	57.741.796	61.738.334	4,00	
síndrome de deleción de 2q33.1	2	196.925.089	205.206.940	8,28	1
monosomía de 2q37	2	239.954.693	243.102.476	3,15	1

Síndrome	**Cromosoma**	**Inicio**	**Fin**	**Intervalo (Mb)**	**Grado**
síndrome de microdeleción de 3q29	3	195.672.229	197.497.869	1,83	
síndrome de microduplicación de 3q29	3	195.672.229	197.497.869	1,83	
síndrome de duplicación de 7q11.23	7	72.332.743	74.616.901	2,28	
síndrome de deleción de 8p23.1	8	8.119.295	11.765.719	3,65	
síndrome de deleción subtelomérica de 9q	9	140.403.363	141.153.431	0,75	1
Leucodistrofia autosómica dominante de inicio en la edad adulta (ADLD)	5	126.063.045	126.204.952	0,14	
Síndrome de Angelman (tipo 1)	15	22.876.632	28.557.186	5,68	1
Síndrome de Angelman (tipo 2)	15	23.758.390	28.557.186	4,80	1
Síndrome ATR-16	16	60.001	834.372	0,77	1
AZFa	Y	14.352.761	15.154.862	0,80	
AZFb	Y	20.118.045	26.065.197	5,95	
AZFb+AZFc	Y	19.964.826	27.793.830	7,83	
AZFc	Y	24.977.425	28.033.929	3,06	
Síndrome del ojo de gato (tipo I)	22	1	16.971.860	16,97	
Síndrome de Charcot-Marie-Tooth tipo 1A (CMT1A)	17	13.968.607	15.434.038	1,47	1
Síndrome de maullido de gato (deleción de 5p)	5	10.001	11.723.854	11,71	1
Enfermedad de Alzheimer de inicio temprano con angiopatía amiloide cerebral	21	27.037.956	27.548.479	0,51	
Adenocarcinoma familiar Poliposis	5	112.101.596	112.221.377	0,12	
Hereditaria con propensión a las parálisis por presión (NHPP)	17	13.968.607	15.434.038	1,47	1
Discondrosteosis de Leri-Weill (LWD) - deleción de SHOX	X	751.878	867.875	0,12	

Síndrome	Cromosoma	Inicio	Fin	Intervalo (Mb)	Grado
Discondrosteosis de Leri-Weill (LWD) - deleción de SHOX	X	460.558	753.877	0,29	
Síndrome de Miller-Dieker (SMD)	17	1	2.545.429	2,55	1
Síndrome de microdeleción de NF1	17	29.162.822	30.218.667	1,06	1
Enfermedad de Pelizaeus-Merzbacher	X	102.642.051	103.131.767	0,49	
Síndrome de Potocki-Lupski (síndrome de duplicación de 17p11.2)	17	16.706.021	20.482.061	3,78	
Síndrome de Potocki-Shaffer	11	43.985.277	46.064.560	2,08	1
Síndrome de Prader-Willi (tipo 1)	15	22.876.632	28.557.186	5,68	1
Síndrome de Prader-Willi (tipo 2)	15	23.758.390	28.557.186	4,80	1
QRD (quistes renales y diabetes)	17	34.907.366	36.076.803	1,17	
Síndrome de Rubinstein-Taybi	16	3.781.464	3.861.246	0,08	1
Síndrome de Smith-Magenis	17	16.706.021	20.482.061	3,78	1
Síndrome de Sotos	5	175.130.402	177.456.545	2,33	1
Malformación de manos/pies divididos 1 (SHFM1)	7	95.533.860	96.779.486	1,25	
Deficiencia de esteroide sulfatasa (STS)	X	6.441.957	8.167.697	1,73	
Síndrome de deleción de 11p13 WAGR	11	31.803.509	32.510.988	0,71	
Síndrome de Williams-Beuren (WBS)	7	72.332.743	74.616.901	2,28	1
Síndrome de Wolf-Hirschhorn	4	10.001	2.073.670	2,06	1
Duplicación de Xq28 (MECP2)	X	152.749.900	153.390.999	0,64	

Las afecciones de grado 1 tienen a menudo una o más de las siguientes características: anomalía patógena; fuerte consenso entre los genetistas; altamente penetrante; todavía puede tener un fenotipo variable, pero algunas características comunes; todos los casos en la bibliografía tienen un fenotipo clínico; ningún caso de individuos sanos con la anomalía; no se notifica en bases de datos de DVG ni se encuentra en poblaciones sanas; datos funcionales que confirman el efecto de dosificación de un solo gen o de múltiples genes; genes candidatos confirmados o fuertes; implicaciones de manejo clínico definidas; riesgo conocido de cáncer con implicación para la vigilancia; múltiples fuentes de información (OMIM, GeneReviews, Orphanet, Unique, Wikipedia); y/o disponibles para uso diagnóstico (asesoramiento sobre reproducción).

Las afecciones de grado 2 tienen a menudo una o más de las siguientes características: probablemente anomalía patógena; altamente penetrante; fenotipo variable sin características sistemáticas distintas de DD; pequeña cantidad de casos/ informes en la bibliografía; todos los casos notificados tienen un fenotipo clínico; sin datos funcionales o genes patógenos confirmados; múltiples fuentes de información (OMIM, Genereviews, Orphanet, Unique, Wikipedia); y/o pueden usarse con fines diagnósticos y asesoramiento sobre reproducción.

Las afecciones grado 3 tienen a menudo una o más de las siguientes características: locus de susceptibilidad; individuos sanos o progenitores no afectados de un probando descrito; presentes en poblaciones de control; no penetrantes; fenotipo leve e inespecífico; características menos sistemáticas; sin datos funcionales o genes patógenos confirmados; fuentes de datos más limitadas; la posibilidad de un segundo diagnóstico sigue siendo una posibilidad para casos que se desvían de la mayoría o si hay nuevos hallazgos clínicos presentes; y/o precaución cuando se usa para fines de diagnóstico y consejos cautos para asesoramiento sobre reproducción.

Preeclampsia

En algunas implementaciones, la presencia o ausencia de preeclampsia se determina usando un método o aparato descrito en el presente documento. La preeclampsia es una afección en la que surge hipertensión durante el embarazo (es decir, hipertensión inducida por el embarazo) y se asocia con cantidades significativas de proteína en la orina. En algunos casos, la preeclampsia se asocia además con niveles elevados de ácido nucleico extracelular y/o alteraciones en los patrones de metilación. Por ejemplo, se ha observado una correlación positiva entre los niveles de RASSF1A hipermetilado derivado del feto extracelular y la gravedad de la preeclampsia. En determinados ejemplos, se observa una mayor metilación del ADN para el gen H19 en la placenta con preeclampsia en comparación con los controles normales.

La preeclampsia es una de las causas principales de morbimortalidad materna y fetal/neonatal en todo el mundo. Los ácidos nucleicos circulantes, libres de células en plasma y suero son biomarcadores novedosos con aplicaciones clínicas prometedoras en diferentes campos médicos, incluyendo el diagnóstico prenatal. Se han notificados cambios cuantitativos del ADN fetal libre de células (flc) en plasma materno como indicador para la preeclampsia inminente en diferentes estudios, por ejemplo, usando PCR cuantitativa en tiempo real para los loci SRA o DYS 14 específicos del sexo masculino. En los casos de preeclampsia de inicio temprano, pueden observarse niveles elevados en el primer trimestre. El aumento de los niveles de ADNflc antes del inicio de los síntomas puede deberse a hipoxia/reoxigenación dentro del espacio intervelloso que conduce al estrés oxidativo tisular y al aumento de la apoptosis y necrosis placentaria. Además de la evidencia de mayor descarga de ADNflc en la circulación materna, también existe evidencia de menor aclaramiento renal de ADNflc en la preeclampsia. Dado que la cantidad de ADN fetal se determina actualmente mediante la cuantificación de secuencias específicas del cromosoma Y, los enfoques alternativos, tales como la medición del ADN libre de células total o el uso de marcadores epigenéticos fetales independientes del sexo, tales como la metilación del ADN, ofrecen una alternativa. EL ARN libre de células de origen placentario es otro biomarcador alternativo que puede usarse para analizar y diagnosticar preeclampsia en la práctica clínica. EL ARN fetal se asocia con partículas placentarias subcelulares que lo protegen frente a la degradación. Los niveles de ARN fetal a veces son diez veces mayores en mujeres embarazadas con preeclampsia en comparación con los controles y, por tanto, es un biomarcador alternativo que puede usarse para analizar y diagnosticar preeclampsia en la práctica clínica.

Patógenos

En algunas implementaciones, la presencia o ausencia de una afección patógena se determina mediante un método o aparato descrito en el presente documento. Una afección patógena puede estar provocada por la infección de un huésped por un patógeno incluyendo, pero sin limitarse a, una bacteria, un virus u hongo. Dado que los patógenos poseen normalmente ácido nucleico (por ejemplo, ADN genómico, ARN genómico, ARNm) que puede distinguirse del ácido nucleico del huésped, los métodos y aparatos proporcionados en el presente documento pueden usarse para determinar la presencia o ausencia de un patógeno. A menudo, los patógenos poseen ácido nucleico con características únicas para un patógeno particular, tal como, por ejemplo, estado epigenético y/o una o más variaciones, duplicaciones y/o deleciones de secuencias. Por tanto, los métodos proporcionados en el presente documento pueden usarse para identificar un patógeno particular o variante de patógenos (por ejemplo, cepa).

Cánceres

En algunas implementaciones, la presencia o ausencia de un trastorno de proliferación celular (por ejemplo, un cáncer) se determina usando un método o aparato descrito en el presente documento. Por ejemplo, los niveles de ácido nucleico libre de células en suero pueden elevarse en pacientes con diversos tipos de cáncer en comparación con pacientes sanos. Los pacientes con enfermedades metastásicas, por ejemplo, algunas veces pueden tener niveles de ADN en suero aproximadamente dos veces más altos que los pacientes no metastásicos. Los pacientes con enfermedades metastásicas pueden identificarse además por marcadores específicos de cáncer y/o determinados polimorfismos de un solo nucleótido o repeticiones cortas en tándem, por ejemplo. Los ejemplos no limitativos de tipos de cáncer que pueden correlacionarse positivamente con niveles elevados de ADN circulante incluyen cáncer de mama, cáncer colorrectal, cáncer gastrointestinal, cáncer hepatocelular, cáncer de pulmón, melanoma, linfoma no hodgkiniano, leucemia, mieloma múltiple, cáncer de vejiga, hepatoma, cáncer cervicouterino, cáncer de esófago, cáncer de páncreas y cáncer de próstata. Diversos cánceres pueden tener, y algunas veces pueden liberarse en el torrente sanguíneo, ácidos nucleicos con características que son distinguibles de los ácidos nucleicos de las células sanas no cancerosas, tales como, por ejemplo, el estado epigenético y/o variaciones, duplicaciones y/o deleciones de secuencia. Tales características pueden ser, por ejemplo, específicas para un tipo particular de cáncer. Por tanto, se contempla además que un método proporcionado en el presente documento puede usarse para identificar un tipo particular de cáncer.

Ejemplos

Los ejemplos expuestos a continuación ilustran determinadas implementaciones y no limitan la tecnología.

Ejemplo 1: Métodos generales para detectar afecciones asociadas con variaciones genéticas.

Los métodos y la teoría subyacente descritos en el presente documento pueden usarse para detectar diversas afecciones asociadas con la variación genética y determinar la presencia o ausencia de una variación genética. Los ejemplos no limitativos de variaciones genéticas que pueden detectarse con los métodos descritos en el presente documento incluyen aberraciones cromosómicas segmentarias (por ejemplo, deleciones, duplicaciones), aneuploidía, sexo, identificación de muestras, afecciones patológicas asociadas con variación genética, similares o combinaciones de lo anterior.

Filtrado de bins

El contenido de información de una región genómica en un cromosoma diana puede visualizarse al representar gráficamente el resultado de la separación promedio entre recuentos de euploide y trisomía normalizados por incertidumbres combinadas, en función de la posición en el cromosoma. Mayor incertidumbre (véase la Fig. 1) o brecha reducida entre triploides y euploides (por ejemplo, embarazos triploides y embarazos euploides)(véase la Fig. 2) ambos dan como resultado valores de Z disminuidos para los casos afectados, a veces reduciendo la potencia predictiva de las puntuaciones Z.

La Fig. 3 ilustra gráficamente un perfil de valor de p, basándose en la distribución t, representado gráficamente en función de la posición en el cromosoma a lo largo del cromosoma 21. El análisis de los datos presentados en la Fig. 3 identifica 36 bins del cromosoma 21 no informativos, cada uno de aproximadamente 50 kilopares de bases (kpb) de longitud. La región no informativa está ubicada en el brazo p, cerca del centrómero (21 p11.2-21 p11.1). La eliminación de los 36 bins del cálculo de las puntuaciones Z, tal como se describe esquemáticamente en la Fig. 4, algunas veces puede aumentar significativamente los valores de Z para todos los casos de trisomía, mientras que introduce solo variaciones aleatorias en los valores de Z euploides.

La mejora en la potencia predictiva proporcionada por la eliminación de los 36 bins no informativos puede explicarse al examinar el perfil de recuento para el cromosoma 21 (véase la Fig. 5). En la Fig. 5, dos muestras seleccionadas arbitrariamente demuestran la tendencia general de los perfiles de recuento frente a (vs) de bins a seguir tendencias prácticamente similares, aparte del ruido de corto alcance. Los perfiles mostrados en la Fig. 5 son sustancialmente paralelos. La región resaltada del gráfico de perfiles presentado en la Fig. 5 (por ejemplo, la región en la elipse), aunque todavía presenta paralelismo, presenta además grandes fluctuaciones en relación con el resto del cromosoma. La eliminación de los bins fluctuantes (por ejemplo, los 36 bins no informativos) puede mejorar la precisión y sistematicidad de las estadísticas Z, en algunas implementaciones.

Normalización de bins

Filtrar los bins no informativos, tal como se describe en el ejemplo 1, a veces no proporciona la mejora deseada en cuanto a la potencia predictiva de los valores de Z. Cuando los datos del cromosoma 18 se filtran para eliminar los bins no informativos, tal como se describe en el ejemplo 1, los valores z no mejoraron sustancialmente (véase la Fig. 6). Tal como se observa con los perfiles de recuento del cromosoma 21 presentados en el ejemplo 1, los perfiles de recuento del cromosoma 18 son además sustancialmente paralelos, sin considerar ruido de corto alcance. Sin embargo, dos muestras del cromosoma 18 usadas para evaluar las incertidumbres de recuento de bins (véase la parte inferior de la Fig. 6) se desvía significativamente del paralelismo general de los perfiles de recuento. Las depresiones en el centro de las dos trazas, resaltadas por la elipse, representan deleciones grandes. Otras muestras examinadas durante el transcurso del experimento no presentaron esta deleción. La deleción coincide con la ubicación de una depresión en perfiles de valor de p para el cromosoma 18, ilustrado en por la elipse que se muestra en la Fig. 7. Es decir, la depresión observada en los perfiles de valor de p para el cromosoma 18 se explican por la presencia de la deleción en las muestras del cromosoma 18, que provoca un aumento en la varianza de ectada. La varianza en recuentos no es aleatoria, pero representa un evento poco común (por ejemplo, la deleción de un segmento del cromosoma 18), que, si se incluye con otras fluctuaciones aleatorias de otras muestras, disminuye el procedimiento de filtrado de bins de potencia predictiva.

Surgen dos cuestiones a partir de este ejemplo; (1) cómo se determina que las señales de valor de p son significativas y/o útiles, y (2) el enfoque de valor de p descrito en el presente documento puede generalizarse para su uso con cualquier dato de bins (por ejemplo, de dentro de cualquier cromosoma, no solo bins de dentro de los cromosomas 13, 18 o 21). Un procedimiento generalizado podría usarse para eliminar la variabilidad en los recuentos totales para todo el genoma, que puede usarse a menudo como el factor de normalización cuando se evalúan las puntuaciones Z. Los datos presentados en la Fig. 8 pueden usarse para investigar las respuestas a las cuestiones anteriores al reconstruir el contorno general de los datos al asignar la mediana del recuento de referencia a cada bin y normalizar cada recuento de bins en la muestra de prueba con respecto a la mediana del recuento de referencia asignada.

Las medianas se extraen de un conjunto de referencias euploides conocidas. Antes de calcular la mediana de recuentos de referencia, se filtran los bins no informativos en todo el genoma. Los recuentos de bins restantes se normalizan con respecto al número residual total de recuentos. La muestra de prueba se normaliza además con respecto a la suma de los recuentos observados para los bins que no se filtran. El perfil de prueba resultante a menudo se centra alrededor de un valor de 1, excepto en áreas de deleciones o duplicación maternas, y áreas en las que el feto es triploide (véase la Fig. 9). El perfil normalizado por bins ilustrado en la Fig. 10 confirma la validez del procedimiento de normalización y revela

claramente la deleción materna heterocigota (por ejemplo, depresión central en el segmento de color gris del trazado del perfil) en el cromosoma 18 y la representación cromosómica elevada del cromosoma 18 de la muestra sometida a prueba (véase el área de color gris del trazado del perfil en la Fig. 10). Tal como puede observarse en la Fig. 10, la mediana del valor para el segmento de color gris del trazado se centra alrededor de aproximadamente 1,1, en el que la mediana del valor para el segmento de color negro del trazado se centra alrededor de 1,0.

Elevación de pico

La Fig. 11 ilustra gráficamente los resultados del análisis de múltiples muestras usando normalización por bins, de un paciente con una característica o rasgo discernible (por ejemplo, duplicación materna, deleción materna, similares o combinaciones de los mismos). A menudo, las identidades de las muestras pueden determinarse mediante la comparación de sus perfiles de recuento normalizados respectivos. En el ejemplo ilustrado en la Fig. 11, la ubicación de la depresión en el perfil normalizado y su elevación, así como su rareza, indican que ambas muestras se originan del mismo paciente. Los datos del panel forense pueden usarse a menudo para confirmar estos hallazgos.

Las Fig. 12 y 13 ilustran gráficamente los resultados del uso de perfiles de bins normalizados para identificar la identidad del paciente, o la identidad de la muestra. Las muestras analizadas en las Fig. 12 y 13 portan aberraciones maternas amplias en los cromosomas 4 y 22, que están ausentes en las otras muestras en los trazados de perfil, lo que confirma el origen compartido de las trazas superior e inferior. Resultados tales como esto pueden conducir a la determinación de que una muestra particular pertenece a un paciente específico, y puede usarse además para determinar si una muestra particular ya se ha analizado.

La normalización por bins facilita la detección de aberraciones; sin embargo, la comparación de picos de diferentes muestras se facilita, a menudo además mediante el análisis de medidas cuantitativas de las elevaciones y ubicaciones de los picos (por ejemplo, bordes de pico). El descriptor más destacado de un pico es a menudo su elevación, seguido por las ubicaciones de sus bordes. Las características de perfiles de recuento diferentes pueden compararse a menudo con el uso del siguiente análisis no limitativo.

(a) Se determina la confianza en unas características de los picos detectados en una sola muestra de prueba. Si la característica puede distinguirse del ruido de fondo o de los artefactos de procesamiento, la característica puede analizarse adicionalmente frente a la población general.

(b) Determinar la prevalencia de la característica detectada en la población general. Si la característica es rara, puede usarse como marcador para aberraciones raras. Las características que se encuentran frecuentemente en la población general son menos útiles para el análisis. Los orígenes étnicos pueden desempeñar un papel en la determinación de la relevancia de una elevación de pico de las características detectadas. Por tanto, algunas características proporcionan información útil para muestras de determinados orígenes étnicos.

(c) Derivar la confianza en la comparación entre características observadas en diferentes muestras.

Se ilustran en la Fig. 14 los recuentos de bins normalizados en el cromosoma 5, de un sujeto euploide. La elevación promedio es generalmente la línea base de referencia a partir de la cual se miden las elevaciones de aberraciones, en algunas implementaciones. Las desviaciones pequeñas y/o estrechas son predictores menos fiables que las aberraciones amplias y pronunciadas. Por tanto, el ruido de fondo o la varianza de baja contribución fetal y/o artefactos de procesamiento son una consideración importante cuando las aberraciones no son grandes o no tienen una elevación de pico significativa por encima del fondo. Un ejemplo de esto se presenta en la Fig. 15, en el que un pico que sería significativo en la traza superior, puede enmascararse en el ruido de fondo observado en la traza del perfil inferior. La confianza en la elevación de pico (véase la Fig. 16) puede determinarse por la desviación promedio de la referencia (mostrada como el símbolo delta), con relación a la anchura de la distribución euploide (por ejemplo, combinada con la varianza (mostrada como el símbolo sigma) en la desviación promedio). El error en la elevación de tramo promedio puede derivarse de la fórmula conocida para el error de la media. Si una extensión más larga de un bin se trata como una muestra aleatoria (no contigua) de todos los bins dentro de un cromosoma, el error en la elevación promedio disminuye con la raíz cuadrada del número de bins dentro de la aberración. Este razonamiento ignora la correlación entre bins vecinos, una suposición confirmada por la función de correlación que se muestra en la Fig. 17 (por ejemplo, la ecuación para G(n)). Los perfiles no normalizados a veces presentan fuertes correlaciones de rango medio (por ejemplo, la variación similar a una onda de la línea base); sin embargo, los perfiles normalizados suavizan la correlación, dejando solamente ruido aleatorio. La estrecha coincidencia entre el error estándar de la media, la corrección para autocorrelación y las estimaciones de muestra reales de la desviación estándar de la elevación media en el cromosoma 5 (véase la Fig. 18) confirma la validez de la falta de correlación supuesta. Pueden evaluarse entonces las puntuaciones Z (véase la Fig. 19) y valores de p calculados a partir de las puntuaciones Z asociadas con desviaciones de la elevación esperada de 1 (véase la Fig. 20) a la luz de la estimación de incertidumbre en la elevación promedio. Los valores de p se basan en una distribución t cuyo orden está determinado por el número de bins en un pico. Dependiendo del nivel de confianza deseado, un punto de corte puede suprimir el ruido y permitir la detección inequívoca de la señal real.

$$Z = \frac{\Delta_1 - \Delta_2}{\sqrt{\sigma_1^2\left(\frac{1}{N_1}+\frac{1}{n_1}\right)+\sigma_2^2\left(\frac{1}{N_2}+\frac{1}{n_2}\right)}} \quad (1)$$

La ecuación 1 puede usarse para comparar directamente la elevación de pico de dos muestras diferentes, en la que N y n se refieren a los números de bins en todo el cromosoma y dentro de la aberración, respectivamente. El orden de la prueba de la t que producirá un valor de p que mide la similitud entre dos muestras se determina por el número de bins en el más corto de los dos tramos desviados.

Borde de pico

Además de comparar las elevaciones promedio de aberraciones en una muestra, el comienzo y el final de los tramos comparados pueden proporcionar además información útil para el análisis estadístico. El límite superior de resolución para comparaciones de los bordes de pico se determina a menudo por el tamaño de bins (por ejemplo, 50 kpbs en los ejemplos descritos en el presente documento). La Fig. 21 ilustra 3 escenarios posibles de bordes de pico; (a) un pico de una muestra puede estar completamente contenido dentro del pico coincidente de otra muestra, (b) los bordes de una muestra pueden solaparse parcialmente con los bordes de otra muestra, o (c) el borde anterior de una muestra puede tocar o solaparse solo ligeramente con el borde posterior de otra muestra.

La Fig. 22 ilustra y ejemplo del escenario descrito en (c) (por ejemplo, véase la traza de color gris claro intermedia, en la que el borde posterior de la traza intermedia toca ligeramentre el borde anterior de la traza superior).

La tolerancia lateral asociada con un borde puede usarse a menudo para distinguir variaciones aleatorias de los bordes de aberración verdaderos. La posición y la anchura de un borde pueden cuantificarse mediante la evaluación numérica de la primera derivada del perfil de recuento aberrante, tal como se muestra en la Fig. 23. Si la aberración se representa como una función compuesta de dos funciones de Heaviside, su derivada será la suma de dos funciones delta de Dirac. El borde inicial corresponde a un pico en forma de absorción ascendente, mientras que el borde final es un pico de absorción descendente desplazado 180 grados. Si la aberración es estrecha, los dos aumentos bruscos están cerca entre sí, lo que forma un contorno similar a una dispersión. Las ubicaciones de los bordes pueden aproximarse por los extremos de los primeros aumentos bruscos de derivada, mientras que la tolerancia de borde está determinada por sus anchuras.

La comparación entre muestras diferentes puede reducirse a menudo para determinar la diferencia entre dos ubicaciones de borde coincidentes, dividida entre las incertidumbres de borde combinadas. Sin embargo, algunas veces las derivadas se pierden en el ruido de fondo, tal como se ilustra en la Fig. 24. Aunque la aberración propiamente dicha se beneficia de la información colectiva contribuida por todos sus bins, la primera derivada solo puede proporcionar información de los pocos puntos en el borde de la aberración, lo que puede ser insuficiente para superar el ruido. El promediado de ventana deslizante, usado para crear la Fig. 24, es de valor limitado en esta situación. El ruido puede eliminarse combinando la primera derivada (por ejemplo, similar a una estimación puntual) con la elevación de pico (por ejemplo, comparable a una estimación integral). En algunas implementaciones la primera derivada y la elevación de pico pueden combinarse al multiplicarlas entre sí, lo que es equivalente a tomar la primera derivada de una potencia de la elevación de pico, tal como se muestra en la Fig. 25. Los resultados presentados en la Fig. 25 suprimen exitosamente el ruido fuera de la aberración; sin embargo, el ruido dentro de la aberración se mejora mediante la manipulación. Los picos de la primera derivada son todavía claramente discernibles, permitiendo que se usen para extraer ubicaciones de borde y tolerancias laterales, permitiendo de ese modo que la aberración se identifique claramente en el trazado del perfil inferior.

Mediana de elevación cromosómica

Se espera que la mediana de la elevación normalizada dentro del cromosoma diana en un paciente euploide permanezca cerca de 1 independientemente de la fracción fetal. Sin embargo, tal como se muestra en las Fig. 9 y 10, la mediana de elevaciones en pacientes con trisomía aumenta con la fracción fetal. El aumento generalmente es sustancialmente lineal con una pendiente de 0,5. Las mediciones experimentales confirman estas expectativas.

La Fig. 26 ilustra un histograma de la mediana de elevaciones para 86 muestras euploides (mostradas en color negro en la Fig. 26). La mediana de valores se agrupa estrechamente alrededor de 1 (mediana = 1,0000, mediana de la desviación absoluta (D.M.A.) = 0,0042, media = 0,9996, desviación estándar (D.E.) = 0,0046). Ninguna de las medianas de elevaciones euploides supera 1,012, tal como se muestra en el histograma presentado en la Fig. 26. Por el contrario, de las 35 muestras de trisomía mostradas (las muestras de color gris) en la Fig. 26, todas menos una tienen medianas de elevaciones que superan 1,02, significativamente por encima del rango euploide. La brecha entre los dos grupos de pacientes en este ejemplo es lo suficientemente grande como para permitir la clasificación como euploide o aneuploide.

Fracción fetal como factor limitante en la precisión de la clasificación

La razón entre la fracción fetal y la anchura de la distribución de la mediana de recuentos normalizados en euploides (por ejemplo, embarazos euploides) puede usarse para determinar la fiabilidad de la clasificación usando la mediana

de elevaciones normalizadas, en algunas implementaciones. Dado que la mediana de recuentos normalizados, así como otros descriptores tales como valores de Z, aumentan linealmente con la fracción fetal con la constante de proporcionalidad de 0,5, la fracción fetal debe superar cuatro desviaciones estándar de la distribución de la mediana de recuentos normalizados para lograr una confianza del 95 % en la clasificación, o seis desviaciones estándar para lograr una confianza del 99 % en la clasificación. Aumentar el número de etiquetas de secuencia alineadas puede servir para disminuir el error en perfiles medidos y afilar la distribución de las medianas de elevaciones normalizadas, en determinadas implementaciones. Por tanto, el efecto de las mediciones cada vez más precisas es mejorar la razón entre la fracción fetal y la anchura de la distribución de las medianas de elevaciones normalizadas euploides.

Razón de área

La mediana de la distribución de los recuentos normalizados es generalmente una estimación puntual y, como tal, es a menudo una estimación menos fiable que las estimaciones integrales, tales como áreas bajo la distribución (por ejemplo, área bajo la curva). Las muestras que contienen fracciones de alto nivel fetal no se ven tan afectadas por el uso de una estimación puntual; sin embargo, a bajos valores de fracción fetal, se hace difícil distinguir un perfil normalizado verdaderamente elevado de una muestra euploide que tiene una mediana de recuento ligeramente aumentada debido a errores aleatorios. Un histograma que ilustra la mediana de distribución de los recuentos normalizados de un caso de trisomía con una fracción fetal relativamente baja (por ejemplo, F = aproximadamente 7 %; F(7 %)) se muestra en la Fig. 27. La mediana de la distribución es 1,021, no lejos de 1 + F/2 = 1,035. Sin embargo, la anchura de la distribución (D.M.A. = 0,054, D.E. = 0,082) supera ampliamente la desviación de la mediana del valor euploide de 1, excluyendo cualquier reivindicación de que la muestra sea anómala. La inspección visual de la distribución sugiere un análisis alternativo: aunque el desplazamiento del pico hacia la derecha es relativamente pequeño, perturba significativamente el equilibrio entre las áreas a la izquierda (de color gris oscuro) y a la derecha (de color gris claro) de la expectativa euploide de 1. Por tanto, la razón entre las dos áreas, que son una estimación integral, puede ser ventajosa en casos en los que la clasificación es difícil debido a bajos valores de fracción fetal. El cálculo de la estimación integral para las áreas de color gris claro y de color gris oscuro bajo la curva se explica con mayor detalle más adelante.

Si se supone una distribución gaussiana de recuentos normalizados, entonces

$$P(q) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-{(q - q_D)}/{(2\sigma^2)}\right] \quad (2).$$

En casos de euploides, la expectativa para los recuentos normalizados es 1. Para los pacientes con trisomía, la expectativa es

$$q_D = 1 + {F}/{2} \quad (3).$$

Dado que el punto de referencia para calcular la razón de área es 1, el argumento a la función exponencial es z^2, donde

$$z = -{F}/{(2\sigma\sqrt{2})} \quad (4).$$

El área a la izquierda del punto de referencia es

$$B = \int_{-\infty}^{1} P(q)dq = \frac{1}{2}[1 + erf(z)] \quad (5).$$

La función de error erf(z) puede evaluarse usando su expansión de Taylor:

$$erf(z) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{n!\,(2n+1)} \quad (6).$$

El área a la derecha desde el punto de referencia es 1 - B. Por tanto, la razón entre las dos áreas es

$$R = \frac{1 - B}{B} = \frac{1 - erf(z)}{1 + erf(z)} = \frac{1 - erf\left[-{F}/{(2\sigma\sqrt{2})}\right]}{1 + erf\left[-{F}/{(2\sigma\sqrt{2})}\right]} \quad (7).$$

La propagación de errores de fracciones fetales medidas en razones de área R puede estimarse simplemente al reemplazar F en la ecuación 7 por F - ΔF y F + ΔF. La Fig. 28 muestra las frecuencias de las razones de área euploide y de trisomía en un conjunto de 480 muestras. El solapamiento entre dos grupos implica muestras de trisomía con bajas fracciones fetales.

Criterios de clasificación combinados

La Fig. 29 ilustra la interrelación y la interdependencia de la mediana de elevaciones y razones de área, ambas describieron fenómenos sustancialmente similares. Las relaciones similares conectan las medianas de elevaciones y razones de área con otros criterios de clasificación, tales como puntuaciones Z, fracciones fetales ajustadas, diversas sumas de residuos al cuadrado y valores de p bayesianos (véase la Fig. 30). Los criterios de clasificación individuales pueden experimentar ambigüedad originada por la superposición parcial entre las distribuciones euploides y de trisomía en regiones de brecha; sin embargo, una combinación de múltiples criterios puede reducir o eliminar cualquier ambigüedad. La dispersión de la señal a lo largo de múltiples dimensiones puede tener el mismo efecto que medir las frecuencias de RMN de diferentes núcleos, en algunas implementaciones, resolver picos solapantes en entidades fácilmente identificables, bien definidas. Dado que no se realiza ningún intento para predecir cuantitativamente cualquier parámetro teórico usando descriptores correlacionados mutuamente, las correlaciones cruzadas observadas entre diferentes criterios de clasificación no interfieren. Definir una región en el espacio multidimensional que está poblada exclusivamente por euploides, permite clasificar cualquier muestra que esté ubicada fuera de la superficie limitante de esa región. Por tanto, el esquema de clasificación se reduce a un voto consenso para euploidía.

En algunas implementaciones que usan un enfoque de criterios de clasificación combinados, los criterios de clasificación descritos en el presente documento pueden combinarse con criterios de clasificación adicionales conocidos en la técnica. Determinadas implementaciones pueden usar un subconjunto de los criterios de clasificación enumerados en el presente documento. Determinadas implementaciones pueden combinar matemáticamente (por ejemplo, sumar, restar, dividir, multiplicar y similares) uno o más criterios de clasificación entre sí y/o con la fracción fetal para derivar nuevos criterios de clasificación. Algunas implementaciones pueden aplicar análisis de componentes principales para reducir la dimensionalidad del espacio de clasificación multidimensional. Algunas implementaciones pueden usar uno o más criterios de clasificación para definir la brecha entre pacientes afectados y no afectados y para clasificar conjuntos de datos nuevos. Cualquier combinación de criterios de clasificación puede usarse para definir la brecha entre pacientes afectados y no afectados y para clasificar nuevos conjuntos de datos. Los ejemplos no limitativos de criterios de clasificación que pueden usarse en combinación con otros criterios de clasificación para definir la brecha entre pacientes afectados y no afectados y para clasificar nuevos conjuntos de datos incluyen: análisis discriminante lineal, análisis discriminante cuadrático, análisis discriminante flexible, análisis discriminante mezcla, k vecinos más cercanos, árbol de clasificación, agregación de tipo *bootstrap*, potenciación, redes neurales, máquinas de vectores de soporte y/o bosque aleatorio.

Ejemplo 2: Métodos para la detección de variaciones genéticas asociadas con aneuploidía fetal usando fracciones fetales medidas y sumas ponderadas por bins de residuos al cuadrado

Las estadísticas del valor de Z y otros análisis estadísticos de los datos de lectura de secuencia son a menudo adecuados para determinar o proporcionar un resultado determinante de la presencia o ausencia de una variación genética con respecto a la aneuploidía fetal, sin embargo, en algunos casos puede ser útil incluir un análisis adicional basándose en la contribución de fracción fetal y las suposiciones de ploidía. Cuando se incluye la contribución de fracción fetal en un esquema de clasificación, se usa generalmente una mediana de perfil de recuento de referencia de un conjunto de euploides conocidos (por ejemplo, embarazos euploides) para la comparación. Una mediana de perfil de recuento de referencia puede generarse al dividir todo el genoma en N bins, donde N es el número de bins. A cada bin i se le asignan dos números: (i) un recuento de referencia F_i y (ii) la incertidumbre (por ejemplo, desviación estándar o σ) para los recuentos de referencia de bins.

La siguiente relación puede utilizarse para incorporar fracción fetal, ploidía materna y mediana de recuentos de referencia en un esquema de clasificación para determinar la presencia o ausencia de una variación genética con respecto a aneuploidía fetal,

$$y_i = (1 - F)M_i f_i + FXf_i \qquad (8)$$

donde Y_i representa los recuentos medidos para un bin en la muestra de prueba correspondiente al bin en la mediana de perfil de recuento, F representa la fracción fetal, X representa la ploidía fetal y M_i representa la ploidía materna asignada a cada bin. Los posibles valores usados para la ecuación Xin (8) son: 1 si el feto es euploide; 3/2, si el feto es triploide; y 5/4, si hay fetos gemelares y uno se ve afectado y el otro no. 5/4 se usa en el caso de gemelos en los que un feto se ve afectado y el otro no, porque el término F en la ecuación (8) representa el ADN fetal total, por tanto, debe tenerse en cuenta todo el ADN fetal. En algunas implementaciones, las deleciones y/o duplicaciones grandes en el genoma materno pueden tenerse en cuenta mediante la asignación de ploidía materna, M_i, a cada bin o sección genómica. La ploidía materna se asigna a menudo como un múltiplo de 1/2, y puede estimarse usando normalización basada en bins, en

algunas implementaciones. Debido a que la ploidía materna es a menudo un múltiplo de 1/2, la ploidía materna puede tenerse en cuenta fácilmente y, por tanto, no se incluirá en ecuaciones adicionales para simplificar las derivaciones.

La ploidía fetal puede evaluarse usando cualquier enfoque adecuado. En algunas implementaciones, la ploidía fetal puede evaluarse usando la ecuación (8), o derivaciones de esta. En determinadas implementaciones, la ploidía fetal puede clasificarse usando uno de los siguientes enfoques basados en la ecuación (8), no limitativos:

1) Medir la fracción fetal F y usar el valor para formar dos sumas de residuos al cuadrado. Para calcular la suma de residuos al cuadrado, se resta el lado derecho (RHS) de la ecuación (8) de su lado izquierdo (LHS), se eleva al cuadrado la diferencia, y se suma por los bins genómicos seleccionados, o en aquellas implementaciones que usan todos los bins, se suma por todos los bins. Este procedimiento se realiza para calcular cada una de las dos sumas de residuos al cuadrado. Una suma de residuos al cuadrado se evalúa con ploidía fetal fijada en 1 (por ejemplo, $X = 1$) y la otra suma de residuos al cuadrado se evalúa con ploidía fetal fijada en 3/2 (por ejemplo, $X = 3/2$). Si el sujeto de prueba fetal es euploide, la diferencia entre las dos sumas de residuos al cuadrado es negativa, de lo contrario, la diferencia es positiva.

2) Se fija la fracción fetal en su valor medido y se optimiza el valor de ploidía. La ploidía fetal puede adoptar generalmente solo 1 de dos valores discretos, 1 o 3/2; sin embargo, la ploidía algunas veces puede tratarse como una función continua. La regresión lineal puede usarse para generar una estimación de ploidía. Si el cálculo resultante del análisis de regresión lineal es próximo a 1, la muestra de prueba fetal puede clasificarse como euploide. Si la estimación está cerca de 3/2, el feto puede clasificarse como triploide.

3) Fijar la ploidía fetal y optimizar la fracción fetal usando el análisis de regresión lineal. La fracción fetal puede medirse y puede incluirse un término de restricción para mantener la fracción fetal ajustada próxima al valor medido de fracción fetal, con una función de ponderación que es proporcional a la inversa del error estimado en la fracción fetal medida. La ecuación (8) se resuelve dos veces, una vez con la ploidía configurada en 3/2, y una vez para la ploidía fetal configurada en 1. Cuando se resuelve la ecuación (8) con la ploidía configurada en 1, no es necesario ajustar la fracción fetal. Se forma una suma de residuos al cuadrado para cada resultado y se resta la suma de residuos al cuadrado. Si la diferencia es negativa, el sujeto de prueba fetal es euploide. Si la diferencia es positiva, el sujeto de prueba fetal es triploide.

Los enfoques generalizados descritos en 1), 2) y 3) se describen con mayor detalle en el presente documento.

Ploidía fija, fracción fetal fija: sumas de residuos al cuadrado

En algunas implementaciones, la aneuploidía fetal puede determinarse usando un modelo que analiza dos variables, ploidía fetal (por ejemplo, *X)* y fracción de ácido nucleico fetal (por ejemplo, fracción fetal; *F).* En determinadas implementaciones, la ploidía fetal puede adoptar valores discretos y, en algunas implementaciones, la fracción fetal puede ser una continuidad de valores. La fracción fetal puede medirse, y el valor medido puede usarse para generar un resultado para la ecuación (8), para cada valor posible para la ploidía fetal. Los valores de ploidía fetal que pueden usarse para generar un resultado para la ecuación (8) incluyen 1 y 3/2 para un embarazo de un solo feto, y en el caso de un embarazo de fetos gemelares en el que un feto está afectado y el otro feto no está afectado, puede usarse 5/4. La suma de residuos al cuadrado obtenidos para cada valor de ploidía fetal mide el éxito con el cual el método reproduce las mediciones, en algunas implementaciones. Cuando se evalúa la ecuación (8) en X = 1, (por ejemplo, suposición euploide), la fracción fetal se cancela y resulta la siguiente ecuación para la suma de residuos al cuadrado:

$$\varphi_E = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}(y_i - f_i)^2 = \sum_{i=1}^{N} \frac{y_i^2}{\sigma_i^2} - 2\sum_{i=1}^{N} \frac{y_i f_i}{\sigma_i^2} + \sum_{i=1}^{N} \frac{f_i^2}{\sigma_i^2} = \Xi_{yy} - 2\Xi_{fy} + \Xi_{ff} \quad (9)$$

Para simplificar la ecuación (9) y los cálculos subsiguientes, se utiliza la siguiente noción:

$$\Xi_{yy} = \sum_{i=1}^{N} \frac{y_i^2}{\sigma_i^2} \quad (10)$$

$$\Xi_{ff} = \sum_{i=1}^{N} \frac{f_i^2}{\sigma_i^2} \quad (11)$$

$$\Xi_{fy} = \sum_{i=1}^{N} \frac{y_i f_i}{\sigma_i^2} \quad (12)$$

Cuando se evalúa la ecuación (8) en X = 3/2 (por ejemplo, suposición triploide), resulta la siguiente ecuación para la suma de los residuos al cuadrado:

$$\varphi_T = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}\left(y_i - f_i - \frac{1}{2}Ff_i\right)^2 = \Xi_{yy} - 2\Xi_{fy} + \Xi_{ff} + F\left(\Xi_{ff} - \Xi_{fy}\right) + \frac{1}{4}F^2\Xi_{ff} \quad (13)$$

La diferencia entre las ecuaciones (9) y (13) forma el resultado funcional (por ejemplo, phi) que puede usarse para someter a prueba la hipótesis nula (por ejemplo, euploide, $X = 1$) contra la hipótesis alternativa (por ejemplo, trisomía única, $X = 3/2$):

$$\varphi = \varphi_E - \varphi_T = F\left(\Xi_{fy} - \Xi_{ff}\right) - \frac{1}{4}F^2\Xi_{ff} \quad (14)$$

El perfil de phi con respecto a *F* es una parábola definida a la derecha de la ordenada (ya que *F* es mayor de o igual a 0). Phi converge al origen a medida que *F* se aproxima a cero, independientemente de los errores experimentales e incertidumbres experimentales en los parámetros del modelo.

En algunas implementaciones, la phi funcional depende de fracción fetal F medida con un coeficiente cuadrático negativo de segundo orden (véase la ecuación (14)). La dependencia de phi de la fracción fetal medida parecería implicar una forma convexa para los casos de euploides y triploides. Si este análisis fuese correcto, los casos de trisomía revertirían el signo a altos valores de *F*; sin embargo, la ecuación (12) depende de F. La combinación de las ecuaciones (8) y (14), sin considerar la ploidía materna, fijando X = 3/2 y despreciando los errores experimentales, la ecuación para casos de trisomía se convierte en:

$$\Xi_{fy} = \sum_{i=1}^{N} \frac{y_i f_i}{\sigma_i^2} = \sum_{i=1}^{N} \frac{f_i}{\sigma_i^2}\left[(1-F)f_i + FXf_i\right] = \left(1 + \frac{1}{2}F\right)\sum_{i=1}^{N} \frac{f_i^2}{\sigma_i^2} = \left(1 + \frac{1}{2}F\right)\Xi_{ff}$$

(15)

La relación entre las ecuaciones (11) y (12) para los triploides sigue siendo cierta en circunstancias ideales, en ausencia de cualquier error de medición. La combinación de las ecuaciones (14) y (15) da como resultado la siguiente expresión, que a menudo produce una parábola cóncava en casos triploides:

$$\varphi = F\left(\Xi_{fy} - \Xi_{ff}\right) - \frac{1}{4}F^2\Xi_{ff} = F\left[\left(1 + \frac{1}{2}F\right)\Xi_{ff} - \Xi_{ff}\right] - \frac{1}{4}F^2\Xi_{ff} = \frac{1}{4}F^2\Xi_{ff} \quad \text{(Trisomy)}$$

(16)

Para los euploides, las ecuaciones (11) y (12) deben tener el mismo valor, con la excepción de errores de medición, que a veces produce una parábola convexa:

$$\varphi = F\left(\Xi_{fy} - \Xi_{ff}\right) - \frac{1}{4}F^2\Xi_{ff} = -\frac{1}{4}F^2\Xi_{ff} \quad \text{(Euploides)} \quad (17)$$

Los perfiles de phi funcionales simulados para los valores típicos de parámetros del modelo se muestran en la Fig. 31, para casos de trisomía (de color gris) y euploides (de color azul). La Fig. 32 muestra un ejemplo usando datos reales. En las Fig. 31 y 32, los puntos de datos debajo de las abscisas representan generalmente casos clasificados como euploides. Los puntos de datos por encima de las abscisas representan generalmente casos clasificados como casos de trisomía 21 (T21). En la Fig. 32, el punto de datos solitarios en el cuarto cuadrante (por ejemplo, cuadrante inferior intermedio) es un embarazo gemelar con un feto afectado. El conjunto de datos usado para generar la Fig. 32 incluye otras muestras de gemelos afectados también, que explican la dispersión de puntos de datos de T21 hacia las abscisas.

Las ecuaciones (9) y (10) pueden interpretarse a menudo de la siguiente manera: Para triploides, el modelo euploide a veces genera errores más grandes, lo que implica que las phi (véase la ecuación (9)) es mayor que phi_T (véase la ecuación (13)). Como resultado, la phi funcional (véase la ecuación (7)) ocupa el primer cuadrante (por ejemplo, cuadrante superior izquierdo). Para los euploides, el modelo de trisomía a veces genera errores más grandes, el rango de las ecuaciones (2) y (6) se invierte y la phi funcional (ecuación (7)) ocupa en el cuarto cuadrante. Por tanto, en principio, la clasificación de una muestra como euploide o triploide a veces se reduce para evaluar la señal de phi.

En algunas implementaciones, la curvatura de los puntos de datos que se muestran en las Fig. 31 y 32 pueden reducirse o eliminarse reemplazando phi funcional (ecuación (7)) por la raíz cuadrada del valor absoluto del phi funcional, multiplicada por su signo. La relación lineal generada con respecto a *F* a veces puede mejorar la separación entre triploides y euploides a

bajos valores de fracción fetal, tal como se muestra en la Fig. 33. Linealizar la relación con respecto a *F* a veces da como resultado un aumento de los intervalos de incertidumbre a bajos valores de fracción fetal (por ejemplo, *F)*, por tanto, las ganancias realizadas a partir de este procedimiento se relacionan con realizar una inspección visual de las diferencias sustancialmente más fácil; el área de color gris permanece sin cambios. La extensión del procedimiento para analizar los embarazos gemelares es relativamente sencilla. El motivo usado para generar la ecuación (9) implica que en un embarazo gemelar con un feto afectado y un feto normal, la phi funcional debe reducirse a cero, más o menos el error experimental, independientemente de *F.* Los embarazos gemelares producen generalmente más ADN fetal que los embarazos únicos.

Ploidía optimizada, fracción fetal fija: regresión lineal

En determinadas implementaciones, la aneuploidía fetal puede determinarse usando un modelo en el cual la fracción fetal se fija en su valor medido y la ploidía se varía para optimizar la suma de residuos al cuadrado. En algunas implementaciones, el valor de fracción fetal ajustada resultante puede usarse para clasificar un caso como trisomía o euploide, dependiendo de si el valor es próximo a 1, 3/2 o 5/4 en el caso de gemelos. Partiendo de la ecuación (8), la suma de los residuos al cuadrado puede formarse de la siguiente manera:

$$\varphi = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} [y_i - (1-F)M_i f_i - FXf_i]^2$$

$$= \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[y_i^2 - 2(1-F)M_i f_i y_i - 2FXf_i y_i + (1-F)^2 M_i^2 f_i^2 + 2F(1-F)XM_i f_i^2 + F^2 X^2 f_i^2 \right]$$

(18)

Para minimizar phi en función de X, se genera la primera derivada de phi con respecto a X, se establece igual a cero, y la ecuación resultante se resuelve para X. La expresión resultante se presenta en la ecuación (19).

$$\frac{1}{2}\left(\frac{d\varphi}{dX}\right) = 0 = XF^2 \sum_{i=1}^{N} \frac{f_i^2}{\sigma_i^2} - F \sum_{i=1}^{N} \frac{f_i y_i}{\sigma_i^2} + F(1-F) \sum_{i=1}^{N} \frac{M_i f_i^2}{\sigma_i^2} \quad (19)$$

El valor de ploidía óptimo a veces viene dado por la siguiente expresión:

$$X = \frac{\sum_{i=1}^{N} \frac{f_i y_i}{\sigma_i^2} - (1-F) \sum_{i=1}^{N} \frac{M_i f_i^2}{\sigma_i^2}}{F \sum_{i=1}^{N} \frac{f_i^2}{\sigma_i^2}} \quad (20)$$

Tal como se mencionó anteriormente, el término para ploidía materna, M_i, puede omitirse de derivaciones matemáticas adicionales. La expresión resultante para X corresponde al caso especial relativamente simple, y a menudo que se produce más frecuentemente, cuando la madre no tiene deleciones o duplicaciones en el cromosoma o cromosomas que se evalúan. La expresión resultante se presenta en la Fig. 21.

$$X = \frac{\Xi_{fy} - (1-F)\Xi_{ff}}{F\Xi_{ff}} = \frac{\Xi_{fy}}{F\Xi_{ff}} - \frac{1-F}{F} = 1 + \frac{1}{F}\left(\frac{\Xi_{fy}}{\Xi_{ff}} - 1\right) \quad (21)$$

Xiff y Xify vienen dados por las ecuaciones (11) y (12), respectivamente. En implementaciones en las que todos los errores experimentales son insignificantes, resolver la ecuación (21) da como resultado un valor de 1 para euploides donde $Xi_{ff} = Xi_{fy}$. En determinadas implementaciones en las que todos los errores experimentales son insignificantes, resolver la ecuación (21) da como resultado un valor de 3/2 para triploides (véase la ecuación (15) para la relación triploide entre Xi_{ff} y Xi_{fy}).

Ploidía optimizada, fracción fetal fija: propagación de errores

La ploidía optimizada es a menudo inexacta debido a diversas fuentes de error. Tres ejemplos no limitativos de fuentes de error incluyen: recuentos de bins de referencia *f,* recuentos de bins medidos y_i y fracción fetal *F.* La contribución de los ejemplos no limitativos de error se examinará por separado.

Errores en las fracciones fetales medidas: calidad de fracción fetal ajustada

Las estimaciones de fracciones fetales basadas en la cantidad de etiquetas de secuencia mapeadas en el cromosoma Y (por ejemplo, recuentos Y) a veces muestran desviaciones relativamente grandes con respecto a los valores de fracciones fetales de FQA (véase la Fig. 34). Los valores de Z para triploide a menudo también presentan una dispersión relativamente ancha alrededor de la diagonal mostrada en la Fig. 35. La línea diagonal en la Fig. 35 representa un aumento esperado teóricamente de la representación cromosómica para el cromosoma 21 con una fracción fetal creciente en los casos de trisomía 21. La fracción fetal puede estimarse usando un método adecuado. Un ejemplo no limitativo de un método que puede usarse para estimar la fracción fetal es el ensayo cuantificador fetal (por ejemplo, FQA). Otros métodos para estimar la fracción fetal se conocen en la técnica. A veces, diversos métodos usados para estimar la fracción fetal también muestran una dispersión sustancialmente similar alrededor de la diagonal central, tal como se muestra en las Fig. 36-39. En la Fig. 36, las desviaciones son sustancialmente similares (por ejemplo, negativas a F_0 alta) a las observadas en la fracción fetal ajustada (véase ecuación (33)). En algunas implementaciones, la pendiente de la aproximación lineal a la fracción fetal del cromosoma Y (por ejemplo, el cromosoma Y) promedio (véase la línea de color gris oscuro en la Fig. 36) en el rango entre el 0 % y el 20 % es de aproximadamente 3/4. En determinadas implementaciones, la aproximación lineal para la desviación estándar (véase la Fig. 36, línea de color gris claro) es de aproximadamente 2/3 + F_0/6. En algunas implementaciones, las estimaciones de fracción fetal basadas en el cromosoma 21 (por ejemplo, el cromosoma 21) son sustancialmente similares a las obtenidas mediante el ajuste de fracciones fetales (véase la Fig. 37). Otro conjunto cualitativamente similar de estimaciones de fracción fetal basadas en el sexo se muestra en la Fig. 38. La Fig. 39 ilustra las medianas de los recuentos de bins normalizados para casos de T21, que se espera que tengan una pendiente cuya aproximación lineal es sustancialmente similar a 1 + F_0/2 (véase la línea de color gris desde el origen hasta el punto medio de la parte superior del gráfico en la Fig. 39).

Las Fig. 36-39 comparten las siguientes características comunes:

a) pendiente no igual a 1 (ya sea mayor o menor de 1, dependiendo del método, con excepción de los valores de Z),

b) estimación de fracción fetal de gran dispersión, y

c) el alcance de la dispersión aumenta con la fracción fetal.

Para tener en cuenta estas observaciones, los errores en la fracción fetal medida se modelarán usando la fórmula ΔF = 2/3 + F_0/6, en algunas implementaciones.

Errores en fracciones fetales medidas: propagación de errores de fracciones fetales medidas con respecto a ploidía ajustada

Si se supone que f_i e y_i no tienen errores, para simplificar el análisis, la fracción fetal medida F se compone de F_v (por ejemplo, la verdadera fracción fetal) y ΔF (por ejemplo, el error en la fracción fetal medida):

$$F = F_v + \Delta F \qquad (22).$$

En algunos casos, las incertidumbres en los valores de X ajustados se originan a partir de errores en la fracción fetal medida, F. Los valores optimizados para X vienen dados por la ecuación (21), sin embargo, el verdadero valor de ploidía viene dado por X_v, donde X_v = 1 o 3/2. X_v varía discretamente, mientras que X varía continuamente y solo se acumula alrededor de X_v en condiciones favorables (por ejemplo, error relativamente bajo).

Suponiendo de nuevo que A e y no tienen errores, la ecuación (8) se convierte en:

$$y_i = (1 - F_V)M_i f_i + F_V X f_i \quad (23).$$

La combinación de las ecuaciones (21) a (23) genera la siguiente relación entre ploidía verdadera X_v y la estimación de ploidía X que incluye el error ΔF. La relación también incluye la suposición de que la ploidía materna es igual a 1 (por ejemplo, euploide), y el término para ploidía materna, M_i, se reemplaza por 1.

$$X = 1 + \frac{1}{F_V + \Delta F}\left\{\frac{\sum_{i=1}^{N} \frac{f_i}{\sigma_i^2}[(1-F_V)f_i + F_V X_V f_i]}{\sum_{i=1}^{N} \frac{f_i^2}{\sigma_i^2}} - 1\right\} = 1 + \frac{F_V(X_V - 1)}{F_V + \Delta F} \qquad (24)$$

En algunos casos, el término X_v- 1 es sustancialmente idéntico a cero en euploides, y ΔF no contribuye a errores en X. En casos triploides, el término de error no se reduce a cero (por ejemplo, no es sustancialmente idéntico a cero). Por tanto, en algunas implementaciones, los cálculos de ploidía pueden considerarse en función del error ΔF.

$$X = g(\Delta F) \qquad (25)$$

Los perfiles simulados de X triploide ajustado en función de F_0 con errores fijos ΔF = más o menos el 0,2 % se muestran en la Fig. 40. Los resultados obtenidos usando datos reales se muestran en la Fig. 41. Los puntos de datos se ajustan generalmente al contorno asimétrico abocinado predicho por la ecuación (24).

A menudo, las fracciones fetales más pequeñas se asocian cualitativamente con errores de ploidía más grandes. La fracción fetal subestimada algunas veces se compensa con sobreestimaciones de ploidía; la fracción fetal sobreestimada a menudo se vincula con subestimaciones en ploidía. A menudo, el efecto es más fuerte cuando la fracción fetal es subestimada. Esto concuerda con la asimetría observada en los gráficos presentados en las Fig. 40 y 41, (por ejemplo, a medida que *F* disminuye, el crecimiento de la rama superior es sustancialmente más rápido que la disminución de la rama inferior). Las simulaciones con diferentes niveles de error en *F* siguen el mismo patrón, aumentando la extensión de las desviaciones de X_v con ΔF.

Puede usarse una distribución de probabilidad para *X* para cuantificar estas observaciones. En algunas implementaciones, la distribución de ΔF puede usarse para derivar la función de densidad para *X* usando la siguiente expresión:

$$f_Y(y) = \left|\frac{1}{g'(g^{-1}(y))}\right| f_X(g^{-1}(y)) \qquad (26)$$

donde,

$f_Y(y)$ es la función de densidad desconocida para $y = g(x)$

$f_x(x)$ es la función de densidad dada para x

$g'(x)$ es la primera derivada de la función dada $y = g(x)$

$g^{-1}(y)$ es la inversa de la función g dada: $x = g^{-1}(y)$

$g'(g^{-1}(y))$ es el valor de la derivada en el punto $g^{-1}(y)$

En la ecuación 26, x es ΔF, y es X (por ejemplo, estimación de ploidía), y $g(x)$ viene dado por la ecuación (24). La derivada se evalúa según la siguiente expresión:

$$\frac{dg}{d\Delta F} = -\frac{F_V(X_V - 1)}{(F_V + \Delta F)^2} \qquad (27)$$

La inversa $g^{-1}(y)$ puede obtenerse a partir de la ecuación (24), en algunas implementaciones:

$$\Delta F = \frac{F_V(X_V - X)}{X - 1} \qquad (28)$$

Si el error en *F* se ajusta a una distribución gaussiana, $f_x(x)$ en la ecuación (26) puede reemplazarse por la siguiente expresión:

$$P(\Delta F) = \frac{exp[-(\Delta F)^2/(2\sigma^2)]}{\sigma\sqrt{2\pi}} \qquad (29)$$

En determinadas implementaciones, combinar las ecuaciones (26) a (29) da como resultado una distribución de probabilidad para X a diferentes niveles de ΔF, tal como se muestra en la Fig. 42.

En algunos casos, un sesgo hacia valores de ploidía más altos, que algunas veces son prominentes a altos niveles de errores en F, a menudo se refleja en la forma asimétrica de la función de densidad: una cola relativamente larga, que disminuye lentamente a la derecha de la línea de color gris claro, verticalmente en línea con *X*, a lo largo del eje X, tal como se muestra en la Fig. 42, paneles A-C. En algunas implementaciones, para cualquier valor de ΔF, el área bajo la función de densidad de probabilidad a la izquierda de la línea de color gris claro (X_v = 3/2) es igual al área a la derecha de la línea de

color gris claro. Es decir, la mitad de todos los valores de ploidía ajustados a menudo son sobreestimaciones, mientras que la otra mitad de todos los valores de ploidía ajustados a veces son subestimaciones. En algunos casos, el sesgo generalmente solo se refiere a la extensión de errores en *X,* no a la prevalencia de una u otra dirección. La mediana de la distribución permanece igual a X_v, en algunas implementaciones. La Fig. 43 ilustra las distribuciones euploides y de trisomía obtenidas para los datos reales. Las incertidumbres en los valores medidos de fracción fetal explican, algunas veces, parte de la varianza observada en los valores de ploidía ajustados para triploides, sin embargo, los errores en los valores de X estimados para euploides requieren a menudo examinar la propagación de errores a partir de los recuentos de bins.

Ploidía fija, fracción fetal optimizada: regresión lineal

Una fracción fetal continuamente variable puede optimizarse a menudo mientras se mantiene la ploidía fija en uno de sus posibles valores (por ejemplo, 1 para euploides, 3/2 para triploides únicos, 5/4 para triploides gemelares), en oposición a la ploidía de ajuste que a menudo puede adoptar un número limitado de valores discretos conocidos. En implementaciones en las que se conoce la fracción fetal medida (F_0), la optimización de fracción fetal puede restringirse de tal manera que *F* ajustada permanezca próxima a F_0, dentro del error experimental (por ejemplo, ΔF). En algunos casos, la fracción fetal observada (por ejemplo, medida) F_0, a veces difiere de fracción fetal, F_v, descrita en las ecuaciones (22) a (28). Un análisis robusto de propagación de errores debe ser capaz de distinguir entre F_0 y F_v. Para simplificar las siguientes derivaciones, se ignorará la diferencia entre la fracción fetal observada y la fracción fetal verdadera.

La ecuación (8) se presenta a continuación en un formato reordenado que también omite el término de ploidía materna (por ejemplo, M_i).

$$y_i = F(X-1)f_i + f_i \quad (30)$$

Un término funcional que necesita minimizarse se define de la siguiente manera, en algunas implementaciones:

$$\varphi(F) = \frac{(F-F_0)^2}{(\Delta F)^2} + \sum_{i=1}^{N} \frac{1}{\sigma_i^2}[y_i - F(X-1)f_i - f_i]^2$$

$$= \frac{(F-F_0)^2}{(\Delta F)^2} + \sum_{i=1}^{N} \frac{1}{\sigma_i^2}\left[y_i^2 + F^2(X-1)^2 f_i^2 + f_i^2 - 2F(X-1)f_i y_i - 2f_i y_i + 2F(X-1)f_i^2\right]$$

$$= \frac{(F-F_0)^2}{(\Delta F)^2} + F^2(X-1)^2 \sum_{i=1}^{N} \frac{f_i^2}{\sigma_i^2} + 2F(X-1)\sum_{i=1}^{N} \frac{f_i^2 - f_i y_i}{\sigma_i^2} + \sum_{i=1}^{N} \frac{(y_i - f_i)^2}{\sigma_i^2}$$

(31)

Cuando la ecuación (31) se evalúa para euploides (por ejemplo, $X = 1$), el término $\frac{(F-F_0)^2}{(\Delta F)^2}$ depende a menudo de *F,* por tanto, *F* ajustada frecuentemente es igual a F_0, En algunos casos, cuando la ecuación (24) se evalúa para euploides, la ecuación algunas veces se reduce a $\sum_{i=1}^{N} \frac{(y_i - f_i)^2}{\sigma_i^2}$.

Cuando la ecuación (24) se evalúa para casos de trisomía únicos (por ejemplo, $X = 3/2$), los coeficientes que multiplican *F* contienen tanto mediciones de fracción fetal como recuentos de bins, por tanto, el valor optimizado para *F* depende a menudo de ambos parámetros. La primera derivada de la ecuación (24) con respecto a *F* se reduce a cero en algunos casos:

$$\frac{1}{2}\left(\frac{d\varphi}{dF}\right) = 0 = \frac{(F-F_0)}{(\Delta F)^2} + F(X-1)^2 \sum_{i=1}^{N} \frac{f_i^2}{\sigma_i^2} + (X-1)\sum_{i=1}^{N} \frac{f_i^2 - f_i y_i}{\sigma_i^2} \quad (32)$$

En algunas implementaciones, reemplazar $X = 3/2$ y resolver la ecuación (32) para *F* produce un valor optimizado para *F:*

$$F = \frac{F_0 + \frac{(\Delta F)^2}{2}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left(f_i y_i - f_i^2\right)}{1 + \frac{(\Delta F)^2}{4}\sum_{i=1}^{N} f_i^2/\sigma_i^2} \quad (33).$$

Para simplificar otros cálculos y/o derivaciones, se utilizarán las siguientes variables auxiliares:

$$S_0 = \frac{(\Delta F)^2}{4}\sum_{i=1}^{N}\frac{1}{\sigma_i^2} \quad (34)$$

$$S_f = \frac{(\Delta F)^2}{4}\sum_{i=1}^{N}\frac{f_i}{\sigma_i^2} \quad (35)$$

$$S_y = \frac{(\Delta F)^2}{4}\sum_{i=1}^{N}\frac{y_i}{\sigma_i^2} \quad (36)$$

$$S_{yy} = \frac{(\Delta F)^2}{4}\sum_{i=1}^{N}\frac{y_i^2}{\sigma_i^2} \quad (37)$$

$$S_{ff} = \frac{(\Delta F)^2}{4}\sum_{i=1}^{N}\frac{f_i^2}{\sigma_i^2} \quad (38)$$

$$S_{fy} = \frac{(\Delta F)^2}{4}\sum_{i=1}^{N}\frac{y_i f_i}{\sigma_i^2} \quad (39)$$

Utilizando las variables auxiliares, la fracción fetal optimizada para $X = 3/2$ para la ecuación (33) entonces se reduce a:

$$F = \frac{F_0 + 2S_{fy} - 2S_{ff}}{1 + S_{ff}} \quad (40)$$

A menudo, *F* ajustada es linealmente proporcional al valor medido F_0, pero algunas veces no es necesariamente igual a F_0. La razón entre errores en las mediciones de fracción fetal y las incertidumbres en los recuentos de bins determina el peso relativo dado al F_0 medido en comparación con los bins individuales, en algunas implementaciones. En algunos casos, cuanto mayor sea el error ΔF, mayor será la influencia que ejercerán los recuentos de bins sobre *F* ajustada. Alternativamente, la ΔF pequeña implica generalmente que el valor de F ajustado estará dominado por F_0. En algunas implementaciones, si un conjunto de datos proviene de una muestra de trisomía, y todos los errores son insignificantes, la ecuación (40) se reduce a identidad entre *F* y F_0. A modo de prueba matemática, el uso de ploidía fetal establecida en X = 3/2, y suponiendo que F_0 (observado) y F_v (verdadero) tienen el mismo valor, la ecuación (30) se convierte en:

$$y_i = \frac{1}{2}F_0 f_i + f_i \quad (41)$$

La suposición de que F_0 y F_v generalmente es una suposición aceptable por motivos del análisis cualitativo presentado en el presente documento. La combinación de las ecuaciones (39) y (41) produce

$$S_{fy} = \frac{(\Delta F)^2}{4}\sum_{i=1}^{N}\frac{y_i f_i}{\sigma_i^2} = \frac{(\Delta F)^2}{4}\sum_{i=1}^{N}\frac{\left(\frac{1}{2}F_0 f_i + f_i\right)f_i}{\sigma_i^2} = \left(\frac{1}{2}F_0 + 1\right)S_{ff} \quad (42)$$

La combinación de las ecuaciones (40) y (42) da como resultado identidad entre F_0 y F_v:

$$F = \frac{F_0 + 2S_{fy} - 2S_{ff}}{1+S_{ff}} = \frac{F_0 + 2\left(\frac{1}{2}F_0 + 1\right)S_{ff} - 2S_{ff}}{1+S_{ff}} = \frac{F_0(1+S_{ff})}{1+S_{ff}} \equiv F_0 \qquad \text{QED} \quad (43)$$

Para ilustrar adicionalmente el modelo teórico, si la ploidía verdadera es 1 (por ejemplo, euploide), pero el valor de ploidía utilizado en la ecuación (40) se establece en X = 3/2 (por ejemplo, único triploide), la *F* ajustada resultante no es igual a F_0, ni reduce a cero, y la siguiente expresión es generalmente verdadera:

$$y_i = f_i \Rightarrow S_{fy} = \frac{(\Delta F)^2}{4}\sum_{i=1}^{N}\frac{y_i f_i}{\sigma_i^2} = \frac{(\Delta F)^2}{4}\sum_{i=1}^{N}\frac{f_i^2}{\sigma_i^2} = S_{ff} \Rightarrow F = \frac{F_0 + 2S_{fy} - 2S_{ff}}{1+S_{ff}} = \frac{F_0}{1+S_{ff}}$$

(44).

Por tanto, la aplicación de ecuaciones triploides cuando se somete a prueba un caso euploide da como resultado generalmente una *F* ajustada distinta de cero que es proporcional a F_0 con un coeficiente de proporcionalidad entre 0 y 1 (excluyente), dependiendo de los recuentos de bins de referencia e incertidumbres asociadas (remítase a la ecuación (38)), en determinadas implementaciones. Un análisis similar se muestra en la Fig. 44, usando datos reales de 86 euploides conocidos como referencia. La pendiente de la línea recta de la ecuación (44) es próxima a 20 grados, tal como se muestra en la Fig. 44.

El punto de datos solitarios entre casos de euploides y T21 (por ejemplo, fracción fetal medida de aproximadamente el 40 %, fracción ajustada de aproximadamente el 20 %) representa un gemelo con T21. Cuando se supone una ΔF constante, la rama euploide del gráfico que se muestra en la Fig. 44 generalmente está inclinada, sin embargo, cuando se usa $\Delta F = 2/3 + F_0/6$, la rama euploide del gráfico a menudo se vuelve sustancialmente horizontal, tal como se describe en el presente documento en la sección titulada "Ploidía fija, fracción fetal optimizada, propagación de errores: fracciones fetales ajustadas".

Ploidía fija, fracción fetal optimizada: sumas de residuos al cuadrado

En algunos casos para los casos de euploides, en los que *F* ajustada para la ecuación (32) es igual a F_0 y X = 1, la suma de residuos al cuadrado para un modelo euploide sigue de la ecuación (31):

$$\varphi_E = \sum_{i=1}^{N}\frac{1}{\sigma_i^2}(y_i - f_i)^2 = \Xi_{yy} - 2\Xi_{fy} + \Xi_{ff} \quad (45)$$

que es prácticamente el mismo resultado que la ecuación (9). En determinados casos para casos de euploides, la ecuación (40) puede combinarse en la ecuación (31). La expresión matemática resultante depende cuadráticamente de F_0, en algunas implementaciones. En determinadas implementaciones, la clasificación de una variación genética se realiza restando la suma triploide de los residuos al cuadrado de la suma euploide de los residuos al cuadrado. El resultado de la clasificación obtenida restando la suma triploide de residuos al cuadrado de la suma euploide de residuos al cuadrado depende además a menudo de F_0:

$$\varphi_E - \varphi_T$$

$$= \frac{-1}{(\Delta F)^2}\left[\left(\frac{F_0+2S_{fy}-2S_{ff}}{1+S_{ff}} - F_0\right)^2 + \left(\frac{F_0+2S_{fy}-2S_{ff}}{1+S_{ff}}\right)^2 \frac{(\Delta F)^2}{4}\sum_{i=1}^{N}\frac{f_i^2}{\sigma_i^2} + \left(\frac{F_0+2S_{fy}-2S_{ff}}{1+S_{ff}}\right)(\Delta F)^2\sum_{i=1}^{N}\frac{f_i^2-f_i y_i}{\sigma_i^2}\right]$$

$$= \frac{-1}{(\Delta F)^2}\left[\left(\frac{F_0+2S_{fy}-2S_{ff}}{1+S_{ff}} - F_0\right)^2 + \left(\frac{F_0+2S_{fy}-2S_{ff}}{1+S_{ff}}\right)^2 S_{ff} + 4\left(\frac{F_0+2S_{fy}-2S_{ff}}{1+S_{ff}}\right)(S_{ff}-S_{fy})\right]$$

$$= \frac{-\left[(2S_{fy}-2S_{ff}-F_0S_{ff})^2+(F_0+2S_{fy}-2S_{ff})^2S_{ff}+4(F_0+2S_{fy}-2S_{ff})(1+S_{ff})(S_{ff}-S_{fy})\right]}{(\Delta F)^2(1+S_{ff})^2}$$

$$= \frac{-1}{(\Delta F)^2(1+S_{ff})^2}\left[(4S_{fy}^2 + 4S_{ff}^2 + F_0^2S_{ff}^2 - 8S_{fy}S_{ff} - 4F_0S_{fy}S_{ff} + 4F_0S_{ff}^2)\right.$$

$$+(F_0^2S_{ff} + 4S_{fy}^2S_{ff} + 4S_{ff}^3 + 4F_0S_{fy}S_{ff} - 4F_0S_{ff}^2 - 8S_{fy}S_{ff}^2)$$

$$+(4F_0S_{ff} + 8S_{fy}S_{ff} - 8S_{ff}^2 - 4F_0S_{fy} - 8F_0S_{fy} + 8S_{fy}S_{ff}$$

$$\left.+4F_0S_{ff}^2 + 8S_{fy}S_{ff}^2 - 8S_{ff}^3 - 4F_0S_{fy}S_{ff} - 8S_{fy}^2S_{ff} + 8S_{fy}S_{ff}^2)\right]$$

$$= \frac{-1}{(\Delta F)^2(1+S_{ff})}\left[F_0^2S_{ff} + 4F_0(S_{ff}-S_{fy}) - 4(S_{ff}-S_{fy})^2\right]$$

(46)

El término depende generalmente de la fracción fetal, como también se observa en la ecuación (14). La dependencia de la $\varphi_E - \varphi_T$, de la fracción fetal medida puede analizarse teniendo en cuenta la fracción fetal, en algunas implementaciones. A menudo, la fracción fetal puede tenerse en cuenta suponiendo que la fracción fetal medida F_0 es igual a la fracción fetal verdadera F_v. En algunas implementaciones, si el cariotipo de la muestra es euploide, S_{fy} y S_{ff} tienen los mismos valores (por ejemplo, excepto los errores experimentales). Como resultado, la diferencia entre las dos sumas de residuos al cuadrado se reduce a menudo a:

$$\varphi_E - \varphi_T = \frac{-F_0^2S_{ff}}{(\Delta F)^2(1+S_{ff})} \qquad \text{(Euploides)} \tag{47}$$

En determinadas implementaciones, si el cariotipo de la muestra es triploide, las ecuaciones (41) y (42) pueden combinarse con la ecuación (46), produciendo:

$$\varphi_E - \varphi_T = \frac{F_0^2S_{ff}}{(\Delta F)^2} \qquad \text{(Triploides)} \tag{48}$$

Por tanto, si la diferencia de $\varphi_E - \varphi_T$, es positiva, el feto es triploide, en algunas implementaciones, y en determinadas implementaciones, si la diferencia es negativa, el feto no se ve afectado. La representación gráfica del resultado positivo o negativo frecuentemente es una parábola; cóncava para triploides y convexa para euploides. Ambas ramas tienden hacia cero a medida que disminuye F_0, teniendo el error experimental poco efecto sobre la forma del gráfico. Ninguna rama tiene un término sustancialmente lineal o libre, pero los coeficientes de segundo orden difieren en tamaño además de tener signos diferentes, en muchos casos. Con ΔF de aproximadamente el 2 %, el valor del término S_{ff} es próximo a 3,7, usando los recuentos de referencia e incertidumbres extraídos del conjunto euploide 86 (véase la Fig. 45).

En el ejemplo mostrado en la Fig. 45, las dos ramas son a menudo asimétricas debido a los diferentes coeficientes que multiplican el cuadrado de fracción fetal medida en las ecuaciones (47) y (48). La rama triploide (por ejemplo, positiva) aumenta relativamente rápido, volviéndose distinguible de cero sustancialmente antes que la rama euploide. La Fig. 46, obtenida usando un conjunto de datos reales, confirma los resultados cualitativos que se muestran en la Fig. 45. En la Fig. 46 el punto de color gris oscuro solitario en el cuarto cuadrante (por ejemplo, cuadrante intermedio inferior) es un gemelo afectado. En el conjunto de datos usado para generar la Fig. 46, ambas ramas euploide y T21 del gráfico muestran curvatura porque ambas muestran dependencia cuadrática de F_0 a partir de la versión para trisomía de la ecuación (31)

En algunas implementaciones, ambas ramas del gráfico pueden linealizarse para facilitar la inspección visual. El valor de la linealización se acondiciona a menudo en el análisis de propagación de errores. Los resultados presentados en las Fig. 45 y 46 se basaron en la suposición de que el error en las fracciones fetales medidas es uniforme en todo el rango de fracciones fetales. Sin embargo, la suposición no siempre es correcta. En algunos casos, la suposición más realista, basada en una relación lineal entre el error ΔF y la fracción fetal medida $F_0(\Delta F = 2/3 + F_0/6)$, produce los resultados presentados en la Fig. 47. En la Fig. 47, la rama euploide es prácticamente plana, casi constante (por ejemplo, el carácter parabólico se pierde sustancialmente), sin embargo, la rama de trisomía permanece parabólica. Los tres puntos de color gris claro intercalados en los puntos de color gris oscuro de la rama de trisomía representan datos de gemelos. Los datos de gemelos a veces se elevan con relación al modelo de error fijo.

La clasificación de si una muestra se ve afectada o no por una variación genética se lleva a cabo a menudo usando uno de tres procedimientos: (1) clasificación basada en las diferencias parabólicas de los cuadrados de residuos sumados, (véanse las Fig. 45 y 46), (2) clasificación basada en diferencias lineales de cuadrados de residuos sumados, (véanse las Fig. 47 y 48), y (3) clasificación basada en la fracción fetal ajustada (véase la ecuación (33)). En algunas implementaciones, el enfoque elegido toma en cuenta la propagación de errores.

Ploidía fija, fracción fetal optimizada: desviación de referencia de errores sistemáticos

Idealmente, los recuentos de bins medidos y de referencia deben contener cero error sistemático (por ejemplo, desviación); sin embargo, en la práctica, los recuentos de bins medidos y de referencia a veces se desplazan uno con respecto a otros. En algunos casos, el efecto del desplazamiento uno con respecto a otros puede analizarse usando la ecuación (33), suponiendo que el desplazamiento Δ es constante a través del cromosoma de interés. Para casos de euploides, si se ignoran errores aleatorios, se mantienen las siguientes relaciones, en algunas implementaciones:

$$f_i = f_i^0 + \Delta \quad (49)$$

$$y_i = f_i^0 = f_i - \Delta \quad (50)$$

f_i^0 representa el recuento verdadero de bins de referencia i y f_i representa los recuentos de bins de referencia usados, incluidos cualquier error sistemático Δ. En determinadas implementaciones, la sustitución de las ecuaciones (49) y (50) en la ecuación (33) genera la siguiente expresión para la rama euploide del gráfico de fracción fetal ajustada:

$$F_E = \frac{F_0 + \frac{(\Delta F)^2}{2}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left(f_i y_i - f_i^2\right)}{1 + \frac{(\Delta F)^2}{4}\sum_{i=1}^{N} f_i^2/\sigma_i^2} = \frac{F_0 + \frac{(\Delta F)^2}{2}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left[\left(f_i^0+\Delta\right)f_i^0 - \left(f_i^0+\Delta\right)^2\right]}{1 + \frac{(\Delta F)^2}{4}\sum_{i=1}^{N}\left(f_i^0+\Delta\right)^2/\sigma_i^2}$$

$$= \frac{F_0 - \frac{(\Delta F)^2}{2}\left(\Delta\sum_{i=1}^{N}\frac{f_i^0}{\sigma_i^2} + \Delta^2\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\right)}{1 + \frac{(\Delta F)^2}{4}\left(\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left(f_i^0\right)^2 + 2\Delta\sum_{i=1}^{N}\frac{f_i^0}{\sigma_i^2} + \Delta^2\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\right)} = \frac{F_0 - 2S_f^0\Delta - 2S_0^0\Delta^2}{1 + S_{ff}^0 + 2S_f^0\Delta + S_0^0\Delta^2} \quad (51)$$

Los coeficientes S_0^0, S_f^0 y S_{ff}^0, se generan a partir de las ecuaciones (33) a (39) al reemplazar f_i por f_i^0, en algunas implementaciones. En determinadas implementaciones, la inversa de la pendiente de la relación funcional lineal entre el valor euploide ajustado F_E y F_0 medida es igual a $1 + S_{ff}^0 + 2\,S_f^0\,\Delta + S_0^0\,\Delta^2$, lo que a menudo permite calcular el error sistemático Δ al resolver una ecuación cuadrática relativamente simple. Para triploides, suponiendo que F_0 es igual a F_v, los recuentos de bins medidos a veces se convierten en:

$$y_i = f_i^0 + \frac{1}{2}F_0 f_i^0 \quad (52)$$

La combinación de las ecuaciones (52), (49) y (33) genera la siguiente expresión para la rama triploide del gráfico de fracción fetal ajustada:

$$F_T = \frac{F_0 + \frac{(\Delta F)^2}{2}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left(f_i y_i - f_i^2\right)}{1 + \frac{(\Delta F)^2}{4}\sum_{i=1}^{N} f_i^2/\sigma_i^2} = \frac{F_0 + \frac{(\Delta F)^2}{2}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left[\left(f_i^0+\Delta\right)\left(f_i^0+\frac{1}{2}F_0 f_i^0\right) - \left(f_i^0+\Delta\right)^2\right]}{1 + \frac{(\Delta F)^2}{4}\sum_{i=1}^{N}\left(f_i^0+\Delta\right)^2/\sigma_i^2}$$

$$= \frac{F_0 + \frac{(\Delta F)^2}{2}\left(\frac{1}{2}F_0\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left(f_i^0\right)^2 + \frac{1}{2}F_0\Delta\sum_{i=1}^{N}\frac{f_i^0}{\sigma_i^2} - \Delta\sum_{i=1}^{N}\frac{f_i^0}{\sigma_i^2} - \Delta^2\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\right)}{1 + \frac{(\Delta F)^2}{4}\left(\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left(f_i^0\right)^2 + 2\Delta\sum_{i=1}^{N}\frac{f_i^0}{\sigma_i^2} + \Delta^2\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\right)} = \frac{F_0\left(1 + S_{ff}^0 + S_f^0\Delta\right) - S_f^0\Delta - S_0^0\Delta^2}{1 + S_{ff}^0 + 2S_f^0\Delta + S_0^0\Delta^2}$$

(53)

En algunas implementaciones, las ecuaciones (51) y (53) predicen que las fracciones fetales triploides y euploides ajustadas se comportarán tal como se muestra en la Fig. 48. En la Fig. 48 las líneas de color negro (por ejemplo, las líneas superiores en cada conjunto de 3 líneas) corresponden a la desviación negativa *Δ,* las líneas de color gris oscuro (por ejemplo, las líneas inferiores en cada conjunto de 3 líneas) corresponden a la desviación positiva *Δ*, y las líneas de color gris claro (por ejemplo, las líneas intermedias en cada conjunto de 3 líneas), corresponden a la ausencia de desviación. La Fig. 49 ilustra los efectos de errores sistemáticos simulados *Δ* impuestos artificialmente sobre datos reales.

La Fig. 50 ilustra la dependencia de fracción fetal ajustada en la desviación de errores sistemáticos para los conjuntos de datos euploides y triploides. Para ambos casos euploide y triploide, las expresiones teóricas de las ecuaciones (51) y (53) capturan a menudo la dependencia cualitativa de fracción fetal ajustada en la fracción fetal medida y en la desviación de error sistemático. Los coeficientes usados para los gráficos en las Fig. 49 y 50 se obtuvieron a partir de recuentos de bins de referencia sin procesar, sin eliminar ningún sesgo sistemático potencial.

Ploidía fija, fracción fetal optimizada, propagación de errores: fracción fetal ajustada

Las contribuciones a los errores en las fracciones fetales ajustadas se encuentran a menudo en uno de dos tipos de errores: 1) a partir de fracciones fetales medidas, y 2) a partir de recuentos de bins de referencia y medidos. Los dos tipos de errores se analizarán por separado, usando diferentes enfoques, y después se combinarán para generar los rangos de error finales. Los errores propagados a partir de la medición de fracciones fetales pueden evaluarse al reemplazar F_0 en la ecuación (40) primero por $F_0 - 2\Delta F$ (por ejemplo, para el límite de error inferior) y después por $F_0 + 2\Delta F$ (por ejemplo, para el límite de error superior). Este enfoque relativamente simple produce un comportamiento cualitativo correcto a intervalos de confianza del 95 %, en determinadas implementaciones. Para un nivel de confianza deseado diferente, puede usarse un par de límites más generales, $F_0 - n\Delta F$ y $F_0 + n\Delta F$. Los términos usados para generar límites de error superior e inferior a veces subestima el error total debido a que las contribuciones por errores en la medida y los recuentos de bins de referencia a menudo se desprecian.

Para evaluar mejor la contribución de los recuentos de bins de referencia y medidos sobre error en la fracción fetal ajustada, pueden usarse las ecuaciones (38) a (40), en algunas implementaciones. En determinadas implementaciones, la ecuación (33) puede expandirse para la fracción fetal ajustada en una serie Taylor con respecto a f_i e y_i, truncada en el primer orden, al cuadrado y promedio. En algunos casos, puede suponerse que las incertidumbres en y_i son a menudo iguales que las incertidumbres en f_i. Para simplificar el análisis, se supone que los términos cruzados y términos de orden superior se reducen a cero al promediarse. Los coeficientes de expansión de Taylor se obtienen a menudo usando la regla de cadena. La variación media al cuadrado en la fracción fetal ajustada viene dada entonces por la ecuación (54) que se muestra a continuación. El modelo representado por la ecuación ignora las contribuciones de las estimaciones para *ΔF,* en algunas implementaciones. Las derivadas parciales pueden evaluarse usando las expresiones presentadas a continuación, ecuación (54).

$$(\delta F)^2 = \sum_{i=1}^{N}\left(\frac{\partial F}{\partial f_i}\right)^2\sigma_i^2 + \sum_{i=1}^{N}\left(\frac{\partial F}{\partial y_i}\right)^2\sigma_i^2$$

$$= \sum_{i=1}^{N}\left[\left(\frac{\partial F}{\partial S_{ff}}\right)\left(\frac{\partial S_{ff}}{\partial f_i}\right) + \left(\frac{\partial F}{\partial S_{fy}}\right)\left(\frac{\partial S_{fy}}{\partial f_i}\right)\right]^2\sigma_i^2 + \sum_{i=1}^{N}\left[\left(\frac{\partial F}{\partial S_{fy}}\right)\left(\frac{\partial S_{fy}}{\partial y_i}\right)\right]^2\sigma_i^2 \Big|$$

(54)

$$\left(\frac{\partial F}{\partial S_{ff}}\right) = -\frac{F_0 + 2S_{fy} + 2}{\left(1 + S_{ff}\right)^2}$$

(55)

$$\left(\frac{\partial F}{\partial S_{fy}}\right) = \frac{2}{1+S_{ff}} \quad (56)$$

$$\left(\frac{\partial S_{ff}}{\partial f_i}\right) = \frac{(\Delta F)^2}{2}\left(\frac{f_i}{\sigma_i^2}\right) \quad (57)$$

$$\left(\frac{\partial S_{fy}}{\partial f_i}\right) = \frac{(\Delta F)^2}{4}\left(\frac{y_i}{\sigma_i^2}\right) \quad (58)$$

$$\left(\frac{\partial S_{fy}}{\partial y_i}\right) = \frac{(\Delta F)^2}{4}\left(\frac{f_i}{\sigma_i^2}\right) \quad (59)$$

La combinación de las ecuaciones (54) a (59) genera la siguiente expresión:

$$(\delta F)^2 = \left[\frac{(\Delta F)^2}{4}\right]^2 \left\{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}\left[\frac{2y_i}{1+S_{ff}} - 2f_i\frac{F_0+2S_{fy}+2}{(1+S_{ff})^2}\right]^2 + \sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left(\frac{2f_i}{1+S_{ff}}\right)^2\right\}$$

$$= \left[\frac{(\Delta F)^2}{4}\right]^2 \sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left[\left(\frac{2y_i}{1+S_{ff}}\right)^2 - 8f_i y_i\frac{F_0+2S_{fy}+2}{(1+S_{ff})^3} + 4f_i^2\frac{(F_0+2S_{fy}+2)^2}{(1+S_{ff})^4} + \left(\frac{2f_i}{1+S_{ff}}\right)^2\right]$$

$$= \left[\frac{(\Delta F)^2}{4}\right]^2 \left\{\frac{4}{(1+S_{ff})^2}\sum_{i=1}^{N}\frac{y_i^2}{\sigma_i^2} - 8\frac{F_0+2S_{fy}+2}{(1+S_{ff})^3}\sum_{i=1}^{N}\frac{f_i y_i}{\sigma_i^2} + 4\left[\frac{(F_0+2S_{fy}+2)^2}{(1+S_{ff})^4} + \frac{1}{(1+S_{ff})^2}\right]\sum_{i=1}^{N}\frac{f_i^2}{\sigma_i^2}\right\}$$

$$= (\Delta F)^2\left\{\frac{S_{yy}}{(1+S_{ff})^2} - 2S_{fy}\frac{F_0+2S_{fy}+2}{(1+S_{ff})^3} + S_{ff}\left[\frac{(F_0+2S_{fy}+2)^2}{(1+S_{ff})^4} + \frac{1}{(1+S_{ff})^2}\right]\right\}$$

(60)

Para evaluar la ecuación (60) a un intervalo de confianza del 95 %, pueden usarse los siguientes límites superior e inferior, en algunas implementaciones:

$$\begin{bmatrix} F_{Lower} \\ F_{Upper} \end{bmatrix} =$$

$$\frac{F_0+2S_{fy}-2S_{ff}}{1+S_{ff}} + \begin{bmatrix} -2 \\ 2 \end{bmatrix}\Delta F\left\{\frac{1}{1+S_{ff}} + \sqrt{\frac{S_{yy}}{(1+S_{ff})^2} - 2S_{fy}\frac{F_0+2S_{fy}+2}{(1+S_{ff})^3} + S_{ff}\left[\frac{(F_0+2S_{fy}+2)^2}{(1+S_{ff})^4} + \frac{1}{(1+S_{ff})^2}\right]}\right\}$$

(61)

En implementaciones en las que prácticamente todas las fuentes de error posibles (por ejemplo, F_0, f_i, y_i) se incluyen en la serie de expansión de Taylor, a menudo se obtiene la misma ecuación. En algunos casos, la dependencia de F de F_0, puede tenerse en cuenta hasta S_{fy}. En algunas implementaciones, los términos de la serie de potencia que corresponden a F_0 adoptan a menudo la forma;

$\left[\left(\frac{\partial F}{\partial F_0}\right)+\left(\frac{\partial F}{\partial S_{fy}}\right)\left(\frac{\partial S_{fy}}{\partial F_0}\right)\right]^2 (\Delta F)^2$, pero $\left[\left(\frac{\partial F}{\partial F_0}\right)+\left(\frac{\partial F}{\partial S_{fy}}\right)\left(\frac{\partial S_{fy}}{\partial F_0}\right)\right]^2$ es igual a 1 para triploides. Por tanto, a menudo, se justifica la relativamente simple resta y suma de ΔF a F_0, aunque ΔF aumenta a menudo con F_0 y se vuelve grande a alta F_0.

El resultado se debe tanto a F como a depender linealmente de F_0, en algunas implementaciones. Las simulaciones basadas en la ecuación (61) se muestran en la Fig. 51, junto con las fracciones fetales ajustadas obtenidas de los datos derivados de sujetos de prueba. En las simulaciones presentadas en la Fig. 51, $\Delta F = 2/3 + F_0/6$, tal como se describe en el presente documento.

Ejemplo 3: Análisis de ventana deslizante y sumas acumulativas en función de la posición genómica

La identificación de características reconocibles (por ejemplo, regiones de variación genética, regiones de variación del número de copias) en un perfil de recuento normalizado a veces es un procedimiento relativamente costoso y/o relativamente lento. El procedimiento para identificar características reconocibles se complica a menudo por los conjuntos de datos que contienen datos con ruido y/o baja contribución de ácido nucleico fetal. La identificación de características reconocibles que representan verdaderas variaciones genéticas o variaciones del número de copias puede ayudar a evitar buscar regiones grandes y sin características de un genoma. La identificación de las características reconocibles puede lograrse eliminando secciones genómicas altamente variables de un conjunto de datos que se busca y obtener, de las secciones genómicas restantes, puntos de datos que se desvían de la elevación de perfil media en un múltiplo predeterminado de la varianza del perfil.

En algunas implementaciones, la obtención de puntos de datos que se desvían de la elevación de perfil media en un múltiplo predeterminado de la varianza del perfil puede usarse para reducir el número de secciones genómicas candidatas de más de 50.000 o 100.000 secciones genómicas candidatas al rango de aproximadamente 100 a aproximadamente 1000 secciones genómicas candidatas que representan señales verdaderas o aumentos bruscos de ruido solitarios (por ejemplo, aproximadamente 100 secciones genómicas, aproximadamente 200 secciones genómicas, aproximadamente 300 secciones genómicas, aproximadamente 400 secciones genómicas, aproximadamente 500 secciones genómicas, aproximadamente 600 secciones genómicas, aproximadamente 700 secciones genómicas, aproximadamente 800 secciones genómicas, aproximadamente 900 secciones genómicas o aproximadamente 1000 secciones genómicas). La reducción del número de secciones genómicas candidatas puede lograrse de manera relativamente rápida y fácil y a menudo acelera la búsqueda y/o identificación de aberraciones genéticas en dos o más órdenes de magnitud. La reducción del número de secciones genómicas buscadas para determinar la presencia o ausencia de regiones candidatas de variación genómica reduce a menudo la complejidad y/o dimensionalidad de un conjunto de datos.

Después de generar un conjunto de datos reducido que contiene puntos de datos que se desvían de la elevación de perfil media en un múltiplo predeterminado de la varianza del perfil, el conjunto de datos reducido se filtra para eliminar los aumentos bruscos de ruido solitarios, en algunas implementaciones. Filtrar un conjunto de datos reducido para eliminar los aumentos bruscos de ruido solitarios genera a menudo un conjunto de datos reducido filtrado. En algunas implementaciones, un conjunto de datos reducido y filtrado retiene agrupaciones contiguas de puntos de datos y, en determinadas implementaciones, un conjunto de datos reducido y filtrado retiene agrupaciones de puntos de datos que son en gran medida contiguas con permiso para un número y/o tamaño predeterminados de huecos. Los puntos de datos del conjunto de datos reducido y filtrado que se desvían de la elevación de perfil promedio en prácticamente la misma dirección se agrupan juntos, en algunas implementaciones.

Debido al ruido de fondo presente a menudo en las muestras de ácido nucleico (por ejemplo, razón de regiones de interés en comparación con el ácido nucleico total en una muestra), distinguir las regiones de variación genética o aberración genética del ruido de fondo es a menudo un desafío. Los métodos que mejoran la relación señal-ruido a menudo son útiles para facilitar la identificación de regiones candidatas representativas de regiones de variación genética y/o aberración genética verdadera. Puede usarse cualquier método que mejore la relación señal-ruido de regiones de variación genética verdadera con respecto al ruido de fondo genómico. Un ejemplo no limitativo de un método adecuado para usar en la mejora de la relación señal-ruido de regiones de variación genética verdadera con respecto al ruido de fondo genómico es el uso de integrales sobre la presunta aberración y sus alrededores inmediatos. En algunas implementaciones, el uso de integrales sobre la presunta aberración y sus alrededores inmediatos es beneficioso, porque la suma cancela el ruido aleatorio. Después de que se ha reducido o eliminado el ruido, incluso señales relativamente menores pueden detectarse fácilmente usando una suma acumulativa del pico candidato y sus alrededores, en algunas implementaciones. Algunas veces, se define una suma acumulativa con respecto a un origen elegido arbitrariamente fuera (por ejemplo, en un lado o en el otro) del pico. Una suma acumulativa es a menudo una estimación numérica de la integral del perfil de recuento normalizado sobre la sección o secciones genéticas seleccionadas.

En ausencia de aberraciones, la suma acumulativa en función de la posición genómica se comporta a menudo como una línea recta con pendiente unitaria (por ejemplo, pendiente igual a 1). Si están presentes deleciones o duplicaciones, el perfil de suma acumulativa consiste a menudo en dos o más segmentos lineales. En algunas implementaciones, las

áreas fuera de las aberraciones se mapean en segmentos de línea con pendientes unitarias. Para áreas dentro de aberraciones, los segmentos lineales están conectados por otros segmentos lineales cuyas pendientes son iguales a la elevación o depresión del perfil de recuento dentro de la aberración, en determinadas implementaciones.

En aquellas muestras que tienen aberraciones maternas, las pendientes (por ejemplo, equivalentes a la elevación de perfil de recuento) se determinan de manera relativamente fácil: 0 para deleciones maternas homocigotas, 0,5 para deleciones maternas heterocigotas, 1,5 para duplicaciones heterocigotas, 2,0 para duplicaciones homocigotas. En aquellas muestras que tienen aberraciones fetales, las pendientes reales dependen tanto del tipo de aberración (por ejemplo, deleción homocigota, deleción heterocigota, duplicación homocigota o duplicación heterocigota) como de la fracción fetal. En algunas implementaciones, también se tiene en cuenta la herencia de una aberración materna por el feto cuando se evalúan las muestras fetales para determinar las variaciones genéticas.

En algunas implementaciones, los segmentos lineales con pendientes unitarias, que corresponden a áreas genómicas normales a la izquierda y a la derecha de una aberración, se desplazan verticalmente uno con respecto a otros. La diferencia (por ejemplo, resultado sustractivo) entre sus ordenadas en el origen es igual al producto entre la anchura de la aberración (número de secciones genómicas afectadas) y el nivel de aberración (por ejemplo, -1 para la deleción materna homocigota, -0,5 para la deleción materna heterocigota, +0,5 para la duplicación heterocigota, +1 para la duplicación homocigota, y similares). Véanse las Figs. 52-61F para ejemplos de conjuntos de datos procesados usando sumas acumulativas en función de la posición genómica (por ejemplo, análisis de ventana deslizante).

Ejemplo 4: Eliminación del error parametrizado y normalización no sesgada (PERUN)

Variabilidad de los recuentos medidos

Idealmente, la elevación cromosómica medida es una línea horizontal recta con una elevación de 1 para los euploides, como en la Fig. 62. Para los embarazos con trisomía, el comportamiento deseado de la elevación cromosómica medida es una función gradual, con la desviación de 1 proporcional a la fracción fetal, tal como se simula en la Fig. 63 para la fracción fetal igual al 15 %. Surgen excepciones de las deleciones/duplicaciones maternas, que se reconocen y distinguen fácilmente de las anomalías fetales basándose en sus magnitudes, que son múltiplos de la mitad.

Lo que realmente se midió no fue ideal. La Fig. 64 muestra recuentos sin procesar superpuestos para los cromosomas 20, 21 y 22 recopilados de 1093 embarazos euploides y la Fig. 65 muestra recuentos sin procesar superpuestos para los cromosomas 20, 21 y 22 recopilados de 134 embarazos con trisomía 21. La inspección visual de los dos conjuntos de perfiles no logró confirmar que las trazas del cromosoma 21 en casos de trisomía estaban elevadas.

Tanto el ruido estocástico como el sesgo sistemático hicieron que la elevación del cromosoma 21 fuese difícil de visualizar. Además, el segmento derecho más lejano del cromosoma 21 sugirió incorrectamente que las trazas del cromosoma euploide 21 estaban elevadas, en lugar de los perfiles de trisomía. Una gran parte del sesgo sistemático se originó a partir del contenido de GC asociado con una región genómica particular.

Los intentos de eliminar el sesgo sistemático debido al contenido de GC incluyeron suavizado de GC con LOESS multiplicativa, enmascaramiento de repeticiones (RM), combinación de LOESS y RM (GCRM), y otros, tales como cQN. La Fig. 66 muestra los resultados de un procedimiento de GCRM tal como se aplica a 1093 trazas euploides y la Fig. 67 muestra los perfiles de GCRM para 134 casos de trisomía. El GCRM aplanó exitosamente el segmento más a la derecha, rico en GC, elevado del cromosoma 21 en euploides. Sin embargo, el procedimiento aumentó evidentemente el ruido estocástico general. Además, creó un nuevo sesgo sistemático, ausente de las mediciones sin procesar (región más a la izquierda del cromosoma 20 (cr20)). Las mejoras que se debieron a GCRM se compensaron por el aumento del ruido y el sesgo, lo que hizo cuestionable la utilidad del procedimiento. La pequeña elevación del cromosoma 21 tal como se observa en la Fig. 63 se perdió en el alto ruido tal como se muestra en la Fig. 66 y la Fig. 67.

PERUN (eliminación del error parametrizado y normalización no sesgada) se desarrolló como una alternativa viable a los métodos de normalización de GC descritos anteriormente. La Fig. 68 y la Fig. 69 contrastan los resultados del método PERUN frente a los presentados en las Fig. 64 a 67. Los resultados de PERUN se obtuvieron en las mismas dos subpoblaciones de datos que se analizaron en las Fig. 64 a 67. La mayor parte del sesgo sistemático no tenía trazas de PERUN, dejando solamente ruido estocástico y variación biológica, tal como la deleción prominente en el cromosoma 20 de una de las muestras euploides (Fig. 68). La deleción del cromosoma 20 también se observó en los perfiles de recuento sin procesar (Fig. 64), pero completamente enmascarado en las trazas de GCRM. La incapacidad del GCRM para revelar esta enorme desviación lo descalifica claramente con el propósito de medir las elevaciones minúsculas de T21 fetal. Las trazas de PERUN contienen menos bins que los perfiles sin procesar o GCRM. Tal como se muestra en las Fig. 62-63, los resultados de PERUN parecen al menos tan buenos como lo permiten los errores de medición.

Normalización con respecto a la mediana de perfil de recuento de referencia

Los procedimientos convencionales de normalización de GC pueden realizarse de manera subóptima. Una parte de la razón ha sido que el sesgo de GC no es la única fuente de variación. Un diagrama apilado de muchos perfiles individuales

de recuento sin procesar reveló paralelismo entre muestras diferentes. Aunque algunas regiones genómicas estaban representadas de manera sistemática, otras estaban subrepresentadas de manera sistemática, tal como se ilustra por las trazas de un estudio de 480v2 (Fig. 6). Aunque el sesgo de GC varió de una muestra a otra, el sesgo sistemático específico de bins observado en estos perfiles siguió el mismo patrón para todas las muestras. Todos los perfiles en la Fig. 6 tenían forma en zigzag de manera coordinada. Las únicas excepciones fueron las porciones intermedias de las dos muestras inferiores, que resultaron originarse de deleciones maternas. Para corregir este sesgo específico de bins, se usó una mediana de perfil de referencia. La mediana de perfil de referencia se construyó a partir de un conjunto de euploides conocidos (por ejemplo, embarazos euploides) o a partir de todas las muestras en una celda de flujo. El procedimiento generó el perfil de referencia al evaluar la mediana de recuentos por bin para un conjunto de muestras de referencia. La D.M.A. asociada con un bin midió la fiabilidad de un bin. Los bins muy variables y bins que tienen de manera sistemática representaciones de fuga se eliminaron del análisis posterior (Fig. 4). Después, los recuentos medidos en un conjunto de datos de prueba se normalizaron con respecto a la mediana de perfil de referencia, tal como se ilustra en la Fig. 8. Los bins altamente variables se eliminan del perfil normalizado, dejando una traza que es de aproximadamente 1 en las secciones diploides, 1,5 en las regiones de duplicación heterocigota, 0,5 en las áreas de deleción heterocigota, etcétera (Fig. 9). Los perfiles normalizados resultantes redujeron razonablemente la variabilidad, lo que permite la detección de deleciones y duplicaciones maternas y el rastreo de las identidades de las muestras (Fig. 12, 22, 13, 11).

La normalización basada en la mediana de perfil de recuento puede aclarar los resultados, pero el sesgo de GC todavía tiene un efecto negativo sobre tales métodos. Los métodos PERUN descritos en el presente documento pueden usarse para abordar el sesgo de GC y proporcionar resultados con mayor sensibilidad y especificidad.

Efectos perjudiciales de la corrección de LOESS multiplicativa

La Fig. 11. ilustró por qué los recuentos de bins fluctúan más después de la aplicación de LOESS de GC o GCRM (Fig. 66-67) que antes (Fig. 64-65). La corrección de LOESS de GC eliminó la tendencia de los recuentos sin procesar (Fig. 70, panel superior) al dividir los recuentos sin procesar con la línea de regresión (línea recta, Fig. 70, panel superior). El punto definido por la mediana de recuentos y la mediana del contenido de GC del genoma se mantuvo inmóvil. En promedio, los recuentos por debajo de la mediana de recuento se dividieron entre números pequeños, mientras que los recuentos que superan la mediana de recuento se dividieron entre números grandes. En cualquier caso, en promedio, los recuentos se aumentaron de escala o se redujeron para coincidir con 1 (Fig. 70, panel inferior). El escalado de recuentos pequeños, además de inflar los recuentos, también infló su variabilidad. El resultado final (Fig. 70, panel inferior) a la izquierda de la mediana del contenido genómico de GC presentó una mayor dispersión que los recuentos sin procesar correspondientes (Fig. 70, panel superior), que conforma la forma triangular típica (Fig. 70, panel inferior, triángulo). Para invertir la tendencia de los recuentos, LOESS de GC/GCRM sacrificó precisión ya que tales procedimientos correctores son generalmente multiplicativos y no aditivos. La normalización proporcionada por PERUN es generalmente de naturaleza aditiva y mejora la precisión con respecto a técnicas multiplicativas.

Inadecuación de un pivote de genoma completo para escalamiento de sesgo de GC

Un enfoque alternativo aplicó la corrección de LOESS por separado a cromosomas individuales en lugar de someter todo el genoma a escalamiento de sesgo de GC colectivo. El escalamiento de cromosomas individuales no fue práctico con el propósito de clasificar muestras como euploides o con trisomía porque canceló la señal de cromosomas sobrerrepresentados. Sin embargo, las conclusiones de este estudio fueron útiles, eventualmente, como catalizadores para desarrollar el algoritmo PERUN. La Fig. 71 ilustra el hecho de que las curvas de LOESS obtenidas para el mismo cromosoma de múltiples muestras comparten una intersección (pivote) común.

La Fig. 72 demostró que inclinar las curvas de LOESS específicas de cromosoma alrededor del pivote en un ángulo proporcional a los coeficientes de sesgo de GC medidos en esas muestras provocó que todas las curvas coalescieran. La inclinación de las curvas de LOESS específicas de cromosoma por los coeficientes de sesgo de GC específicos de muestra redujo significativamente la dispersión de la familia de curvas de LOESS obtenidas para múltiples muestras, tal como se muestra en la Fig. 73 (línea de color negro con forma de V (antes de la inclinación) y línea de fondo de color gris (después de la inclinación)). El punto en el que las curvas de color negro y de color gris se tocaron coincidió con el pivote. Además, resultó evidente que la ubicación en el eje de contenido de GC del pivote específico de cromosoma coincidía con la mediana del contenido de GC del cromosoma dado (Fig. 74, línea de color gris vertical izquierda: mediana, línea en negrita vertical derecha: media). Se obtuvieron resultados similares para todos los cromosomas, tal como se muestra en la Fig. 75A a la Fig. 75F (línea de color gris vertical izquierda: mediana, línea en negrita vertical derecha: media). Todos los autosomas y el cromosoma X se ordenaron según su mediana de contenido de GC.

El escalamiento de LOESS de GC de genoma completo pivotó la transformación en la mediana del contenido medio de GC de genoma completo, tal como se muestra en la Fig. 76. Ese pivote fue aceptable para los cromosomas que tienen una mediana de contenido de GC similar al contenido de GC de todo el genoma, pero se volvió subóptimo para los cromosomas con contenidos de GC extremos, tales como los cromosomas 19, 20, 17 y 16 (contenido de GC extremadamente alto). El pivotado de esos cromosomas centrados en la mediana del contenido de GC de todo el genoma mantuvo la dispersión observada dentro del recuadro izquierdo en la Fig. 76, faltando la región de baja variabilidad encerrada por el recuadro derecho en la Fig. 76 (el pivote específico de cromosoma).

El pivotado en la mediana del contenido medio de GC específico de cromosoma, sin embargo, redujo significativamente la variabilidad (Fig. 75). Se realizaron las siguientes observaciones:

1) La corrección de GC debe realizarse en secciones o segmentos genómicos pequeños, en lugar de en todo el genoma, para reducir la variabilidad. Cuanto menor sea la sección o el segmento, mayor será la corrección de GC enfocada, minimizando el error residual.

2) En este caso particular, esas secciones o segmentos genómicos pequeños son idénticos a los cromosomas. En principio, el concepto es más general: las secciones o los segmentos podrían ser cualquier región genómica, incluyendo bins de 50 kpb.

3) El sesgo de GC dentro de las regiones genómicas individuales puede rectificarse usando el coeficiente de GC de todo el genoma específico de muestra evaluado para todo el genoma. Este concepto es importante: aunque algunos descriptores de las secciones genómicas (tales como la ubicación del punto de pivote, la distribución del contenido de GC, la mediana del contenido de GC, la forma de la curva de LOESS, etcétera) son específicos para cada sección e independientes de la muestra, el valor del coeficiente de GC usado para rectificar el sesgo es el mismo para todas las secciones y diferente para cada muestra.

Estas conclusiones generales guiaron el desarrollo de PERUN, tal como resultará evidente a partir de la descripción detallada de sus procedimientos.

Separabilidad de fuentes de sesgo sistemático

La inspección cuidadosa de una multitud de perfiles de recuento sin procesar medidos usando diferentes químicas de preparación de bibliotecas, entornos de agrupamiento, tecnologías de secuenciación y cohortes de muestras confirmaron de manera sistemática la existencia de al menos dos fuentes independientes de variabilidad sistemática:

1) el sesgo específico de muestra basado en el contenido de GC, que afecta a todos los bins dentro de una muestra dada de la misma manera, variando de una muestra a otra, y

2) patrón de atenuación específico de bins común a todas las muestras.

Las dos fuentes de variabilidad se entremezclan en los datos. La eliminación completa de ambos requirió su desconvolución. Las deficiencias de los procedimientos de eliminación de errores que precedieron a PERUN provienen del hecho de que solo corrigen para una de las dos fuentes de sesgo sistemático, mientras se desprecia la otra.

Por ejemplo, el método de GCRM (o LOESS de GC) trató idénticamente todos los bins con valores de contenido de GC comprendidos dentro de un rango estrecho de contenido de GC. Los bins pertenecientes a ese subconjunto pueden caracterizarse por un amplio rango de elevaciones *intrínsecas* diferentes, tal como se refleja por la mediana de perfil de recuento de referencia. Sin embargo, GCRM era ciego a sus propiedades inherentes aparte de su contenido de GC. Por tanto, GCRM mantiene (o incluso amplía) la dispersión ya presente en el subconjunto de bins.

Por otra parte, la mediana de recuentos de referencia de bins no consideró la modulación del patrón de atenuación específico de bins por el sesgo de GC, manteniendo la dispersión provocada por el contenido variable de GC.

La aplicación secuencial de los métodos que tratan los extremos opuestos del espectro de error intenta sin éxito resolver los dos sesgos globalmente (genoma completo), ignorando la necesidad de disociar los dos sesgos en la elevación de bins. Sin desear restringirse por la teoría, PERUN aparentemente debe su éxito al hecho de que separa las dos fuentes de sesgo localmente, en la elevación de bins.

Eliminación de bins no informativos

Múltiples intentos para eliminar los bins no informativos han indicado que la selección de bins tiene el potencial de mejorar la clasificación. El primer enfoque de este tipo evaluó los recuentos medios del cromosoma 21, del cromosoma 18 y del cromosoma 13 por bin para todos los casos de trisomía 480v2 y lo comparó con los recuentos medios por bin para todos los euploides 480v2. La brecha entre los casos afectados y no afectados se ajustó a escala con la incertidumbre por bins combinada derivada de los recuentos de bins medidos en ambos grupos. El estadístico t resultante se usó para evaluar el perfil del valor de p por bins, que se muestra en la Fig. 77. En el caso del cromosoma 21, el procedimiento identificó 36 bins no informativos (panel central, etiquetado con la elipse en la Fig. 77). La eliminación de esos bins del cálculo de las puntuaciones Z aumentó notoriamente los valores de Z para los casos afectados, mientras que perturbó aleatoriamente las puntuaciones Z no afectadas (Fig. 78), aumentando así la brecha entre los casos euploides y de trisomía 21.

En el cromosoma 18, el procedimiento solamente mejoró las puntuaciones Z para dos casos afectados (Fig. 79).

Un análisis *post-hoc* mostró que la mejora de las puntuaciones Z en esas dos muestras resultó de la eliminación de la deleción materna grande en el cromosoma 18 (Fig. 11) y que las dos muestras provienen realmente del mismo paciente. Estas mejoras fueron específicas de muestra, sin potencia generalizadora. En el cromosoma 13, el procedimiento no condujo a ninguna mejora de las puntuaciones Z.

Un esquema alternativo de filtrado de bins elimina los bins con un contenido de GC extremadamente bajo o extremadamente alto. Este enfoque produjo resultados mixtos, con una varianza notablemente reducida en los cromosomas 9, 15, 16, 19 y 22 (dependiendo de los puntos de cortes), pero efectos adversos en los cromosomas 13 y 18.

Otro esquema simple de selección de bins elimina bins con recuentos sistemáticamente bajos. El procedimiento corrigió dos falsos negativos del cromosoma 18 LDTv2CE (Fig. 80) y dos falsos negativos del cromosoma 21 (Fig. 81). Además, corrigió al menos tres falsos positivos del cromosoma 18, pero creó al menos un nuevo falso positivo de cromosoma 18 (Fig. 80):

En conclusión, los diferentes criterios usados para filtrar los bins no informativos evidenciaron que el procesamiento de datos se beneficiará de la selección de bins basándose en con cuánta información útil contribuyen a la clasificación.

Separación de sesgo de GC del sesgo sistemático por bins

Para resolver y eliminar los diferentes sesgos sistemáticos encontrados en los recuentos medidos, el flujo de trabajo del procesamiento de datos necesitaba combinar óptimamente los procedimientos parciales descritos de la sección anterior titulada "Normalización con respecto a la mediana de perfil de recuento de referencia" a la sección titulada "Eliminación del sesgo no informativo". El primer paso es ordenar muestras diferentes según sus valores del coeficiente de sesgo de GC y luego apilar sus gráficos de recuentos frente al contenido de GC. El resultado es una superficie tridimensional que se retuerce como una hélice, mostrada esquemáticamente en la Fig. 82.

Dispuestas de ese modo, las mediciones sugieren que un conjunto de coeficientes de sesgo de GC específicos de muestra puede aplicarse para rectificar errores dentro de una sección o segmento genómico individual. En la Fig. 82, las secciones o los segmentos se definen por su contenido de GC. Una división alternativa del genoma proporciona bins contiguos, no solapantes. Las ubicaciones iniciales sucesivas de los bins cubren uniformemente el genoma. Para un bin de 50 kpb de largo, la Fig. 83 explora el comportamiento de los valores de recuento medidos dentro de ese bin para un conjunto de muestras. Los recuentos se representan gráficamente frente a los coeficientes de sesgo de GC observados en esas muestras. Los recuentos dentro del bin aumentan evidentemente de manera lineal con el sesgo de GC específico de muestra. El mismo patrón en observa en una gran mayoría de bins. Las observaciones pueden modelarse usando la relación lineal simple:

$$M = LI + GS \qquad \text{(A)}$$

Los diversos términos en la ec. A tiene los siguientes significados:

- ***M****:* recuentos medidos, que representan la información primaria contaminada por variación no deseada.
- ***L*** elevación cromosómica - este es el resultado deseado del procedimiento de procesamiento de datos. *L* indica aberraciones fetales y/o maternas de euploidía. Esta es la cantidad enmascarada tanto por errores estocásticos como por sesgos sistemáticos. La elevación cromosómica *L* es específica de muestra y específica de bin.
- ***G****:* Coeficiente de sesgo de GC medido usando un modelo lineal, LOESS o cualquier enfoque equivalente. *G* representa la información secundaria, extraída de M y de un conjunto de valores de contenido de GC específicos de bins, derivados habitualmente del genoma de referencia (pero también pueden derivarse del contenido de GC realmente observado). *G* es específica de la muestra y no varía a lo largo de la posición genómica. Encapsula una porción de la variación no deseada.
- ***I***. Ordenada en el origen del modelo lineal (línea de color verde en la Fig. 83). Este parámetro del modelo es fijo para una configuración experimental dada, independiente de la muestra, y específico de bin.
- ***S*****:** La pendiente del modelo lineal (línea de color verde en la Fig. 83). Este parámetro del modelo es fijo para una configuración experimental dada, independiente de la muestra, y específico de bin.

Se miden las cantidades *M* y *G*. Inicialmente, los valores específicos de bin *I* y S son desconocidos. Para evaluar *I* y S desconocidos, debemos suponer que *L* = 1 para todos los bins en muestras euploides. La suposición no siempre es verdadera, pero puede esperarse razonablemente que cualquier muestra con deleciones/duplicaciones esté repleta de muestras con elevaciones cromosómicas normales. Un modelo lineal aplicado a las muestras euploides extrae los valores de los parámetros *I* y S específicos para el bin seleccionado (suponiendo *L* = 1). El mismo procedimiento se aplica a todos los bins en el genoma humano, lo que produce un conjunto de ordenadas en el origen *I* y pendientes S para cada ubicación genómica. La validación cruzada selecciona aleatoriamente un conjunto de trabajo que contiene el 90 % de los euploides LDTv2CE y usa ese subconjunto para entrenar el modelo. La selección aleatoria se repite 100 veces,

produciendo un conjunto de 100 pendientes y 100 ordenadas en el origen para cada bin. La sección anterior titulada "Validación cruzada de parámetros de PERUN" describe el procedimiento de validación cruzada con mayor detalle.

Las Fig. 84-85 muestran 100 valores de ordenada en el origen y 100 valores de pendiente, respectivamente, evaluados para el bin n.° 2404 en el cromosoma 2. Las dos distribuciones corresponden a 100 subconjuntos diferentes del 90 % de 1093 euploides LDTv2CE mostrados en la Fig. 83. Ambas distribuciones son relativamente estrechas y de forma irregular. Sus dispersiones son similares a los errores en el coeficiente según lo notificado por el modelo lineal. Como norma, la pendiente es menos fiable que la ordenada en el origen porque menos muestras completan las secciones extremas del rango de sesgo de GC.

Interpretación de los parámetros I y S de PERUN

El significado de la ordenada en el origen *I* se ilustra en la Fig. 86. El gráfico correlaciona las ordenadas en el origen de bin estimadas con los datos extraídos de un conjunto de réplicas técnicas, obtenidos cuando una celda de flujo de LDTv2CE se sometió a tres ciclos de secuenciación independientes. El eje y contiene la mediana de valores de recuento de bins de esas tres mediciones. Estas medianas de valores se relacionan conceptualmente con la mediana de perfil de referencia, usada anteriormente para normalizar perfiles tal como se describe en la sección titulada "Normalización con respecto a la mediana de perfil de recuento de referencia". Las ordenadas en el origen por bins se representan gráficamente a lo largo del eje x. La correlación sorprendente entre las dos cantidades revela el significado verdadero de las ordenadas en el origen como los recuentos esperados por bin en ausencia de sesgo de GC. El problema con la mediana de perfil de recuento de referencia es que no tiene en cuenta el sesgo de GC (véase la sección titulada "Normalización con respecto a la mediana de perfil de recuento de referencia"). En PERUN, sin limitarse por la teoría, la tarea de una ordenada en el origen *I* es tratar la atenuación específica de bin, mientras que el sesgo GC se relega al otro parámetro del modelo, la pendiente S.

La Fig. 86 excluye el cromosoma Y de la correlación porque el conjunto de réplicas técnicas no refleja la población general de embarazos de sexo masculino.

La distribución de la pendiente S (Fig. 87) ilustra el significado de ese parámetro del modelo.

La unidad marcada entre la distribución de la Fig. 87 y la distribución del contenido de GC de genoma completo (Fig. 88) indica que la pendiente S se aproxima al contenido de GC de un bin, desplazado por la mediana del contenido de GC del cromosoma que lo contiene. La línea vertical delgada en la Fig. 88 marca la mediana del contenido de GC de todo el genoma.

La Fig. 89 reafirma la estrecha relación entre la pendiente S y el contenido de GC por bin. Aunque ligeramente curvada, la tendencia observada es extremadamente ajustada y sistemática, con solo un puñado de bins atípicos notables.

Extracción de la elevación cromosómica a partir de recuentos medidos

Suponiendo que los valores *I* y S de los parámetros del modelo están disponibles para cada bin, las mediciones *M* recopiladas en una nueva muestra de prueba se usan para evaluar la elevación cromosómica según la siguiente expresión:

$$L = (M-GS)/I \qquad \text{(B)}$$

Como en la ec. A, el coeficiente de sesgo de GC *G* se evalúa como la pendiente de la regresión entre los recuentos sin procesar *M* medidos por bins y el contenido de GC del genoma de referencia. La elevación cromosómica *L* se usa después para análisis adicionales (valores de Z, deleciones/duplicaciones maternas, microdeleciones/microduplicaciones fetales, sexo del feto, aneuploidías sexuales, etcétera). El procedimiento encapsulado por la ec. B se denomina Eliminación de error parametrizado y normalización no sesgada (PERUN).

Validación cruzada de parámetros de PERUN

Tal como se infiere en la sección titulada "Separación de sesgo de GC del sesgo sistemático por bins", la evaluación de *I* y S selecciona aleatoriamente el 10 % de euploides conocidos (un conjunto de 1093 LDTv2 en la Fig. 83) y los aparta para la validación cruzada. El modelo lineal aplicado al 90 % restante de los euploides extrae los valores de los parámetros *I* y S específicos para el bin seleccionado (suponiendo $L = 1$). Después, la validación cruzada usa las estimaciones de *I* y S para un bin dado para reproducir los valores de *M* medidos a partir de los valores de *G* medidos tanto en el conjunto de trabajo como en el 10 % restante de euploides (suponiendo de nuevo $L = 1$). La selección aleatoria del subconjunto de validación cruzada se repite muchas veces (100 veces en la Fig. 83, aunque 10 repeticiones serían suficientes). 100 líneas diagonales rectas en la Fig. 83 representan los modelos lineales para 100 selecciones diferentes del subconjunto de trabajo del 90 %. El mismo procedimiento se aplica a todos los bins en el genoma humano, lo que produce un conjunto de ordenadas en el origen *I* y pendientes S para cada ubicación genómica.

Para cuantificar el éxito del modelo y evitar el sesgado de los resultados, usamos el factor R, definido de la siguiente manera:

$$R = \frac{\Sigma_{i=1}^{N}|M_i - P_i|}{\Sigma_{i=1}^{N}|M_i|} \qquad (C)$$

El numerador en la ec. B suma las desviaciones absolutas de los valores de recuento predichos (*P*, ec. B) a partir de las mediciones reales *(M).* El numerador simplemente suma las mediciones. El factor *R* puede interpretarse como el error residual en el modelo, o la variación no explicada. El factor *R* está directamente tomado de la práctica de refinamiento del modelo cristalográfico, que es vulnerable al sesgo. En cristalografía, el sesgo se detecta y mide por el factor R evaluado dentro del subconjunto de validación cruzada de observables. Los mismos conceptos se aplican en el contexto de la eliminación de sesgo de recuento de genoma completo.

La Fig. 90 muestra los factores R evaluados para el subconjunto de validación cruzada (eje y) representados gráficamente frente a factores *R* evaluados para el conjunto de trabajo (entrenamiento) para el bin n.°2404 del cromosoma 2. Existen 100 puntos de datos ya que la selección aleatoria del subconjunto de validación cruzada se repitió 100 veces. Se observa una relación lineal típica, con los valores de R_{cv} crecientes (desviación de medición) que acompañan al $R_{trabajo}$ decreciente.

La Fig. 90 puede interpretarse en términos del error porcentual (o error relativo) del modelo para este bin particular. R_{cv} siempre supera a $R_{trabajo}$, habitualmente, en ~ el 1 %. En este caso, tanto R_{cv} como $R_{trabajo}$ permanecen por debajo del 6 %, lo que significa que puede esperarse un error de ~ el 6 % en los valores *M* predichos usando el coeficiente de sesgo de GC medido *G* y los parámetros del modelo *I* y S del procedimiento descrito anteriormente.

Valores de error de validación cruzada

Las Fig. 90-91 muestran errores de validación cruzada para los bins cr2_2404 y cr2_2345, respectivamente. Para estos y muchos otros bins, los errores nunca superan el 6 %. Algunos bins, tales como cr1_31 (Fig. 92) tienen errores de validación cruzada que se aproximan al 8 %. Aún otros (Fig. 93-95) tienen errores de validación cruzada mucho mayores, a veces superiores al 100 % (40 % para cr1_10 en la Fig. 93, 350 % para cr1_9 en la Fig. 94, y 800 % para cr1_8 en la Fig. 95).

La Fig. 96 muestra la distribución de máx.(R_{cv}, $R_{trabajo}$) para todos los bins. Solo un puñado de bins tiene errores por debajo del 5 %. La mayoría de los bins tienen errores por debajo del 7 % (48956 autosomas de 61927 en total incluyendo X e Y). Algunos bins tienen errores de entre el 7 % y el 10 %. La cola consiste en bins con errores que superan el 10 %.

La Fig. 97 correlaciona los errores de validación cruzada con los errores relativos por bin estimados a partir del conjunto de réplicas técnicas. Los puntos de datos en la región de color azul corresponden a errores de validación cruzada de entre el 7 % y el 10 %. Los puntos de datos en la región de color rojo indican bins con un error de validación cruzada que supera el 10 %. Los puntos de datos en la región de color gris (error < 7 %) representan la mayoría de bins.

En las Fig. 91-95, el número entre paréntesis después del nombre del bin por encima del recuadro superior derecho indica la razón entre la ordenada en el origen hallada para ese bin específico y la mediana de recuento de genoma completo por bin. Los errores de validación cruzada aumentan evidentemente con el valor decreciente de esa razón. Por ejemplo, el bin cr1_8 nunca obtiene más de 3 recuentos y su error relativo se aproxima al 800 %. Cuanto menor sea el número esperado de recuentos para un bin dado, menos fiable se volverá ese bin.

Selección de bins basada en la validación cruzada

Basándose en las observaciones descritas en la sección anterior titulada “Eliminación de bins no informativos” (Fig. 78 y Fig. 80-81), se usaron errores de validación cruzada como criterio para el filtrado de bins. El procedimiento de selección elimina todos los bins con errores de validación cruzada que superan el 7 %. El filtrado también elimina todos los bins que contienen de manera sistemática recuentos de valor cero. El subconjunto restante contiene 48956 bins autosómicos. Estas son los bins usados para evaluar representaciones cromosómicas y para clasificar las muestras como afectadas o euploides. El punto de corte de 7 % se justifica por el hecho de que la brecha que separa las puntuaciones Z euploides de las puntuaciones Z de la trisomía forma una meseta en el error de validación cruzada del 7 % (Fig. 98).

La Fig. 99A (todos los bins) y la Fig. 99B (bins validados de forma cruzada) demuestran que la selección de bins descrita en el ejemplo 4 elimina principalmente los bins con baja capacidad de mapeo.

Tal como se esperaba, la mayoría de los bins eliminados tienen ordenadas en el origen mucho más pequeñas que la mediana de recuento de bins de genoma completo. Sin sorpresa, la selección de bins se superpone en gran medida con la selección descrita en la sección anterior titulada “Eliminación de bins no informativos” (Fig. 25 y 27-28).

Errores en los parámetros de modelo

Las Fig. 100-101 muestran los intervalos de confianza del 95 % (líneas curvas) del modelo lineal ajustado (línea recta delgada) para dos bins (cr18_6 y cr18_8). Las líneas rectas de color gris gruesas se obtienen al reemplazar el parámetro S por la diferencia entre el contenido de GC de los mismos dos bins y la mediana del contenido de GC del

cromosoma 18. El rango de errores se evalúa basándose en los errores en los parámetros del modelo I y S para esos dos bins, según lo notificado por el modelo lineal. Además, los coeficientes de sesgo de GC más grandes contienen además errores más grandes. La gran incertidumbre que corresponde a coeficientes de sesgo de GC extremadamente grandes sugiere que el intervalo de aplicabilidad del PERUN no modificado se limita a coeficientes de sesgo de GC moderados. Más allá de ese rango, deben tomarse medidas adicionales para eliminar el sesgo de GC residual. Afortunadamente, solo se ven afectadas muy pocas muestras (aproximadamente el 10 % de la población LDTv2CE).

Las Fig. 102-104 muestran los errores en los parámetros I y S del modelo y la correlación entre el error en S y el valor de la ordenada en el origen.

Normalización secundaria

Los altos valores de coeficientes de sesgo de GC superan el rango lineal supuesto por el modelo de PERUN y son remediados por una etapa adicional de normalización de LOESS de GC después de la normalización PERUN. La naturaleza multiplicativa del procedimiento de LOESS no infla significativamente la variabilidad ya que los recuentos normalizados ya están muy cerca de 1. Alternativamente, LOESS puede reemplazarse por un procedimiento aditivo que resta los residuos. La normalización secundaria opcional que se usa a menudo solo se requirió para una minoría de muestras (aproximadamente el 10 %).

Relleno de agujeros (relleno)

Las Fig. 68-69 confirman la presencia de un gran número de deleciones y duplicaciones maternas que tienen el potencial de crear falsos positivos o falsos negativos, dependiendo de sus tamaños y ubicaciones. Se ha diseñado un procedimiento opcional denominado relleno de agujeros para eliminar las interferencias de estas aberraciones maternas. El procedimiento simplemente proporciona el perfil normalizado para que permanezca cerca de 1 cuando se desvía por encima de 1,3 o por debajo de 0,7. En LDTv2CE, el relleno de agujeros (es decir, relleno) no afectó significativamente a la clasificación. Sin embargo, la Fig. 105 muestra un perfil WI que contiene una deleción grande en el cromosoma 4. El relleno de agujeros convierte ese perfil del falso positivo de cromosoma 13 en verdadero negativo del cromosoma 13.

Resultados

Esta sección comenta los resultados de PERUN para trisomía 13, trisomía 18 y trisomía 21 (T13, T18 y T21, respectivamente), determinación del sexo, y aneuploidía sexual.

Variación reducida

La Fig. 106 compara la distribución de desviaciones estándar de los perfiles de recuento de bins antes y después de la normalización PERUN. Las distribuciones resultantes de representaciones cromosómicas para casos de euploides y de trisomía se muestran en la Fig. 107.

Clasificación mejorada de T13, T18 y T21

Las Fig. 108-111 comparan los resultados de la clasificación de PERUN de LDTv2CE con los obtenidos usando recuentos de GCRM. Además de eliminar dos falsos positivos del cromosoma 18, dos falsos negativos del cromosoma 18, y dos falsos negativos del cromosoma 21, PERUN casi duplica la brecha entre los euploides y los casos afectados, a pesar del hecho de que la mayor elevación de plexación disminuyó el número de recuentos por muestra (datos de ELAND). Se obtienen resultados similares cuando se aplican parámetros de PERUN entrenados en datos de Eland de LDTv2CE a mediciones WI. Las alineaciones de Bowtie requieren un conjunto diferente de parámetros y filtrado adicional de bins, lo que representa una baja capacidad de mapeo en algunos bins, pero sus resultados se aproximan a los que se observan con las alineaciones de ELAND.

Ejemplo 5: Descripción adicional de PERUN

Los ejemplos de métodos de eliminación del error parametrizado y normalización no sesgada (PERUN) se describen en el ejemplo 4, y una descripción adicional de tales métodos se proporciona en este ejemplo 5.

La secuenciación masiva en paralelo del ADN circulante, libre de células (por ejemplo, a partir de plasma materno) puede cuantificar, en condiciones ideales, las elevaciones cromosómicas al contar las lecturas secuenciadas si se alinean de manera inequívoca con un genoma humano de referencia. Tales métodos que incorporan cantidades masivas de datos replicados pueden mostrar, en algunos casos, desviaciones estadísticamente significativas entre las elevaciones cromosómicas medidas y esperadas que pueden implicar aneuploidía [Chiu *et al*., Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc.Natl.Acad.Sci USA.* 2008; 105: 20458-20463; Fan *et al*., Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc.Natl.Acad.Sci USA.* 2008; 105: 16266-16271; Ehrich *et al*., Noninvasive detection of fetal trisomy 21 by sequencing of DNA in maternal blood: a study in a clinical setting, *American Journal of*

Obstetrics and Gynecology - AMER J OBSTET GYNECOL, vol. 204, n.° 3, págs. 205.e1-205.e11,2011 DOI: 10.1016/j.ajog.2010,12.060]. De manera ideal, la distribución de lecturas alineadas debe cubrir las secciones euploides del genoma a un nivel constante (Fig. 62 y Fig. 63). En la práctica, la uniformidad puede ser difícil de alcanzar porque las mediciones multiplexadas de secuenciación de próxima generación (NGS) producen normalmente una baja cobertura (aproximadamente 0,1) con posiciones iniciales de lectura escasamente dispersas. En algunas implementaciones, este problema se supera parcialmente dividiendo el genoma en secciones (bins) no superpuestas de las mismas longitudes y asignando a cada bin el número de lecturas que se alinean dentro del mismo. En algunas implementaciones, la irregularidad residual que proviene del sesgo de GC [Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high- throughput DNA sequencing. Nucleic Acids Res. Sep. de 2008; 36(16): e105. publicación electrónica del 26 de julio de 2008.] se suprime en gran medida usando la inversión de tendencias multiplicativas con respecto al contenido de GC por bins (Fan HC, Quake SR (2010) Sensitivity of Noninvasive Prenatal Detection of Fetal Aneuploidy from Maternal Plasma Using Shotgun Sequencing Is Limited Only by Counting Statistics. PLoS ONE 5(5): e10439. doi:10.1371/journal.pone.0010439). En algunas implementaciones, el aplanamiento resultante del perfil de recuento permite la clasificación exitosa de trisomías fetales en un entorno clínico usando codificación de barras cuádruplex [Palomaki *et al*., DNA sequencing of maternal plasma to detect Down syndrome: an international clinical validation study. *Genet Med.,* nov. de 2011; 13(11): 913-20,].

La transición de un cuádruplex (es decir, 4 lecturas simultáneas de muestra) a mayores niveles de plexación de muestra (por ejemplo, dodecaplex (es decir, 12 lecturas simultáneas de muestra)) empuja los límites de detección basada en NGS de variaciones genéticas (por ejemplo, aneuploidía, trisomía y similares) en un sujeto de prueba (por ejemplo, una mujer embarazada), lo que reduce tanto el número de lecturas por muestra como la brecha que separa las variaciones genéticas (por ejemplo, muestras de euploide de las de trisomía). El submuestreo impulsado por multiplexación aumentada puede imponer requisitos nuevos y más estrictos en los algoritmos de procesamiento de datos (Fig. 64, Fig. 65 y ejemplo 4). En algunas implementaciones, la inversión de tendencia de GC, incluso cuando se acopla con enmascaramiento repetido, requiere cierta mejora (Fig. 66, Fig. 67 y ejemplo 4). En algunas implementaciones, para mantener la sensibilidad lograda con codificación de barras cuádruplex (por ejemplo, indexación cuádruplex), se presentan métodos y algoritmos que son capaces de extraer una señal mínima de interés a partir de un ruido de fondo abrumador tal como se ilustra y describe más adelante y en la Fig. 7, Fig. 8 y ejemplo 4. En algunas implementaciones, se describe un método novedoso denominado ‘‘PERUN’’ (eliminación del error parametrizado y normalización no sesgada).

La eliminación de tendencia de GC convencional puede ser de naturaleza multiplicativa (Fig. 17 y ejemplo 4) y puede no abordar fuentes adicionales de sesgo sistemático, ilustradas en la Fig. 6. En algunos casos, una mediana de perfil de recuento de referencia construida a partir de un conjunto de muestras euploides conocidas puede eliminar el sesgo adicional y conducir a mejoras cualitativas. En algunos casos, una mediana de perfil de recuento de referencia construida a partir de un conjunto de muestras euploides conocidas puede heredar una mezcla de sesgos de GC residuales de las muestras de referencia. En algunas implementaciones, una normalización elimina uno o más tipos ortogonales de sesgo al separarlos entre sí en la elevación de bins, en lugar de conectarlos en el volumen. En algunas implementaciones, se elimina el sesgo de GC y se logra la separación por bins del sesgo de GC de la atenuación dependiente de la posición (Fig. 68, Fig. 69 y ejemplo 4). En algunas implementaciones, se obtienen brechas sustancialmente mayores entre las puntuaciones Z euploides y de trisomía con relación a los resultados de GCRM cuádruplex y dodecaplex. En algunas implementaciones, se detectan microdeleciones y duplicaciones maternas y fetales. En algunas implementaciones, las fracciones fetales se miden con precisión. En algunas implementaciones, el sexo se determina de manera fiable. En algunas implementaciones, se identifica la aneuploidía sexual (por ejemplo, aneuploidía sexual fetal).

Método y definiciones de PERUN

En algunas implementaciones todo el genoma de referencia se divide en un conjunto ordenado B de *J* bins:

$$B = \{b_j | j = 1,...,J\} \quad \text{(D)}$$

Las longitudes de bins pueden restringirse para albergar tramos genómicos de contenido de GC relativamente uniforme. En algunas implementaciones, los bins adyacentes pueden solaparse. En algunas implementaciones, los bins adyacentes no se solapan. En algunas implementaciones, los bordes de bins pueden ser equidistantes o pueden variar para compensar sesgos sistemáticos, tales como la composición de nucleótidos o la atenuación de la señal. En algunas implementaciones, un bin comprende posiciones genómicas dentro de un solo cromosoma. Cada bin b_j se caracteriza por el contenido de GC g_j^0 de la porción correspondiente del genoma de referencia. En algunas implementaciones, se asigna a todo el genoma un perfil de contenido de GC de referencia:

$$g^0=[g_1{}^0\ g_2{}^0 \cdots g_J{}^0] \quad \text{(E)}$$

El mismo perfil g^0 puede aplicarse a todas las muestras alineadas con el genoma de referencia elegido.

Un subconjunto adecuado o trivial de bins b,

$$b \subseteq B \quad \text{(F)}$$

puede seleccionarse para satisfacer determinados criterios, tales como excluir los bins con $g_j^0 = 0$, bins con valores extremos de g_j^0, bins caracterizados por una baja complejidad o baja capacidad de mapeo (Derrien T, Estelle' J, Marco Sola S, Knowles DG, Raineri E, *et al*. (2012) Fast Computation and Applications of Genome Mappability. PLoS ONE 7(1): e30377, doi:10.1371/journal.pone.0030377), bins altamente variables o de otra manera no informativos, regiones con señal atenuada de manera sistemática, aberraciones maternas observadas, o cromosomas enteros (cromosomas X, Y, triploides, y/o cromosomas con contenido extremo de GC). El símbolo $||b||$ indica el tamaño de b.

Todas las lecturas secuenciadas de la muestra *i* alineada de manera inequívoca dentro de un bin b_j forman un conjunto a_{ij} cuya cardinalidad M_{ij} representa recuentos medidos sin procesar asignados a ese bin. En algunas implementaciones, el vector de recuentos de bins medidos para la muestra *i* constituye el perfil de recuento sin procesar para esa muestra. En algunas implementaciones, esta es la observación principal para los fines del PERUN:

$$M_i = [M_{i1}\ M_{i2} \cdots M_{iJ}] \quad \text{(G)}$$

Para permitir comparaciones entre diferentes muestras, la constante de escalado N_i, se evalúa como la suma de los recuentos de bins sin procesar sobre un subconjunto de los bins:

$$N_i = \sum_{b \subseteq B} M_{ij} \quad \text{(H)}$$

En algunas implementaciones *b* en la ec. H se restringe a bins autosómicos. En algunas implementaciones *b* en la ec. H no se restringe a bins autosómicos. La división de M_i entre los recuentos totales N_i produce los recuentos de bins sin procesar escalados m_{ij}:

$$m_i = [m_{i1}\ m_{i2} \cdots m_{iJ}] = M_i/N_i \quad \text{(I)}$$

La composición de nucleótidos del conjunto se describe por el contenido de GC observado del bin g_{ij}. El perfil de contenido de GC observado específico de muestra g_i, recopila el contenido de GC específico de bin individual en un vector:

$$g_i = [g_{i1}\ g_{i2} \cdots g_{iJ}] \quad \text{(J)}$$

En algunas implementaciones, $g_i \neq g^0$ y $g_{i1} \neq g_{i2 \neq i1}$. El símbolo *g* indica el perfil de contenido de GC independientemente de su origen, es decir, si se deriva del genoma de referencia o de las alineaciones de lectura específicas de muestra. En algunas implementaciones, las ecuaciones modelo usan *g*. En algunas implementaciones, las implementaciones reales pueden sustituir *g* por g^0 o g_i.

Para una sola muestra *i*, se supone una relación lineal entre m_i y *g*, donde G_i y r_i indican la pendiente específica de muestra de la línea de regresión y el conjunto de residuos, respectivamente:

$$m_i = G_i g + r_i \quad \text{(K)}$$

La regresión puede extenderse por todo el conjunto *B* (ec. D) o su subconjunto *b* apropiado (ec. F). La pendiente G_i observada también se denomina coeficiente de sesgo de GC escalado. G_i expresa el volumen de la vulnerabilidad de la muestra *i* con respecto al sesgo de GC sistemático. En algunas implementaciones, para minimizar la cantidad de parámetros del modelo, los términos de orden superior, vinculados con la curvatura de la razón $m_i(g)$ y encapsulados en los residuos r_i no se abordan explícitamente. En algunas implementaciones, dado que los recuentos totales específicos de muestra N_i confundieron las interacciones entre observables registradas en diferentes muestras, el equivalente no sellado de G_i, que relaciona M_i con *g,* es menos útil y no se considerará.

El vector de verdaderas elevaciones cromosómicas l_{ij} que corresponden a bins $bj \in b$ en la muestra *i* forma el perfil de elevación cromosómica específico de muestra:

$$l_i = [l_{i1}\ l_{i2} \cdots l_{iJ}] \quad \text{(L)}$$

En algunas implementaciones, el objetivo es derivar estimaciones de l_i a partir de m_i eliminando los sesgos sistemáticos presentes en m_i.

Los valores de l_{iJ} son específicos de bin y además específicos de muestra. Comprenden contribuciones maternas y fetales, de manera proporcional a sus respectivas ploidías P_{ij}^M y P_{ij}^F. La ploidía P_{iJ} específica de bin y específica de muestra puede definirse como un múltiplo entero de una mitad, representando los valores de 1, 1/2, 0, 3/2 y 2 euploidía, deleción heterocigota, deleción homocigota, duplicación heterocigota y duplicación homocigota, respectivamente. En algunos casos, la trisomía de un cromosoma dado implica valores de ploidía de 3/2 a lo largo de todo el cromosoma o su porción sustancial. Cuando tanto la madre como el feto son diploides ($P_{ij}^M = P_{ij}^F = 1$), l_{iJ} es igual a alguna elevación euploide elegida arbitrariamente *E*. En algunas implementaciones, una elección conveniente establece *E* en $1/||b||$, asegurando así que el

perfil l_i, esté normalizado. En ausencia de selección de bins, $||b|| = ||B|| = J \Rightarrow E = 1/J$. En algunas implementaciones, E puede establecerse en 1 para la visualización. En algunas implementaciones, se satisface la siguiente relación:

$$l_{ij} = E\left[(1-f_i)P_{ij}^{M} + f_i P_{ij}^{F}\right] \quad \text{(M)}$$

El símbolo f_i representa la fracción del ADN fetal presente en el ADN circulante, libre de células del plasma materno en la muestra i. Cualquier desviación de la euploidía, ya sea fetal ($P_{ij}^{F} \neq 1$) o materno ($P_{ij}^{M} \neq 1$), provoca diferencias entre l_{iJ} y E que pueden aprovecharse para estimar f_i y detectar microdeleciones/microduplicaciones o trisomía.

Para alcanzar el objetivo de extraer l_i a partir de m_i, se postula una relación lineal entre los recuentos sin procesar escalados específicos de bin m_{ij} medidos en una muestra dada y los coeficientes de sesgo de GC específicos de muestra:

$$m_i = l_i I + G_i S \quad \text{(N)}$$

La matriz diagonal I y el vector S recopilan ordenadas en el origen específicas de bin y pendientes del conjunto de ecuaciones lineales resumidas por la ec. N:

$$I = \begin{bmatrix} I_1 & 0 & \cdots & 0 \\ 0 & I_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & I_J \end{bmatrix} \quad \text{(O)}$$

$$S = [S_1\ S_2\ \cdots\ S_J] \quad \text{(P)}$$

Tanto I como S son independientes de la muestra. Las ordenadas en el origen I_j pueden considerarse valores euploides esperados para recuentos de filas escalados en ausencia de sesgo de GC (es decir, cuando $G_i = 0$). Sus valores reales reflejan la convención adoptada para E *(véase anteriormente).* Las ordenadas en el origen S' no están relacionadas linealmente con las diferencias $g_j^0 - \left\langle g_k^0 \right\rangle$, donde $\left\langle g_k^0 \right\rangle$ representa la mediana del contenido de GC del cromosoma que contiene el bin j.

Una vez que se conocen los valores para los parámetros I y S, el verdadero perfil de elevación cromosómica l_i, se calcula a partir del perfil de recuento sin procesar escalado m_i y el coeficiente de sesgo de GC escalado G_i reorganizando la ec. N:

$$l_i = (m_i - G_i S) I^{-1} \quad \text{(Q)}$$

El carácter diagonal de la matriz de ordenada en el origen I proporciona la inversión de matriz en la ec. Q.

<u>*Estimación de parámetros*</u>

Los parámetros del modelo I y S se evalúan de un conjunto de N perfiles de recuento sin procesar escalados recopilados en muestras cariotipadas como embarazos euploides. N es del orden de 10^3. Los coeficientes de sesgo de GC escalados G_i se determinan para cada muestra ($i = 1,\ldots, N$). Todas las muestras se segregan en un número pequeño de clases según los tamaños y signos de sus valores de G_i. La estratificación equilibra las necesidades opuestas de incluir números suficientemente grandes de representantes y un rango suficientemente pequeño de valores de G_i dentro de cada cubierta exterior. El compromiso de cuatro estratos alberga sesgos de GC negativos, próximos a cero, moderadamente positivos y positivos extremos, con la cubierta próxima a cero que está más densamente poblada. Una fracción de muestras (normalmente, el 10 %) de cada estrato puede seleccionarse aleatoriamente y dejarse aparte para la validación cruzada. Las muestras restantes componen el conjunto de trabajo, usado para entrenar el modelo. Tanto el entrenamiento como la validación cruzada subsiguiente suponen que todas las muestras están libres de deleciones o duplicaciones maternas y fetales a lo largo de todo el genoma:

$$P_{ij}^{M} = P_{ij}^{F} = 1, \forall i = 1,\ldots N, \forall j = 1,\ldots,J \quad \text{(R)}$$

El gran número de muestras compensa las desviaciones maternas ocasionales de la suposición R. Para cada bin j, l_{ij} se establece en E, lo que permite la evaluación de la ordenada en el origen I_j y la pendiente S_j como los coeficientes de la regresión lineal aplicados al conjunto de entrenamiento según la ec. N. También se registran las estimaciones de incertidumbre para I_j y S_j.

La división aleatoria en subconjuntos de trabajo y de validación cruzada se repite múltiples veces (por ejemplo, 10^2), produciendo distribuciones de valores para los parámetros l_j y S_j. En algunas implementaciones, la división aleatoria se repite entre aproximadamente 10 y aproximadamente 10^5 veces. En algunas implementaciones, la división aleatoria se repite aproximadamente 10, aproximadamente 10^2, aproximadamente 10^3, aproximadamente 10^4 o aproximadamente 10^5 veces.

Validación cruzada

Una vez derivados del conjunto de trabajo, los parámetros del modelo l_j y S_j se usan para retrocalcular los recuentos sin procesar escalados a partir de los coeficientes de sesgo de GC escalados usando la ec. N y la suposición R. El símbolo p_{ij} indica los recuentos sin procesar predichos escalados para el bin b_j en la muestra i. Los índices W y CV en texto adicional designan los subconjuntos de trabajo y de validación cruzada, respectivamente. El retrocálculo se aplica a todas las muestras, tanto de W como de CV. Los factores R, prestados de la práctica de refinamiento de estructuras cristalográficas (Brünger, Free R value: a novel statistical quantity for assessing the accuracy of crystal structures, *Nature* 355, 472 - 475 (30 de enero de 1992); doi:10.1038/355472a0), se definen por separado para los dos subconjuntos de muestras:

$$R_j^W = \frac{\sum_{i \in W} \left| m_{ij} - p_{ij} \right|}{\sum_{i \in W} \left| m_{ij} \right|} \qquad \text{(S)}$$

$$R_j^{CV} = \frac{\sum_{i \in CV} \left| m_{ij} - p_{ij} \right|}{\sum_{i \in CV} \left| m_{ij} \right|} \qquad \text{(T)}$$

Ambos factores R son específicos de bin. Como en la cristalografía, los factores R 16-17 pueden interpretarse como errores relativos residuales en el modelo. Habiéndose excluido de la estimación de parámetros, el factor R de validación cruzada R_j^{CV} proporciona una medida verdadera del error para la división W/CV dada, mientras que la diferencia entre R_j^{CV} y R_j^W refleja el sesgo del modelo para el bin j. Un par independiente de valores R se evalúa para cada bin y para cada división aleatoria del conjunto de muestras en W y CV. El máximo de todos los valores de R_j^{CV} y R_j^W obtenidos para las diferentes divisiones aleatorias en W y CV se asigna al bin j como su error de modelo global ε_j.

Selección de bins

Todos los bins con contenido de GC cero g_j^0 se eliminan de la consideración adicional, como lo hace el conjunto $\{b_j : M_{ij} \equiv 0, \forall i = 1, \ldots, N\}$ de bins que reciben de manera sistemática recuentos de valor cero en un gran número de muestras. Además, un valor máximo del error de validación cruzada tolerable ε puede imponerse en todos los bins. En algunas implementaciones, los bins con errores de modelo ε_j que superan el límite superior ε se rechazan. En algunas implementaciones, el filtrado usa puntuaciones de capacidad de mapeo de bins $\mu_j \in [0,1]$ e impone una capacidad de mapeo mínima aceptable μ, rechazando bins con $\mu_j < \mu$ (Derrien T, Estelle' J, Marco Sola S, Knowles DG, Raineri E, *et al.* (2012) Fast Computation and Applications of Genome Mappability. PLoS ONE 7(1): e30377, doi:10.1371/journal.pone.0030377). Para los propósitos de determinar la trisomía fetal de los cromosomas 21, 18 y 13, también pueden excluirse los cromosomas sexuales. El subconjunto p de bins que sobreviven a todas las fases de la selección de bins puede someterse a cálculos adicionales. En algunas implementaciones, el mismo subconjunto p se usa para todas las muestras.

Normalización y estandarización

En algunas implementaciones, para una muestra i dada, las elevaciones cromosómicas l_{ij} correspondientes a la selección de bins β se estiman según la ec. Q. En algunas implementaciones, se aplica una normalización secundaria para eliminar cualquier curvatura de la correlación de l_{ij} frente a contenido de GC. En algunas implementaciones l_{ij} ya es casi sin sesgo, la eliminación de tendencia secundaria es robusta y es inmune a la intensificación de errores. En algunas implementaciones, los procedimientos de libro de texto estándar son suficientes.

En algunas implementaciones, los resultados de la normalización se suman dentro de cada cromosoma:

$$L_{in} = \sum_{b_j \in \beta \cap Chr_n} l_{ij} \;, \quad n = 1, \ldots, 22 \qquad \text{(U)}$$

El material autosómico total en la muestra / puede evaluarse como la suma de todos los términos individuales L_{in}:

$$L_i = \sum_{n-1}^{22} L_{in} \quad \text{(V)}$$

La representación cromosómica de cada cromosoma de interés puede obtenerse dividiendo L_{in} entre L_I:

$$\chi_{in} = L_{in}/L_i \quad \text{(W)}$$

La variabilidad σ_n de la representación del cromosoma n puede estimarse como una D.M.A. sin censurar de valores de χ_{in} a través de una selección de muestras que abarcan múltiples celdas de flujo. En algunas implementaciones, la expectativa $\langle \chi_n \rangle$ se evalúa como la mediana de valores de χ_{in} correspondientes a una selección de muestras de la misma celda de flujo que la muestra sometida a prueba. Ambas selecciones de muestras pueden excluir controles positivos altos, controles positivos bajos, controles negativos altos, blancos, muestras que no cumplen con los criterios de QC y muestras con SD(l_i) que superan un punto de corte predefinido (normalmente 0,10). En conjunto, los valores de σ_n y $\langle \chi_n \rangle$ pueden proporcionar el contexto de estandarización y comparación de representaciones cromosómicas entre diferentes muestras usando puntuaciones Z:

$$Z_{in} = (\chi_{in} - \langle \chi_n \rangle)/\sigma_n \quad \text{(X)}$$

En algunas implementaciones, las aberraciones tales como las trisomías 13, 18 y 21 se indican por valores de Z que superan un valor de predefinido, dictado por el nivel de confianza deseado.

Ejemplo 6: Ejemplos de fórmulas

A continuación se proporcionan ejemplos no limitativos de fórmulas matemáticas y/o estadísticas que pueden usarse en los métodos descritos en el presente documento.

$$Z = \frac{\Delta_1 - \Delta_2}{\sqrt{\sigma_1^2\left(\frac{1}{N_1}+\frac{1}{n_1}\right)+\sigma_2^2\left(\frac{1}{N_2}+\frac{1}{n_2}\right)}}$$

$$P(q) = \frac{1}{\sigma\sqrt{2\pi}} exp[-(q-q_0)/(2\sigma^2)]$$

$$q_0 = 1 + F/2$$

$$z = -F/(2\sigma\sqrt{2})$$

$$B = \int_{-\infty}^{1} P(q)dq = \frac{1}{2}[1 + erf(z)]$$

$$erf(z) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{n!(2n+1)}$$

$$R = \frac{1-B}{B} = \frac{1-erf(z)}{1+erf(z)} = \frac{1-erf[-F/(2\sigma\sqrt{2})]}{1+erf[-F/(2\sigma\sqrt{2})]}$$

Ejemplo 7: Identificación y ajuste de elevaciones (relleno)

Las deleciones y duplicaciones maternas, a menudo representadas como primeras elevaciones en un perfil, pueden eliminarse de perfiles de recuento normalizados con PERUN para reducir la variabilidad cuando se detecta T21, T18 o T13. La eliminación de deleciones y duplicación de un perfil puede reducir la variabilidad (por ejemplo, variabilidad biológica) encontrada en representaciones cromosómicas medidas que se originan a partir de aberraciones maternas.

Primero se identifican todos los bins que se desvían significativamente de la elevación cromosómica esperada de 1. En este ejemplo, algunos bins aislados se eliminan de la selección. Esto es opcional. En este ejemplo, solo se conservan grupos suficientemente grandes de bins atípicos contiguos. Esto también es opcional. Dependiendo de la elevación asignada a un bin atípico o a un grupo de bins atípicos contiguos, se suma un factor de corrección a la elevación medida para ajustarla más próxima a la elevación esperada de 1. Los valores de PAV usados en este ejemplo son +1 (para deleciones maternas homocigotas), +0,5 (para deleciones maternas heterocigotas), -0,5 (para duplicaciones heterocigotas), -1 (para duplicaciones homocigotas) o más (para grandes aumentos bruscos). Los aumentos bruscos grandes no se identifican a menudo como deleciones ni duplicaciones maternas.

Este procedimiento de relleno corrigió la clasificación (por ejemplo, la clasificación como aneuploidía, por ejemplo, una trisomía) para muestras que contiene aberraciones maternas grandes. El relleno convirtió la muestra de WI de T13 falso positivo a T13 verdadero negativo debido a la eliminación de una deleción materna grande en cr4 (Fig. 112-115).

Simulaciones anteriores con datos experimentales han demostrado que dependiendo del cromosoma, la fracción fetal y el tipo de aberración (homocigota o heterocigota, duplicación o deleción), las aberraciones maternas en 20-40 bins de largo puede empujar el valor de Z por encima del borde de clasificación (por ejemplo, umbral) y dar como resultado un falso positivo o un falso negativo. El relleno (por ejemplo, ajuste) puede evitar este riesgo.

Este procedimiento de relleno puede eliminar aberraciones maternas poco interesantes (un factor de confusión), reducir la variabilidad euploide, crear valores sigma más ajustados usados para estandarizar las puntuaciones Z y, por tanto, ampliar la brecha entre los casos de euploides y de trisomía.

Ejemplo 8: Determinación de fracciones fetales a partir de las variaciones en el número de copias materno y/o fetal

Una característica distintiva del método descrito en el presente documento es el uso de aberraciones maternas (por ejemplo, variaciones del número de copias materno y/o fetal) como una sonda que proporciona información sobre la fracción fetal en el caso de una mujer embarazada que porta un feto (por ejemplo, un feto euploide). La detección y cuantificación de aberraciones maternas se asiste normalmente mediante la normalización de los recuentos sin procesar. En este ejemplo, los recuentos sin procesar se normalizan usando PERUN. Alternativamente, la normalización con respecto a una mediana de perfil de recuento de referencia puede usarse de manera similar y con el mismo propósito.

La normalización de los recuentos sin procesar produce niveles cromosómicos por bins específicos de muestra l_{ij} *(i* muestras de recuentos, *j* bins de recuentos). Comprenden contribuciones maternas y fetales, de manera proporcional a su respectiva ploidía P_{ij}^{M} y P_{ij}^{F}. La ploidía específica de bin y específica de muestra P_{ij} se define como un múltiplo entero de 1/2, representando los valores de 1, 1/2, 0, 3/2 y 2 euploidía, deleción heterocigota, deleción homocigota, duplicación heterocigota y duplicación homocigota, respectivamente. Particularmente, la trisomía de un cromosoma dado implica valores de ploidía de 3/2 a lo largo de todo el cromosoma o su porción sustancial.

Cuando tanto la madre como el feto son diploides ($P_{ij}^{M} = P_{ij}^{F} = 1$), l_{ij} es igual a algún nivel euploide elegido arbitrariamente *E*. Una elección conveniente establece *E* en $1/||b||$, donde *b* indica un subconjunto adecuado o trivial del conjunto de todos los bins (*B*), asegurando así que el perfil l_i esté normalizado. En ausencia de selección de bins, $||b|| = ||B|| = J \Rightarrow E = 1/J$. Alternativa y preferiblemente, *E* puede establecerse en 1 para su visualización. En general, se satisface la siguiente relación:

$$l_{ij} = E\left[\left(1 - f_i\right)P_{ij}^{M} + f_i P_{ij}^{F}\right] \qquad (Y)$$

El símbolo f_i representa la fracción del ADN fetal presente en el ADN circulante, libre de células del plasma materno en la muestra *i*. Cualquier desviación de la euploidía, ya sea fetal $(P_{ij}^{F} \neq 1)$ o materna $(P_{ij}^{M} \neq 1)$, provoca diferencias entre l_{IJ} y *E* que pueden aprovecharse para estimar / y detectar microdeleciones/microduplicaciones o trisomía.

Cuatro tipos diferentes de aberraciones maternas se consideran por separado. Las cuatro tienen en cuenta posibles genotipos fetales, ya que el feto puede (o en casos homocigotos debe) heredar la aberración materna. Además, el feto también puede heredar una aberración coincidente del padre. Generalmente, la fracción fetal solo puede medirse cuando $P_{ij}^{M} \neq P_{ij}^{F}$.

A) Deleción materna homocigota $(P_{ij}^{M} = 0)$. Dos ploidías fetales acompañantes posibles incluyen:

   a. $P_{ij}^{F} = 0$, en cuyo caso $l_{ij} = 0$ y la fracción fetal no puede evaluarse a partir de la deleción.

b. P_{ij}^{F} = 1/2, en cuyo caso $l_{ij} = f_i/2$ y la fracción fetal se evalúa como el doble de la elevación promedio dentro de la deleción.

B) Deleción materna heterocigota (P_{ij}^{M} = 1/2). Tres ploidías fetales acompañantes posibles incluyen:

a. P_{ij}^{F} = 0, en cuyo caso $l_{ij} = (1 - f_i)/2$ y la fracción fetal se evalúa como el doble de la diferencia entre el % y la elevación promedio dentro de la deleción.

b. P_{ij}^{F} = 1/2, en cuyo caso l_{ij} = 1/2 y la fracción fetal no puede evaluarse a partir de la deleción.

c. P_{ij}^{F} = 1, en cuyo caso $l_{ij} = (1 + f_i)/2$ y la fracción fetal se evalúa como el doble de la diferencia entre 14 y la elevación promedio dentro de la deleción.

C) Duplicación materna heterocigota (P_{ij}^{M} = 3/2). Tres ploidías fetales acompañantes posibles incluyen:

a. P_{ij}^{F} = 1, en cuyo caso $l_{ij} = (3 - f_i)/2$ y la fracción fetal se evalúa como el doble de la diferencia entre 3/2 y la elevación promedio dentro de la duplicación.

b. P_{ij}^{F} = 3/2, en cuyo caso l_{ij} = 3/2 y la fracción fetal no puede evaluarse a partir de la duplicación.

c. P_{ij}^{F} = 2, en cuyo caso $l_{ij} = (3 + f_i)/2$ y la fracción fetal se evalúa como el doble de la diferencia entre 3/2 y la elevación promedio dentro de la duplicación.

D) Duplicación materna homocigota (P_{ij}^{M} = 2). Dos ploidías fetales acompañantes posibles incluyen:

a. P_{ij}^{F} = 2, en cuyo caso l_{ij} = 2 y la fracción fetal no puede evaluarse a partir de la duplicación.

b. P_{ij}^{F} = 3/2, en cuyo caso $l_{ij} = 2 - f_i/2$ y la fracción fetal se evalúa como el doble de la diferencia entre 2 y la elevación promedio dentro de la duplicación.

Las siguientes muestras de LDTv2CE (Fig. 116 - 131) ilustran la aplicación de determinar la fracción fetal a partir de las variaciones del número de copias maternas y/o fetales. Los pacientes no se seleccionaron al azar y cualquier concordancia con valores de fracción fetal de FQA no debe interpretarse como la medida de mérito de ninguna técnica.

La cita de las patentes, solicitudes de patente, publicaciones y documentos anteriores no es una admisión de que ninguno de los anteriores es técnica anterior pertinente, ni constituye ninguna admisión en cuanto al contenido o fecha de estas publicaciones o documentos.

El término “un(o)” o “una” puede referirse a uno o a una pluralidad de los elementos que modifica (por ejemplo, “un reactivo” puede significar uno o más reactivos) a menos que se aclare contextualmente que se describe o bien uno de los elementos o bien más de uno de los elementos. El término “aproximadamente”, tal como se usa en el presente documento, se refiere a un valor dentro de 10 % del parámetro subyacente (es decir, más o menos el 10 %), y el uso del término “aproximadamente” al comienzo de una cadena de valores modifica cada uno de los valores (es decir, “aproximadamente 1, 2 y 3” se refiere a aproximadamente 1, aproximadamente 2 y aproximadamente 3). Por ejemplo, un peso de “aproximadamente 100 gramos” puede incluir pesos entre 90 gramos y 110 gramos.

Determinadas realizaciones de la tecnología se exponen en la(s) reivindicación/reivindicaciones que sigue(n).

REIVINDICACIONES

1. Un método implementado por ordenador para calcular con sesgo reducido niveles de sección genómica para una muestra de prueba, que comprende:

(a) obtener recuentos de lecturas de secuencia mapeadas en porciones de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante extracelular a partir de una muestra de prueba;
(b) determinar un sesgo de guanina y citosina (GC) para cada una de las porciones del genoma de referencia para múltiples muestras a partir de una relación lineal ajustada para cada muestra entre (i) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, y (ii) el contenido de GC para cada una de las porciones, en donde cada sesgo de GC es un coeficiente de sesgo de GC, coeficiente de sesgo de GC que es la pendiente de la relación lineal entre (i) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, y (ii) el contenido de GC para cada una de las porciones; y
(c) calcular un nivel de sección genómica para cada una de las porciones del genoma de referencia a partir de una relación lineal ajustada entre (i) el sesgo de GC y (ii) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, proporcionándose de este modo niveles de sección genómica calculados, en donde el nivel de sección genómica L_i se determina para cada una de las porciones del genoma de referencia según la ecuación α:

$$L_i = (m_i - G_i S)\, I^{-1} \qquad \text{Ecuación α}$$

en donde G_i es el sesgo de GC, I es la ordenada en el origen de la relación ajustada en (c), S es la pendiente de la relación en (c), m_i es los recuentos medidos mapeados en cada porción del genoma de referencia e i es una muestra,
mediante lo cual el sesgo en los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia se reduce en los niveles de sección genómica calculados.

2. Un sistema que comprende uno o más procesadores y memoria,

memoria que comprende instrucciones ejecutables por el uno o más procesadores y memoria que comprende recuentos de lecturas de secuencia mapeadas en porciones de un genoma de referencia, lecturas de secuencia que son lecturas de ácido nucleico circulante extracelular de una muestra de prueba; e

instrucciones ejecutables por el uno o más procesadores que están configuradas para:

(b) determinar un sesgo de guanina y citosina (GC) para cada una de las porciones del genoma de referencia para múltiples muestras a partir de una relación lineal ajustada para cada muestra entre (i) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, y (ii) el contenido de GC para cada una de las porciones, en donde cada sesgo de GC es un coeficiente de sesgo de GC, coeficiente de sesgo de GC que es la pendiente de la relación lineal entre (i) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, y (ii) el contenido de GC para cada una de las porciones; y
(c) calcular un nivel de sección genómica para cada una de las porciones del genoma de referencia a partir de una relación lineal ajustada entre (i) el sesgo de GC y (ii) los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia, proporcionándose de este modo niveles de sección genómica calculados, en donde el nivel de sección genómica L_i se determina para cada una de las porciones del genoma de referencia según la ecuación α:

$$L_j = (m_i - G_i S)\, I^{-1} \qquad \text{Ecuación α}$$

en donde G_i es el sesgo de GC, I es la ordenada en el origen de la relación ajustada en (c), S es la pendiente de la relación en (c), m_i es los recuentos medidos mapeados en cada porción del genoma de referencia e i es una muestra,
mediante lo cual el sesgo en los recuentos de las lecturas de secuencia mapeadas en cada una de las porciones del genoma de referencia se reduce en los niveles de sección genómica calculados.

3. El método de la reivindicación 1 o sistema de la reivindicación 2, en donde cada una de la relación ajustada de (b) y la relación ajustada de (c) se ajustan independientemente mediante una regresión lineal.

4. El sistema de la reivindicación 2, en donde las instrucciones ejecutables por el uno o más procesadores están configuradas para determinar la presencia o ausencia de una aneuploidía cromosómica fetal para la muestra de prueba según los niveles de sección genómica calculados.

5. El método de la reivindicación 1, que comprende determinar la presencia o ausencia de una aneuploidía cromosómica fetal para la muestra de prueba según los niveles de sección genómica calculados.

6. El método de la reivindicación 5 o sistema de la reivindicación 4, en donde la aneuploidía cromosómica fetal es una trisomía.

7. El método o sistema de la reivindicación 6, en donde la trisomía se selecciona de una trisomía del cromosoma 21, el cromosoma 18, el cromosoma 13 o combinación de los mismos.

8. El método de la reivindicación 1, que comprende, antes de (b), calcular una medida de error para los recuentos de lecturas de secuencia mapeadas en algunas o todas las porciones del genoma de referencia y eliminar o ponderar los recuentos de lecturas de secuencia para determinadas porciones del genoma de referencia según un umbral de la medida de error.

9. El método de la reivindicación 8, en donde el umbral se selecciona según una brecha de desviación estándar entre un primer nivel de sección genómica y un segundo nivel de sección genómica de 3,5 o mayor.

10. El método de la reivindicación 8, en donde la medida de error es un factor R.

11. El método de la reivindicación 10, en donde los recuentos de lecturas de secuencia para una porción del genoma de referencia que tiene un factor R de aproximadamente el 7 % a aproximadamente el 10 % se eliminan antes de (b).

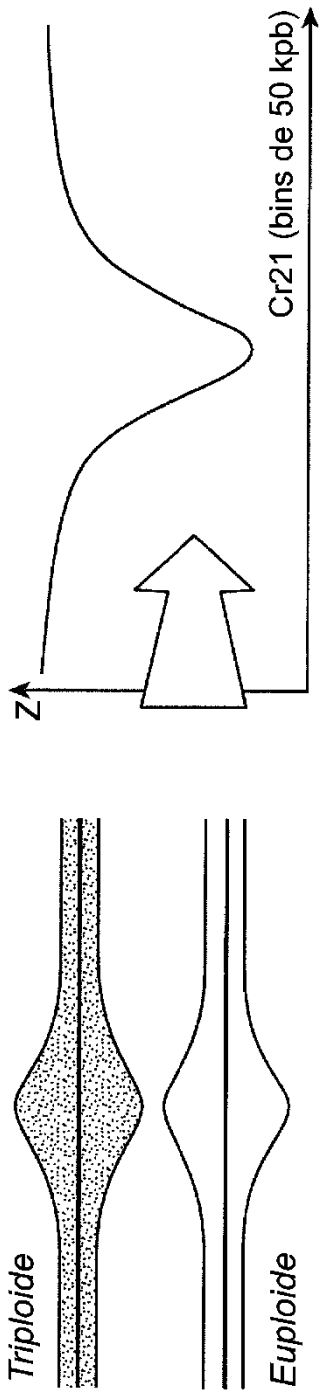
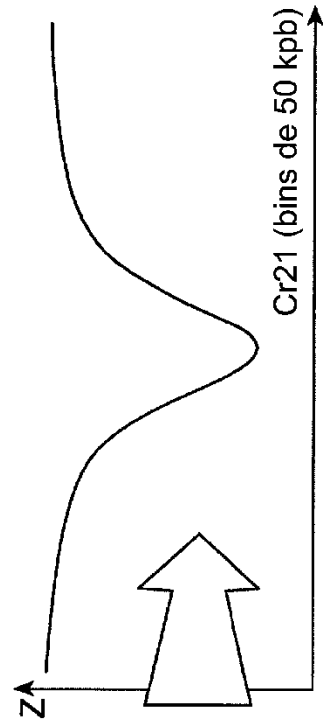


FIG. 1



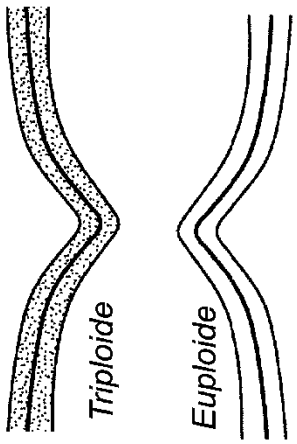


FIG. 2

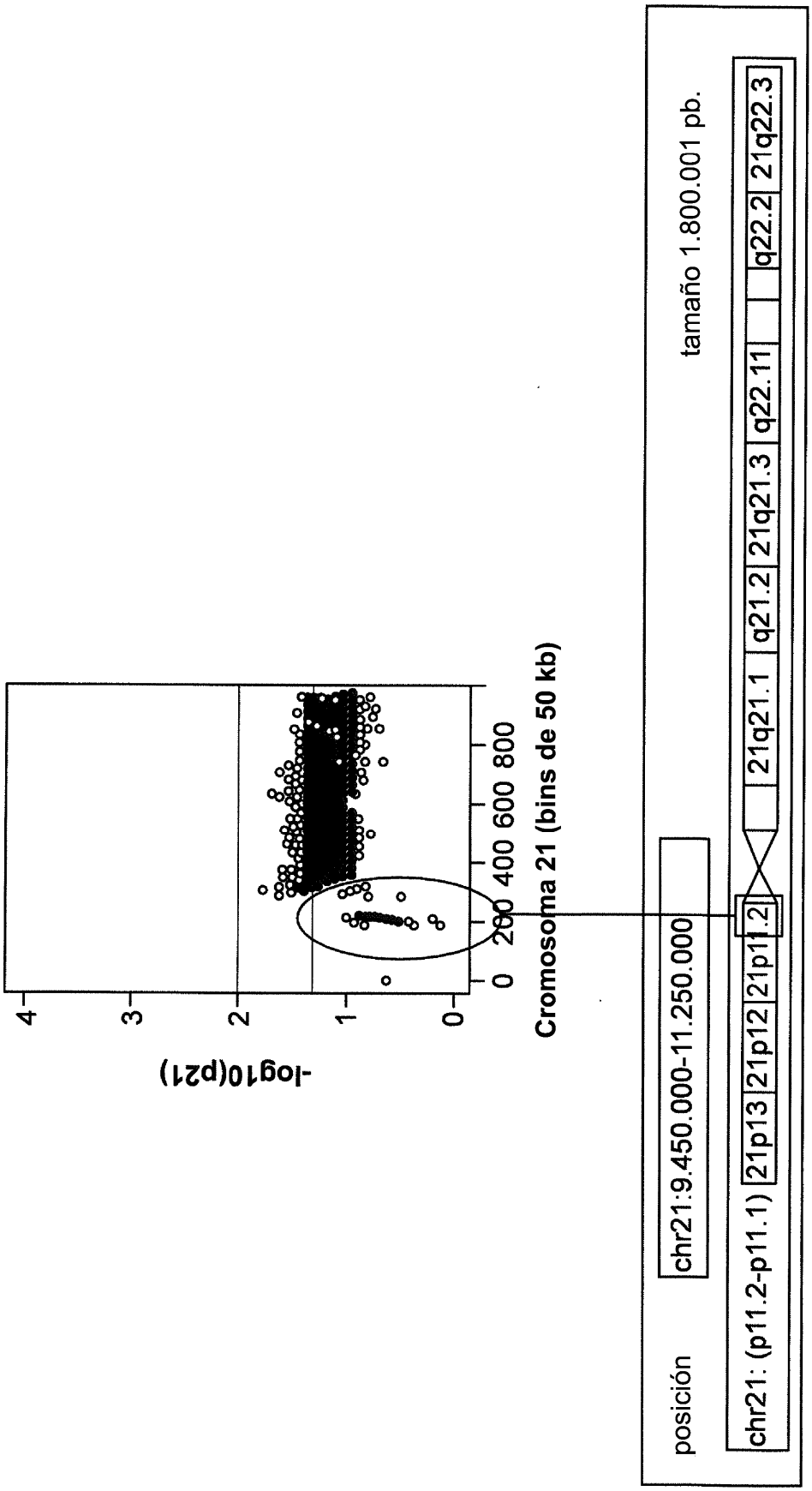

-log10(p21)
0
1
2
3
4
0
200
400
600
800
Cromosoma 21 (bins de 50 kb)
posición
chr21:9.450.000-11.250.000
tamaño 1.800.001 pb.
chr21: (p11.2-p11.1)
21p13
21p12
21p11.2
21q11.1
q21.2
21q21.3
q22.11
q22.2
21q22.3

FIG. 3

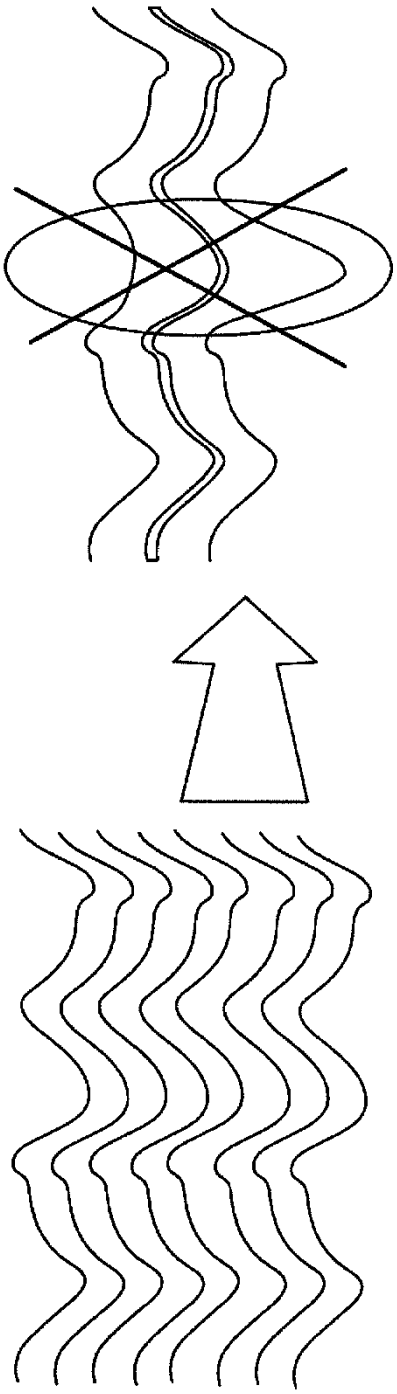


FIG. 4

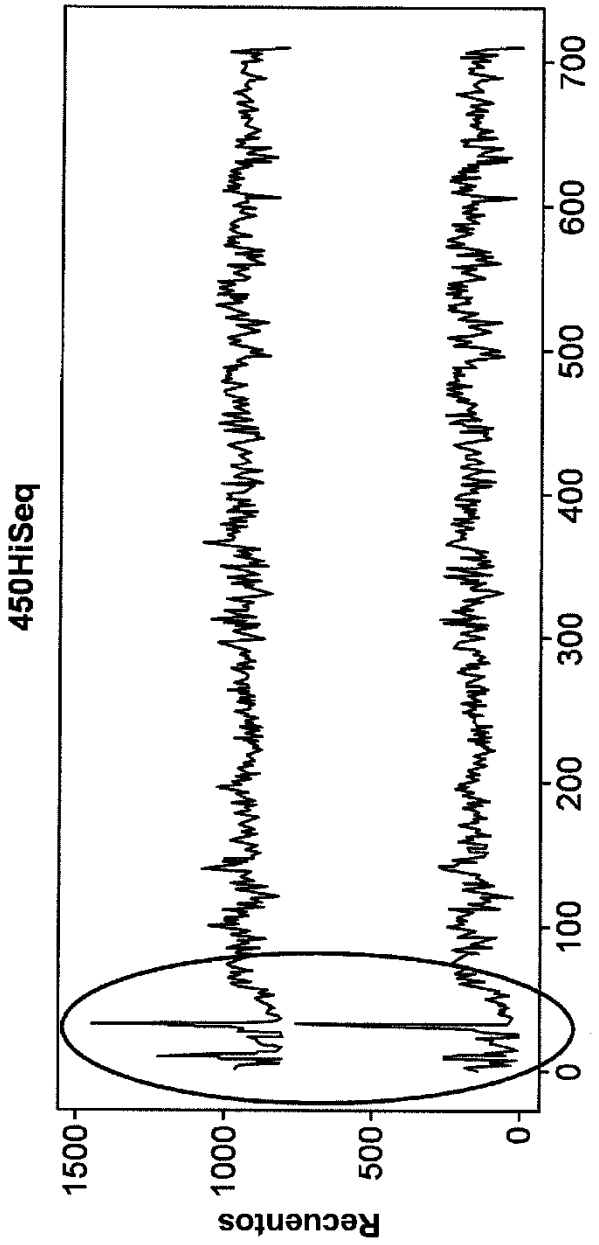

450HiSeq
Recuentos
1500
1000
500
0
0
100
200
300
400
500
600
700
Cr21 (bins de 50 kpb)

FIG. 5

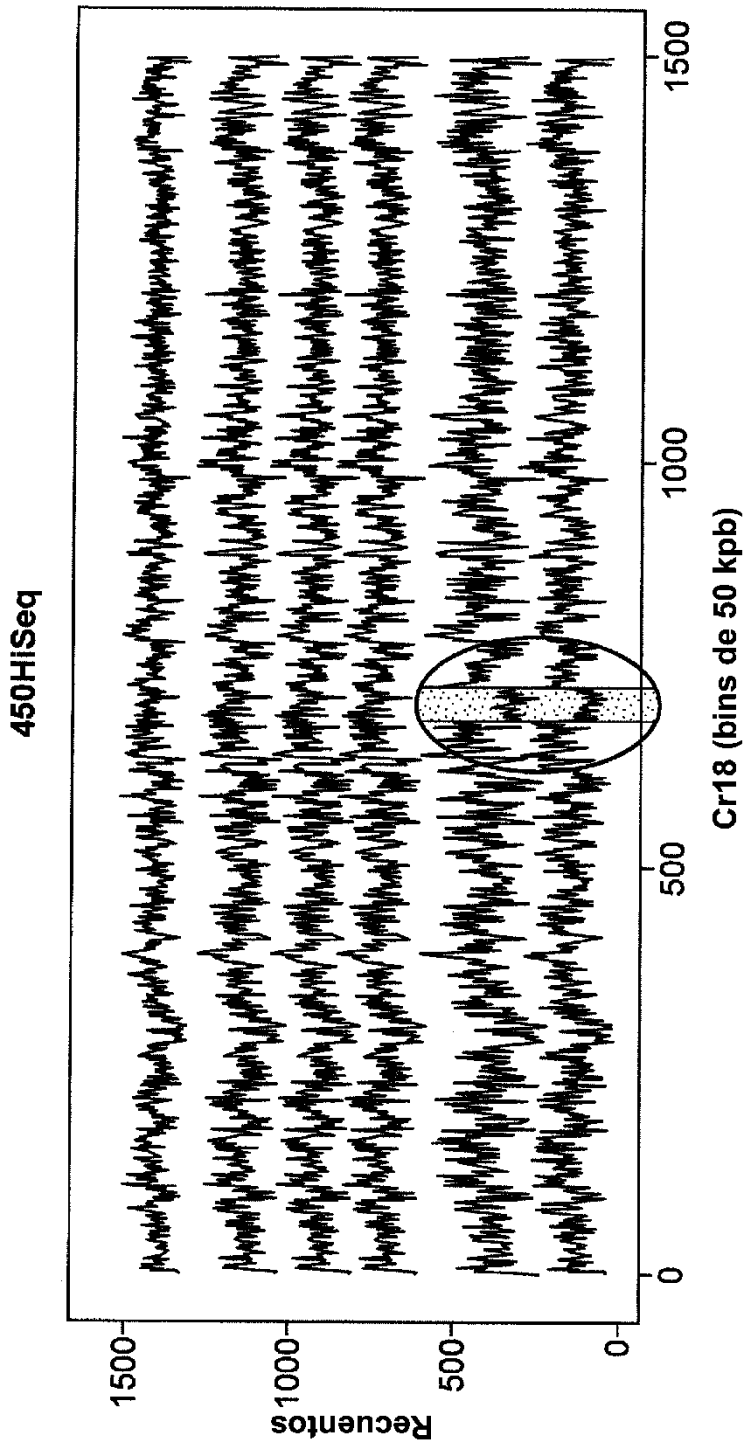

450HiSeq
Recuentos
1500
1000
500
0
0
500
1000
1500
Cr18 (bins de 50 kpb)

FIG. 6

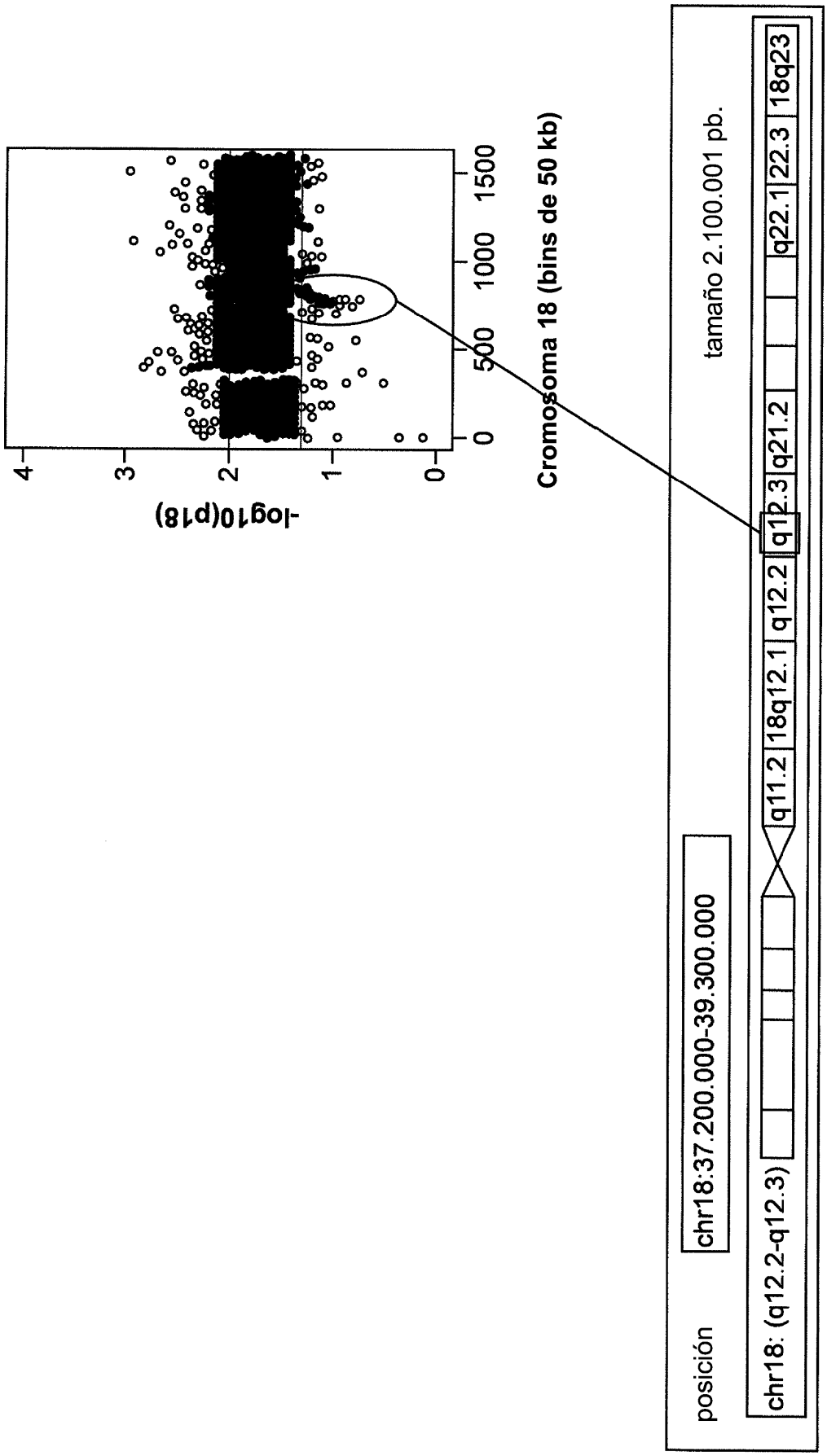

-log10(p18)
4
3
2
1
0
0
500
1000
1500
Cromosoma 18 (bins de 50 kb)
tamaño 2.100.001 pb.
posición
chr18:37.200.000-39.300.000
chr18: (q12.2-q12.3)
q11.2
18q12.1
q12.2
q12.3
q21.2
q22.1
22.3
18q23

FIG. 7

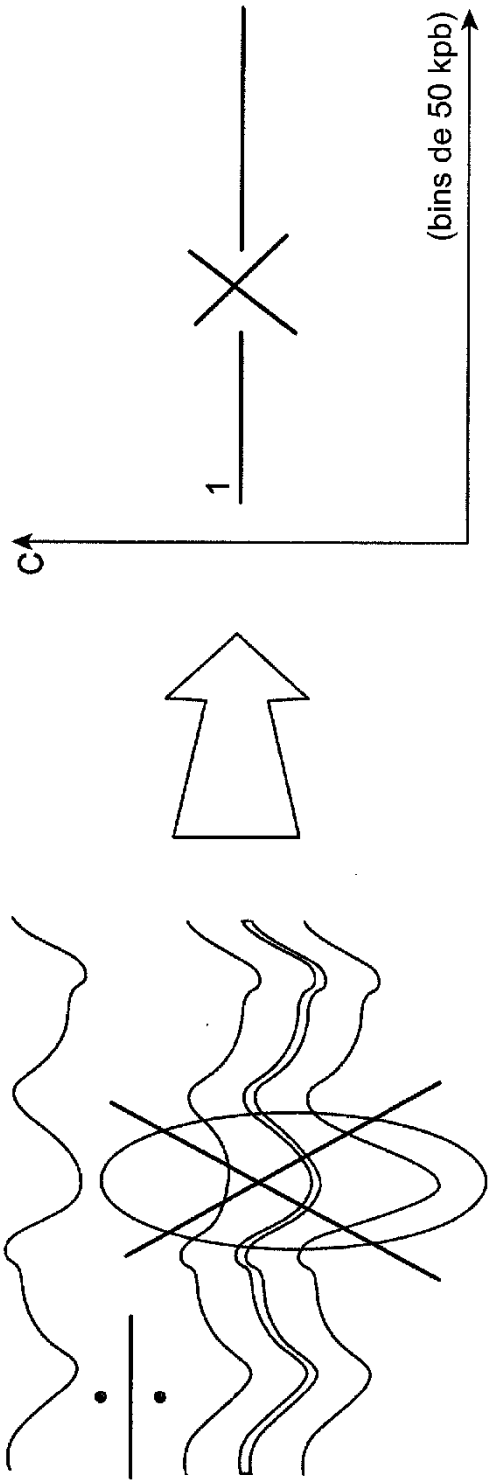


FIG. 8

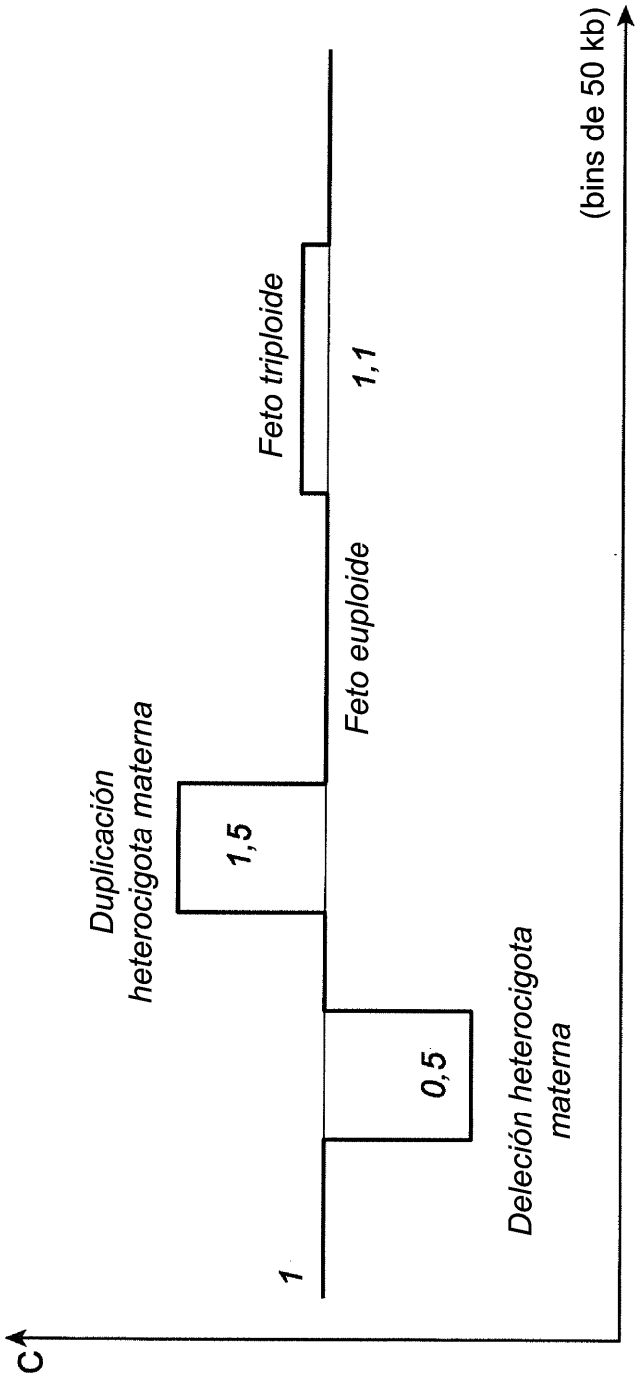

C
(bins de 50 kb)
1
1,5
0,5
1,1
Duplicación heterocigota materna
Feto euploide
Feto triploide
Deleción heterocigota materna

FIG. 9

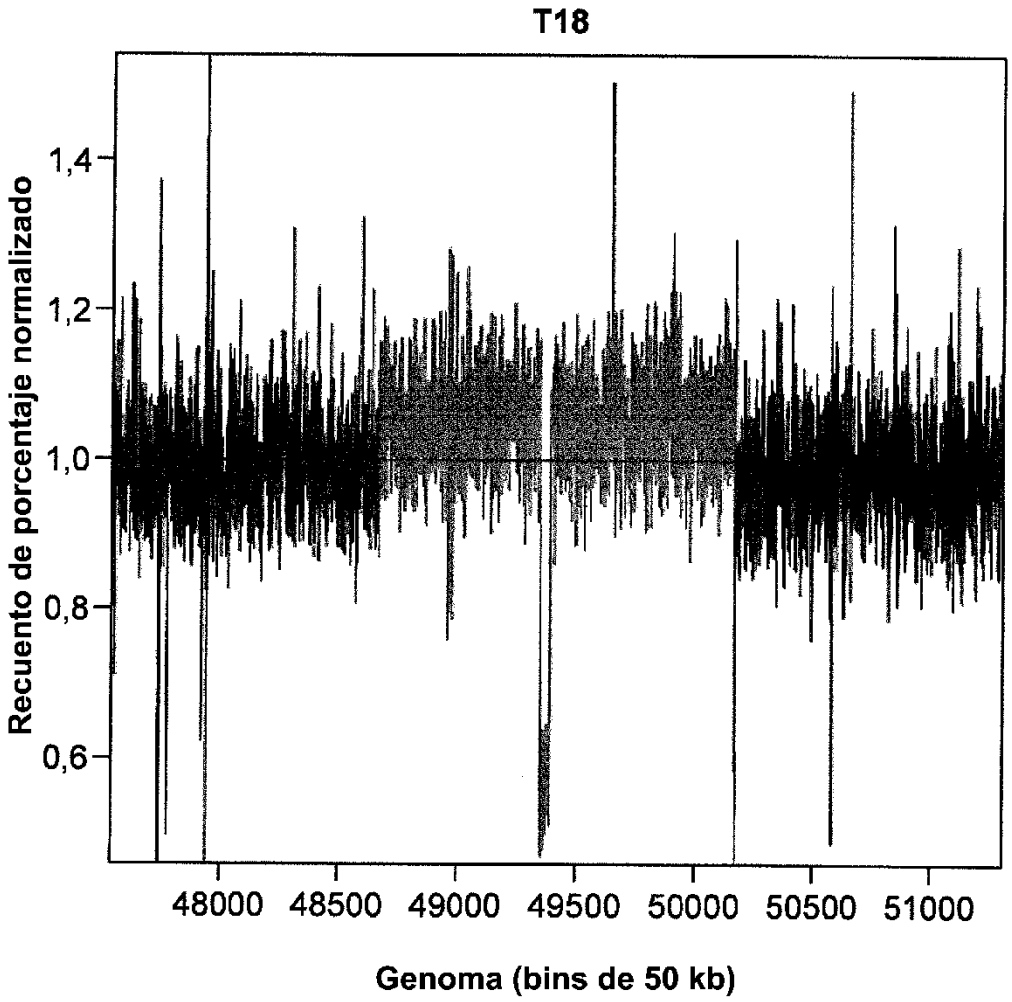

T18
Recuento de porcentaje normalizado
1,4
1,2
1,0
0,8
0,6
48000
48500
49000
49500
50000
50500
51000
Genoma (bins de 50 kb)

FIG. 10

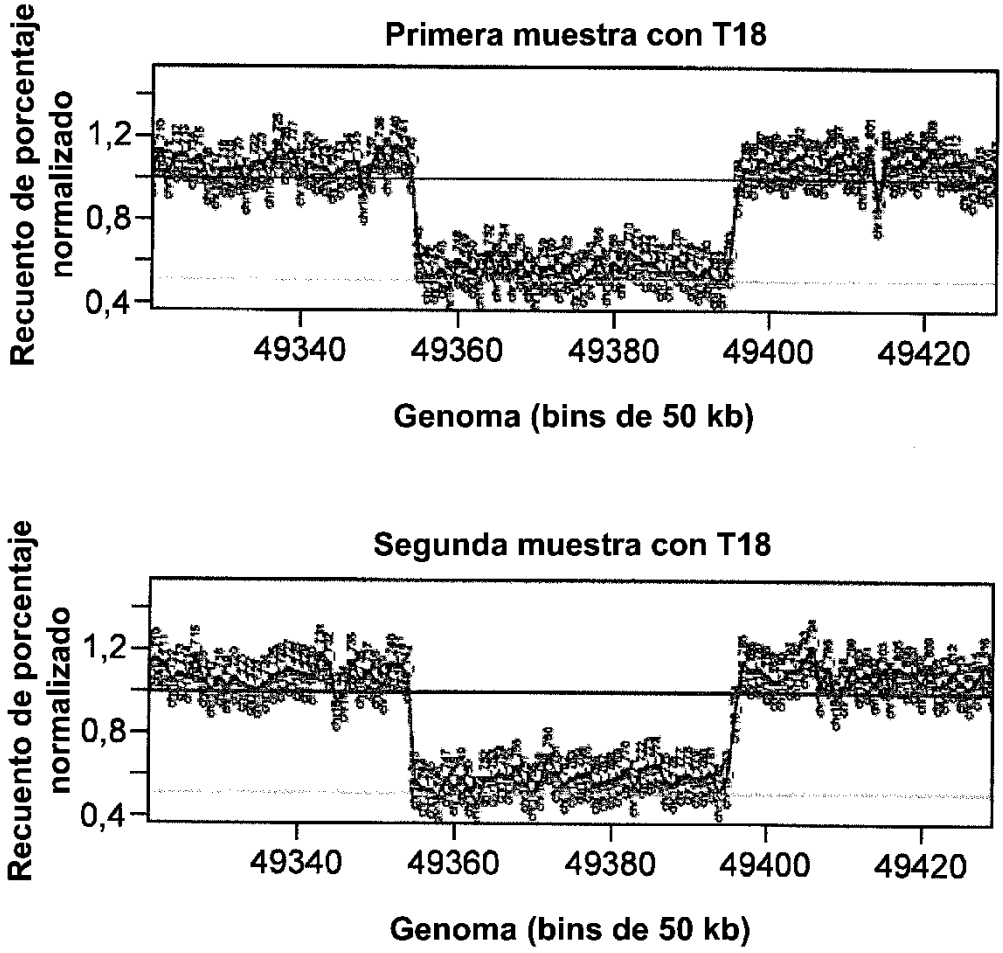

Primera muestra con T18
Recuento de porcentaje normalizado
1,2
0,8
0,4
49340
49360
49380
49400
49420
Genoma (bins de 50 kb)
Segunda muestra con T18
Recuento de porcentaje normalizado
1,2
0,8
0,4
49340
49360
49380
49400
49420
Genoma (bins de 50 kb)

FIG. 11

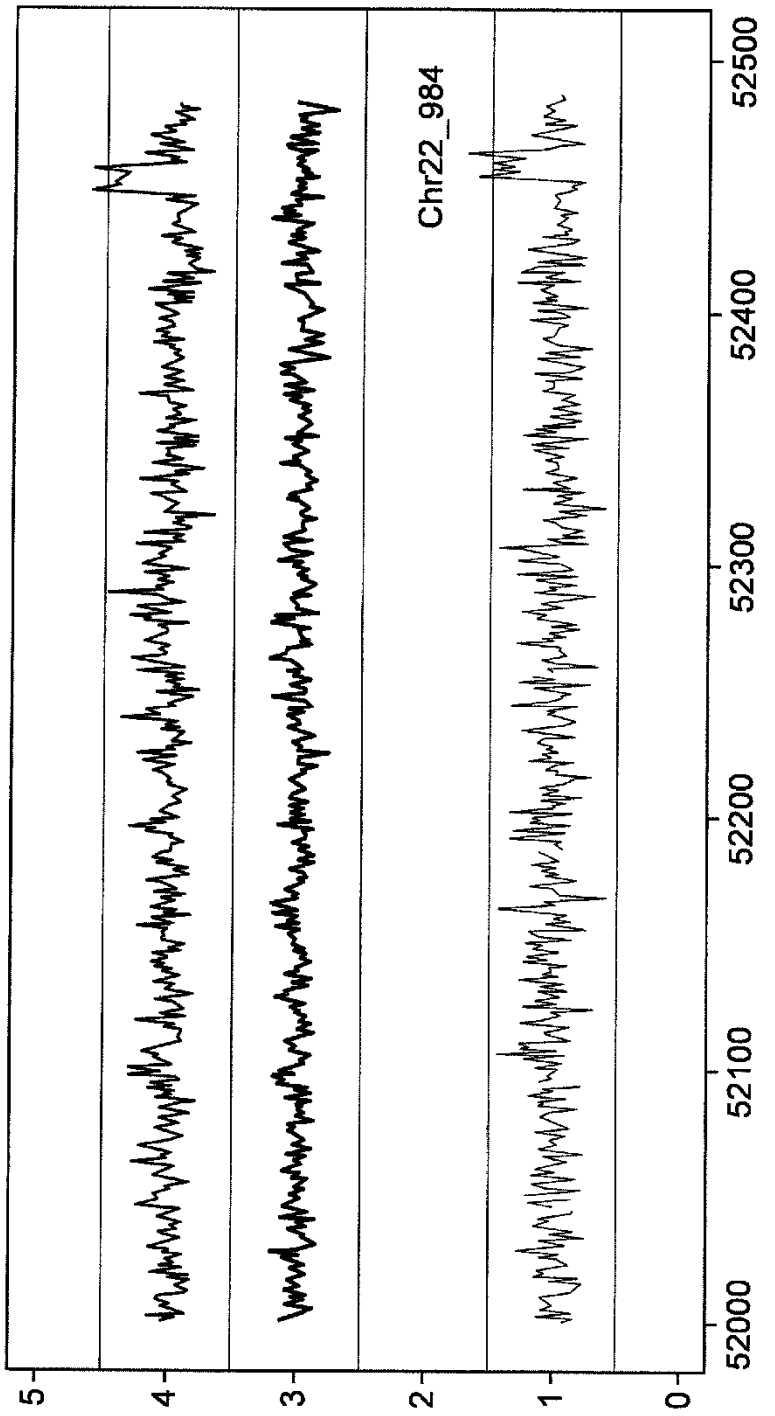

Chr22_984
5
4
3
2
1
0
52000
52100
52200
52300
52400
52500
bins de 50 kpb

FIG. 12

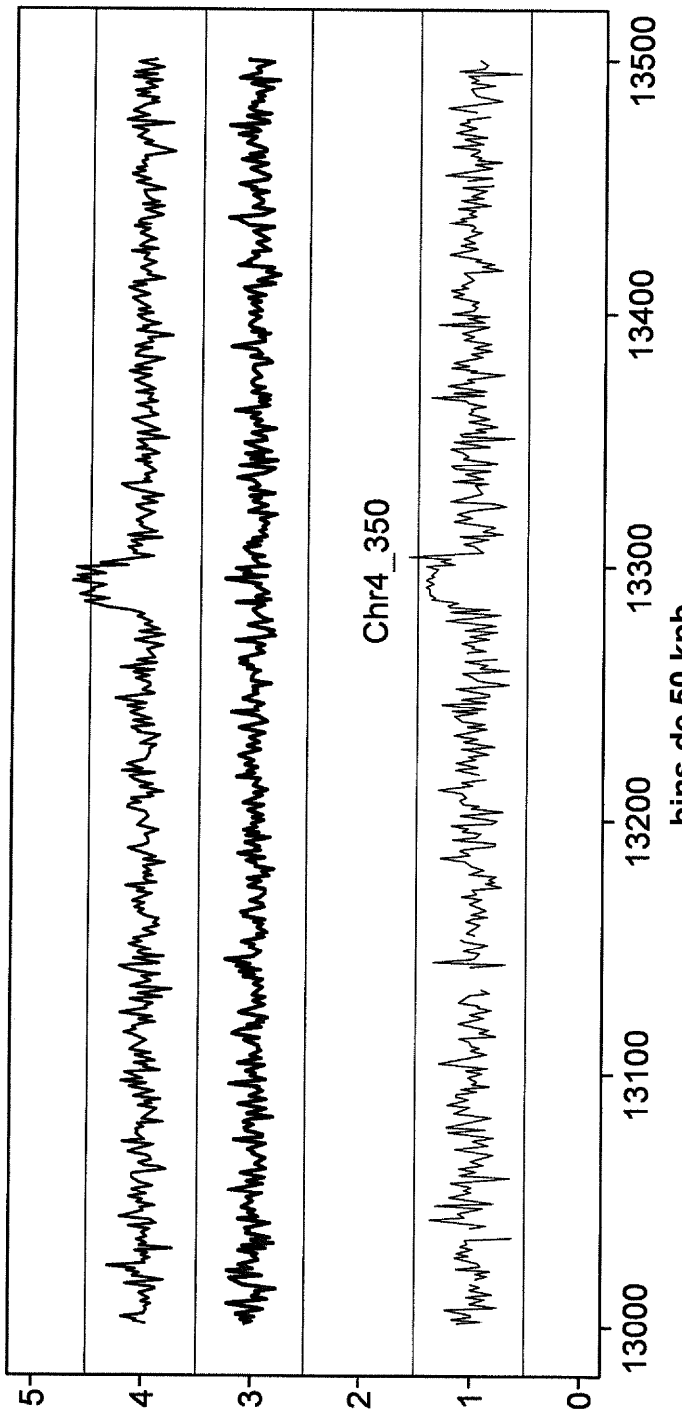

5
4
3
2
1
0
13000
13100
13200
13300
13400
13500
Chr4_350
bins de 50 kpb

FIG. 13

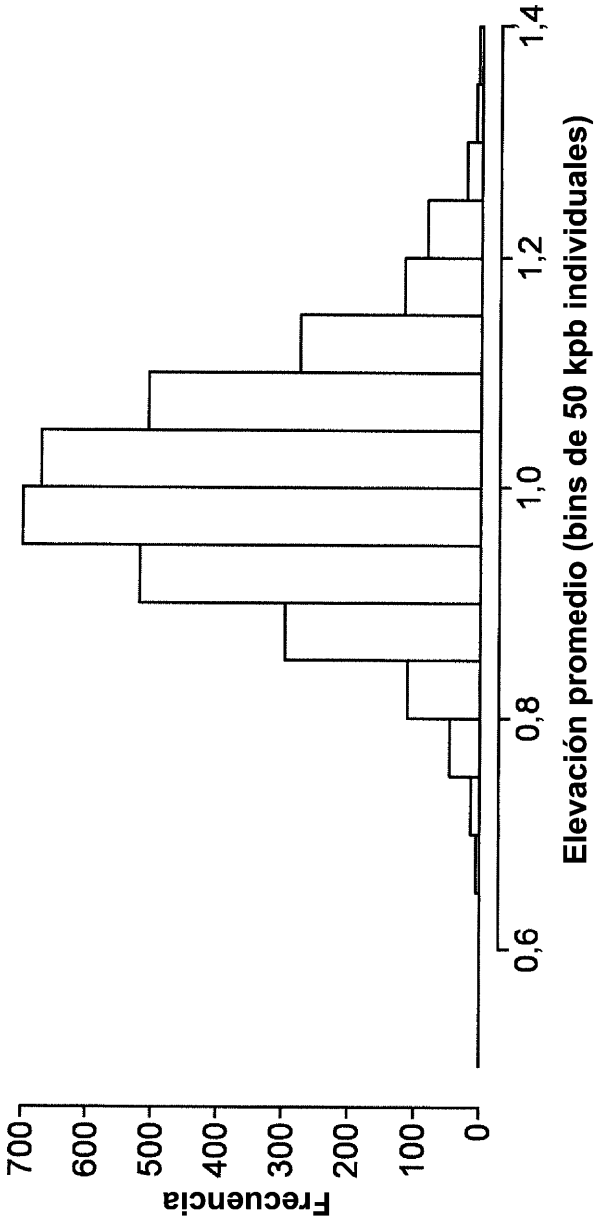

Frecuencia
0
100
200
300
400
500
600
700
0,6
0,8
1,0
1,2
1,4
Elevación promedio (bins de 50 kpb individuales)

FIG. 14

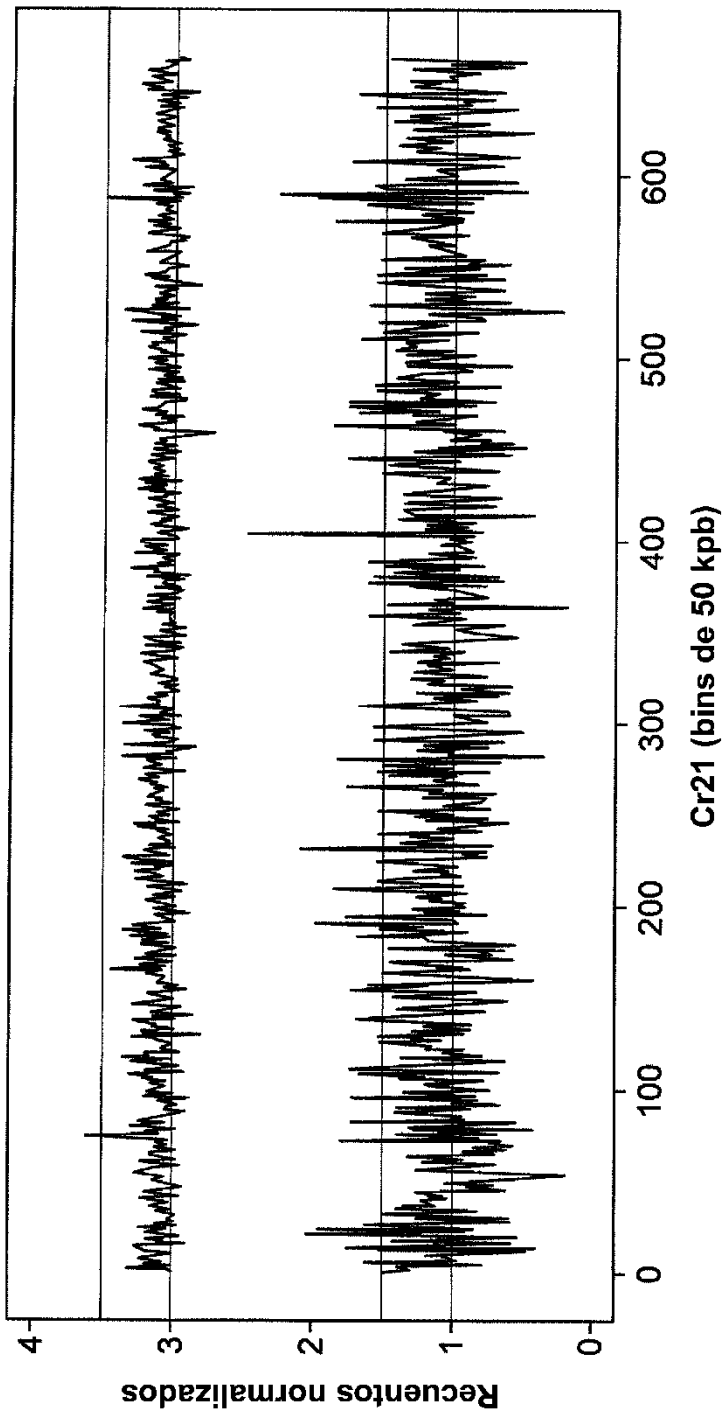

Recuentos normalizados
0
1
2
3
4
0
100
200
300
400
500
600
Cr21 (bins de 50 kpb)

FIG. 15

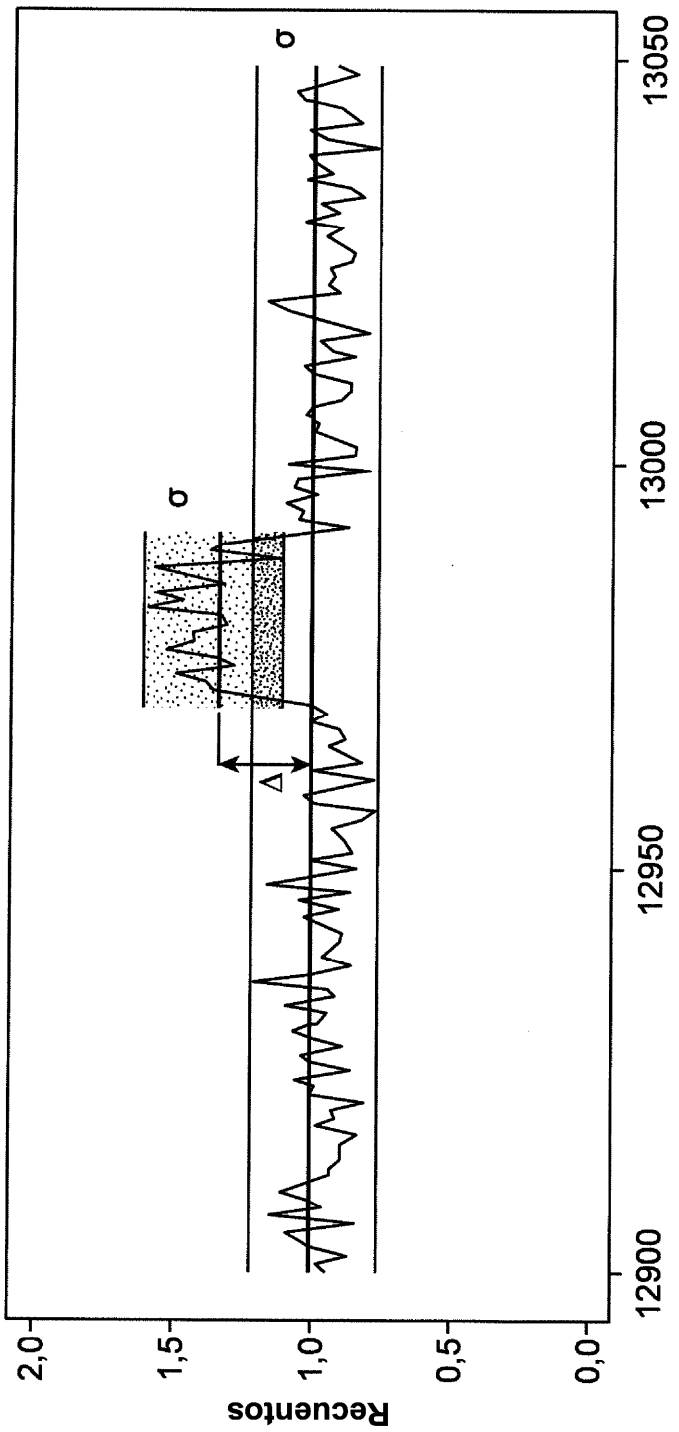

σ
σ
Δ
2,0
1,5
1,0
0,5
0,0
Recuentos
12900
12950
13000
13050
Cr4 (bins de 50 kpb)

FIG. 16

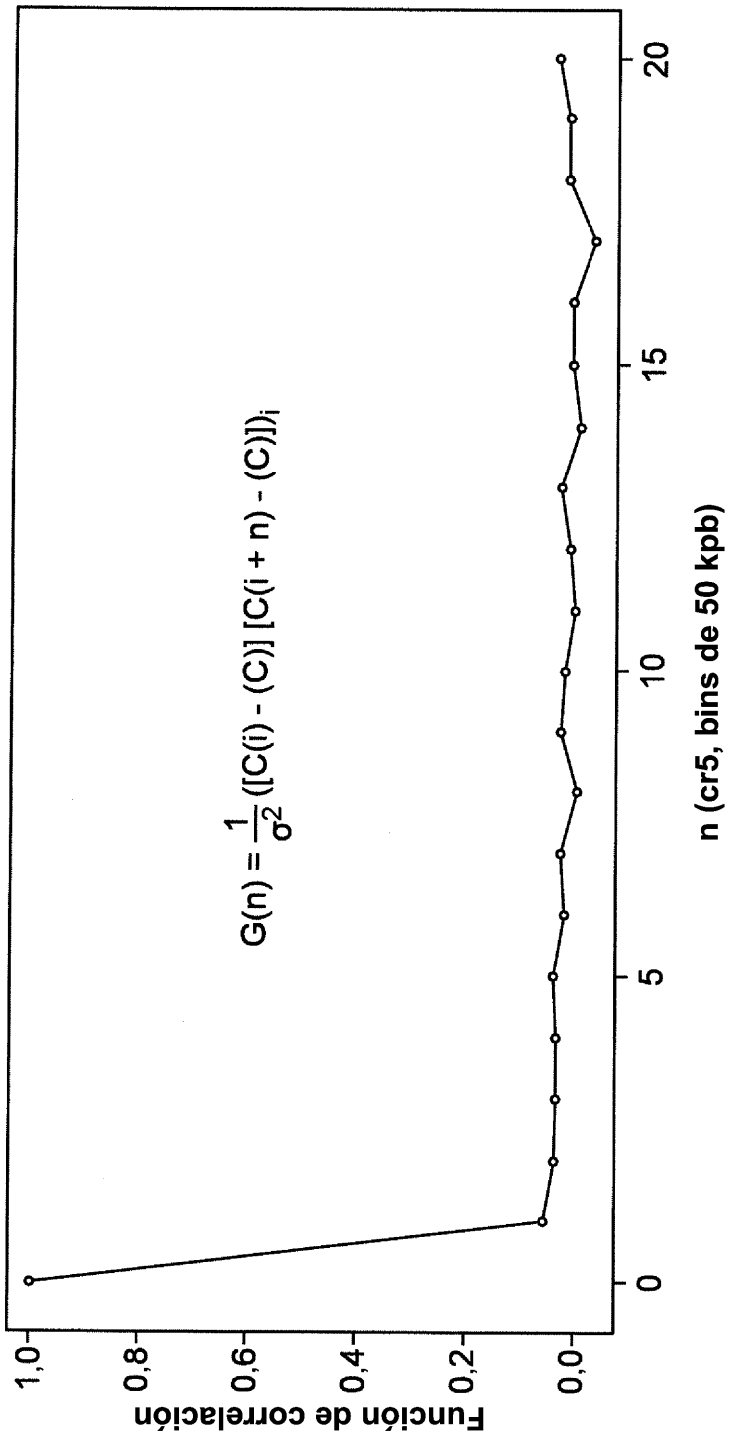

$G(n) = \frac{1}{\sigma^2}([C(i) - (C)][C(i+n) - (C)])_i$
Función de correlación
1,0
0,8
0,6
0,4
0,2
0,0
0
5
10
15
20
n (cr5, bins de 50 kpb)

FIG. 17

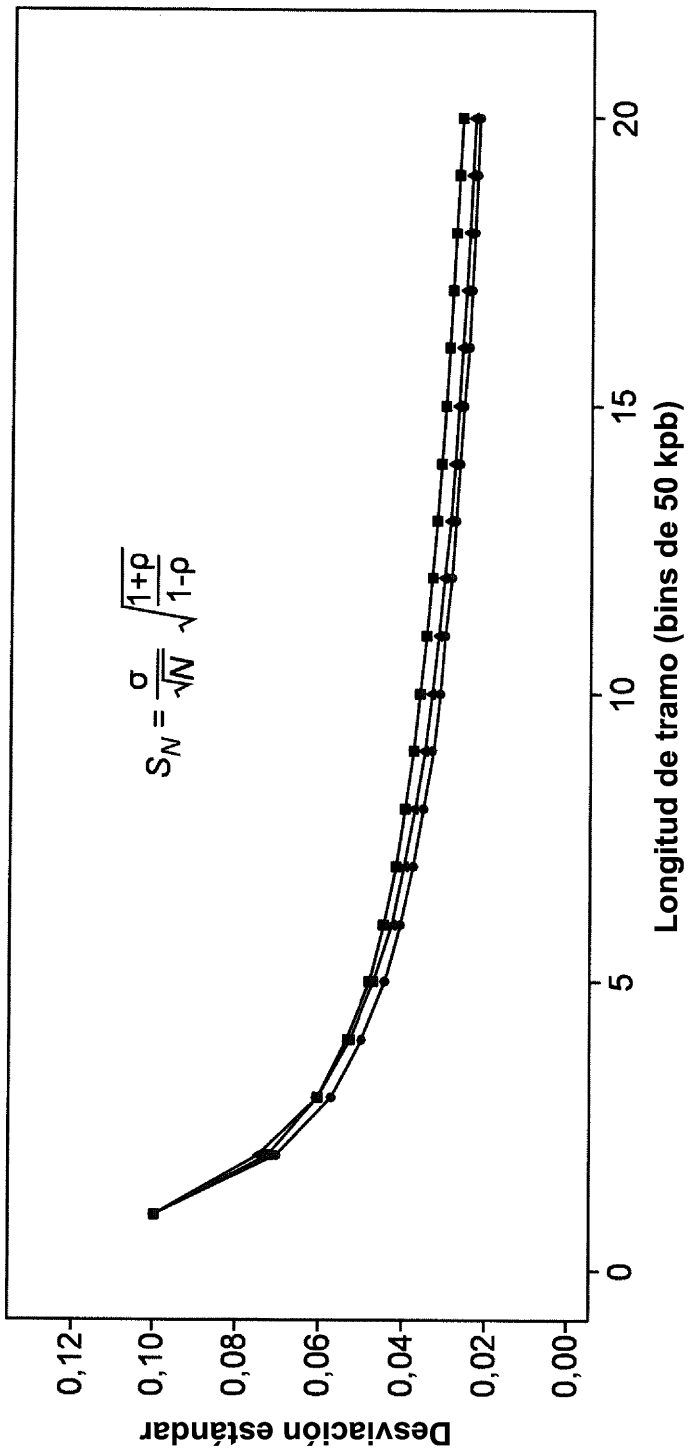

Desviación estándar
0,00
0,02
0,04
0,06
0,08
0,10
0,12
0
5
10
15
20
Longitud de tramo (bins de 50 kpb)
$S_N = \frac{\sigma}{\sqrt{N}} \sqrt{\frac{1+\rho}{1-\rho}}$

FIG. 18

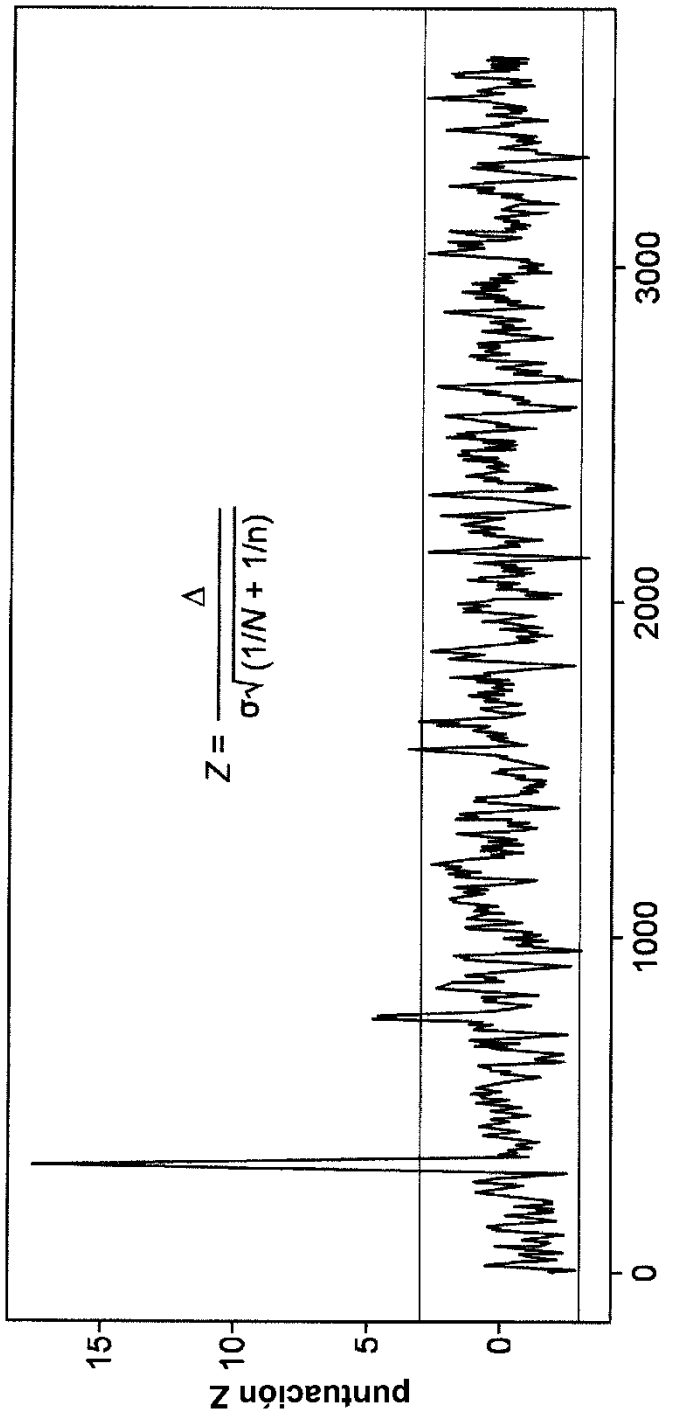

puntuación Z
15
10
5
0
0
1000
2000
3000
Cr4 (bins de 50 kpb)
$Z = \frac{\Delta}{\sigma\sqrt{(1/N + 1/n)}}$

FIG. 19

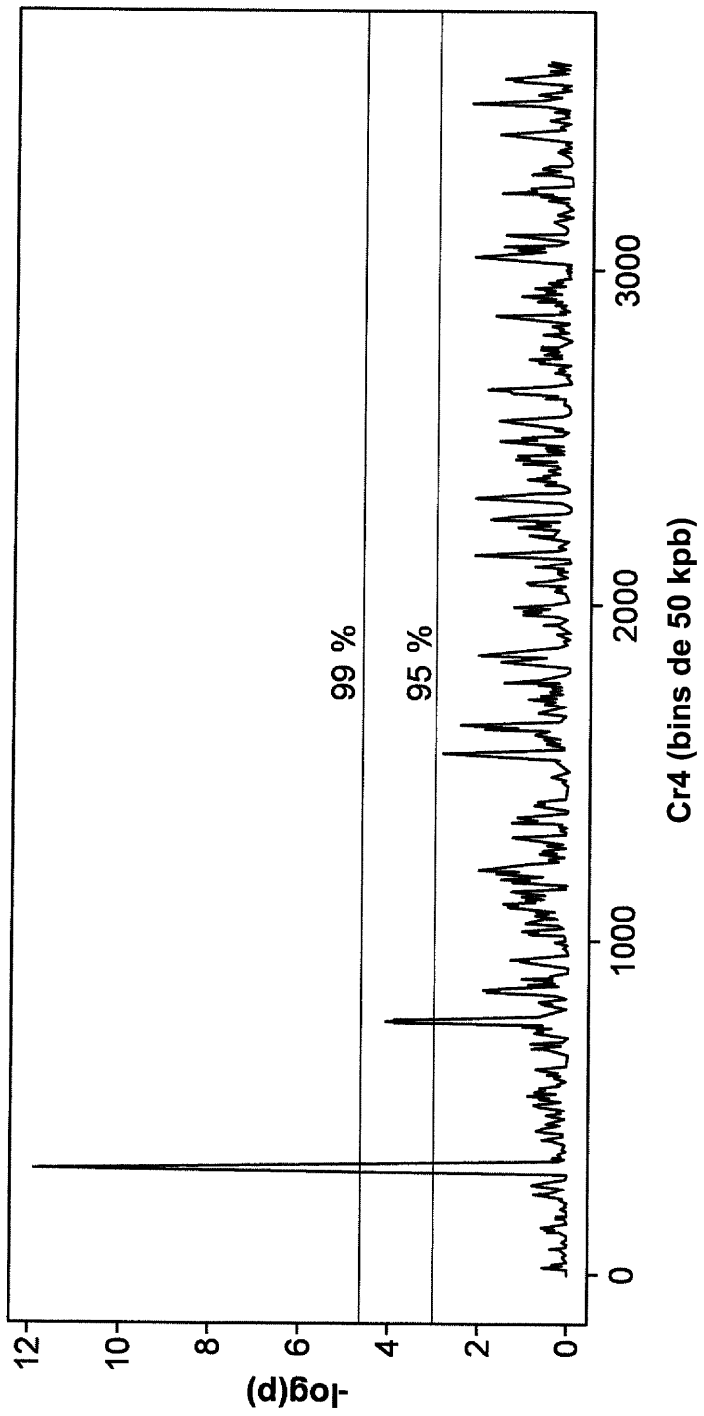


FIG. 20

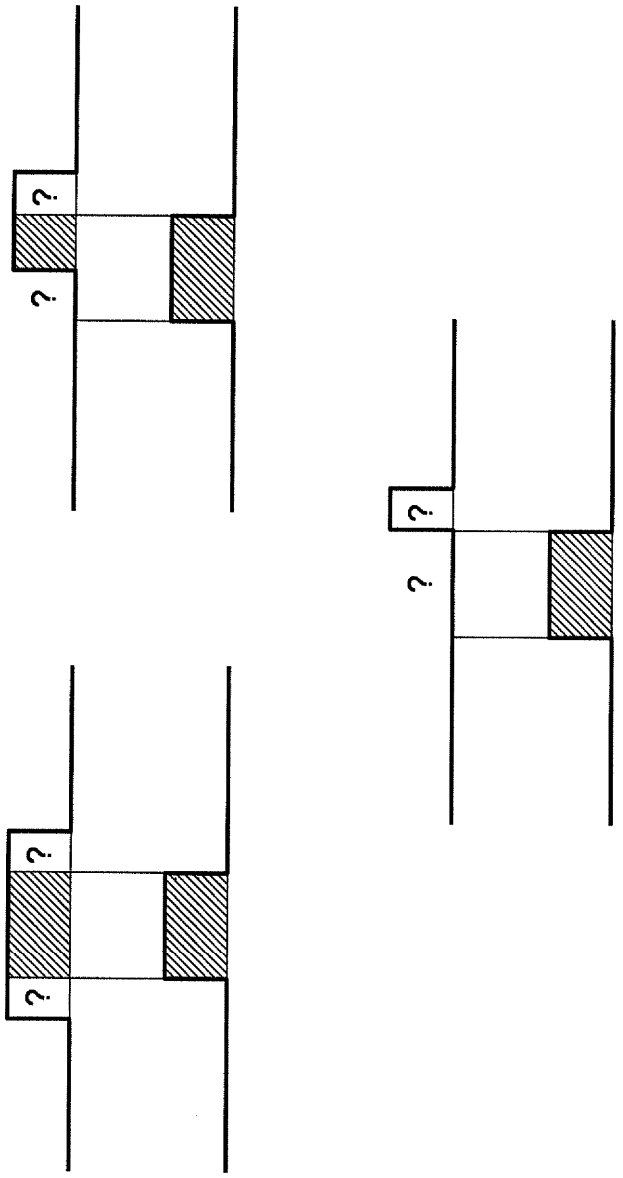


FIG. 21

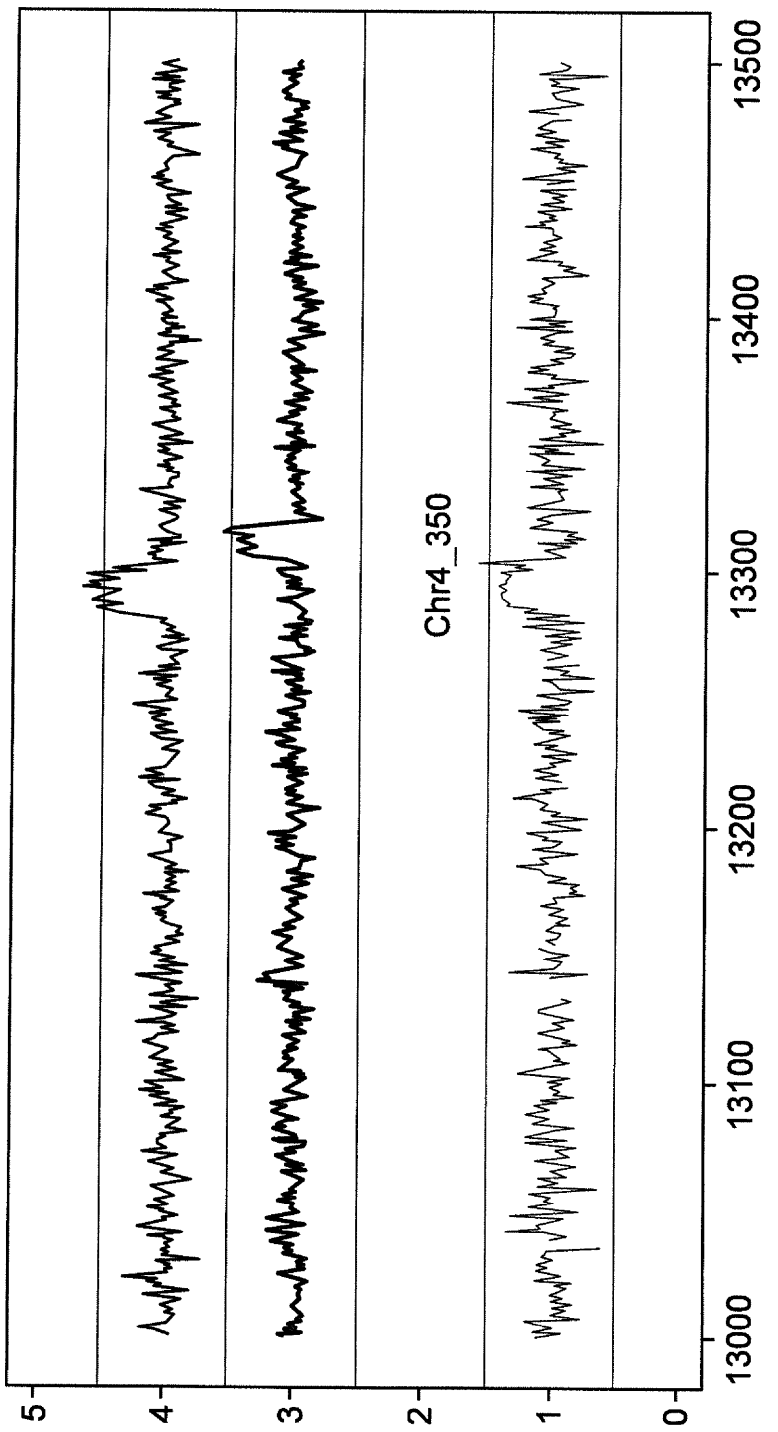

5
4
3
2
1
0
13000
13100
13200
13300
13400
13500
Chr4_350
Cr4 (bins de 50 kpb)

FIG. 22

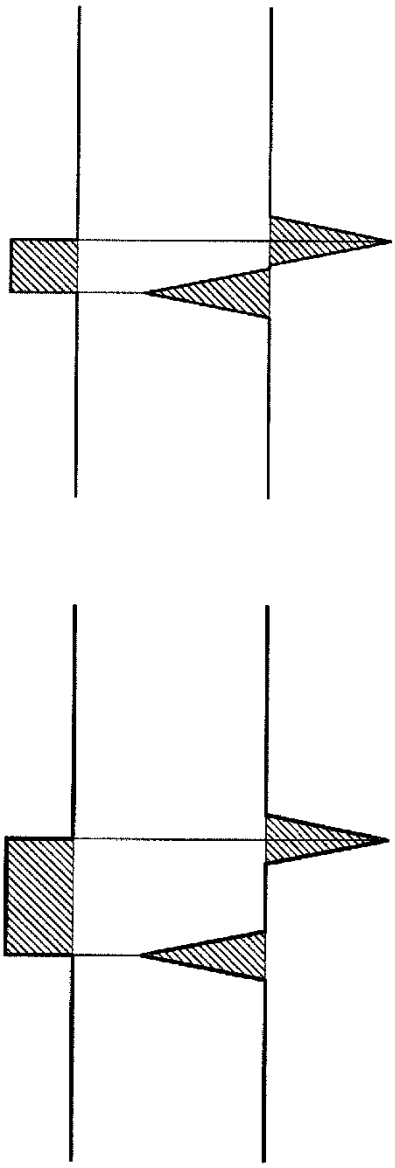


FIG. 23

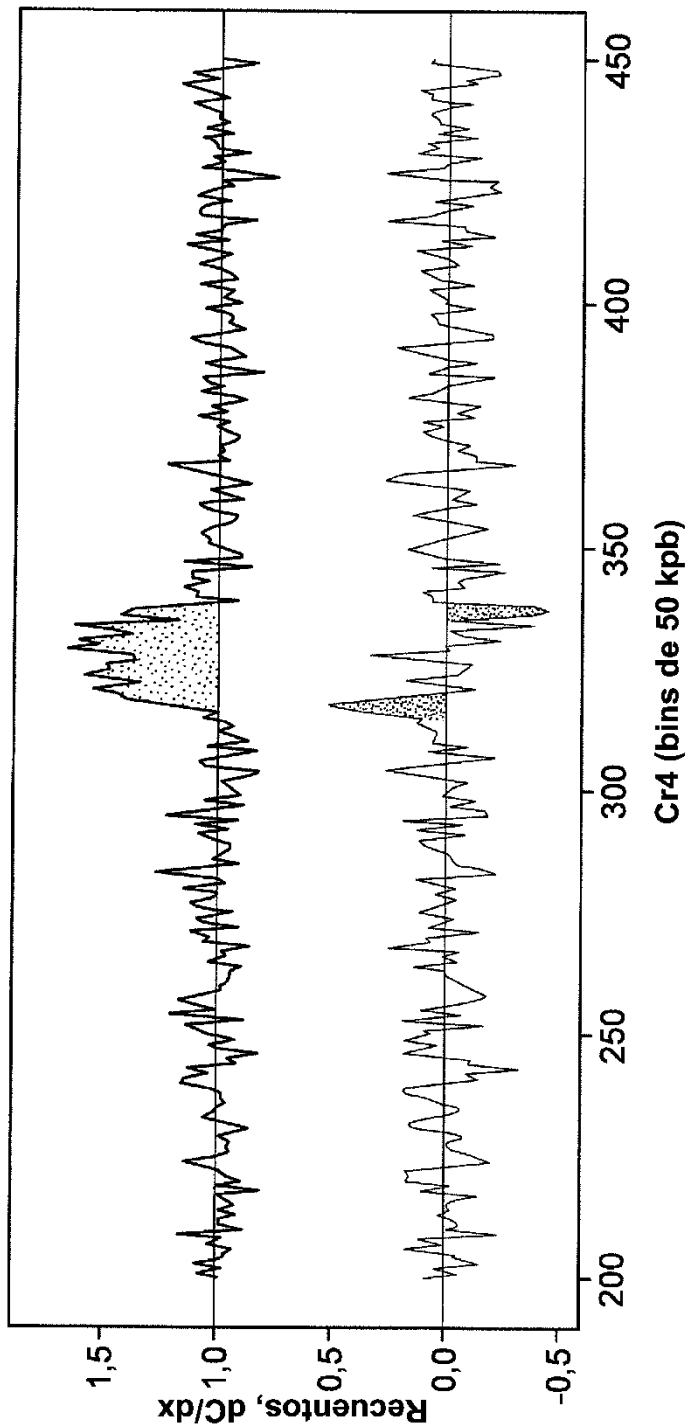

Recuentos, dC/dx
1,5
1,0
0,5
0,0
-0,5
200
250
300
350
400
450
Cr4 (bins de 50 kpb)

FIG. 24

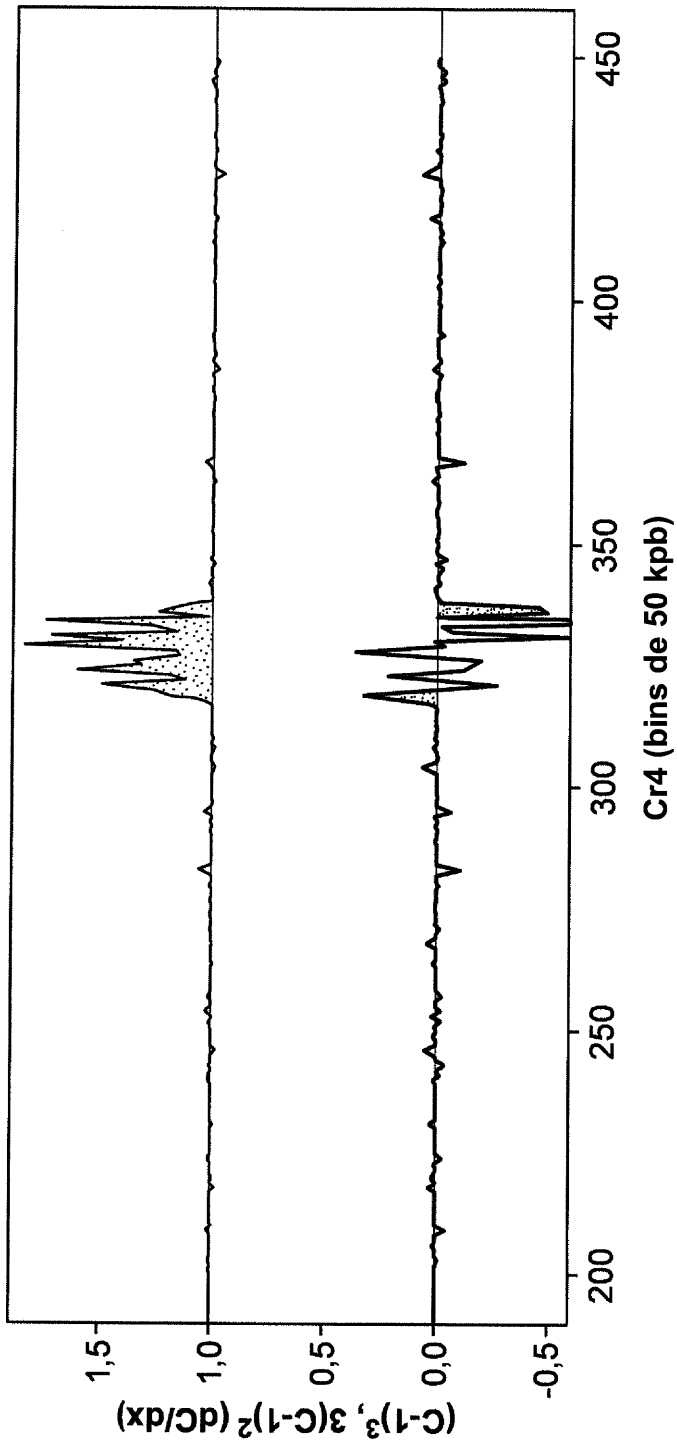

(C-1)3, 3(C-1)2 (dC/dx)
1,5
1,0
0,5
0,0
-0,5
200
250
300
350
400
450
Cr4 (bins de 50 kpb)

FIG. 25

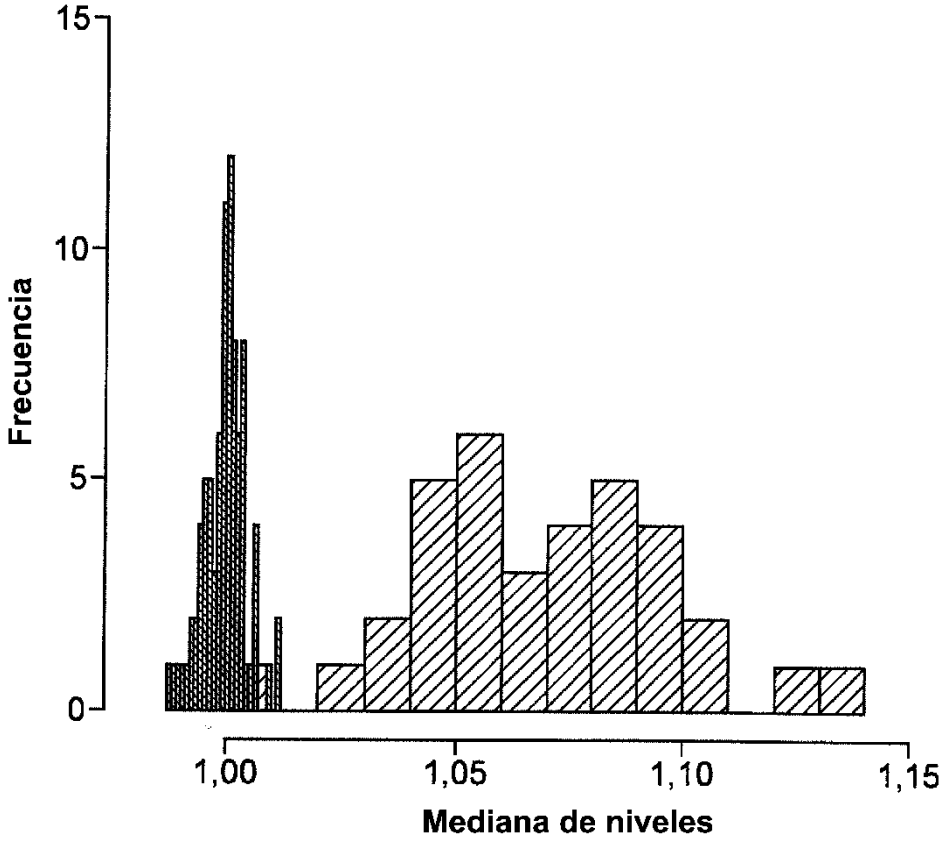


FIG. 26

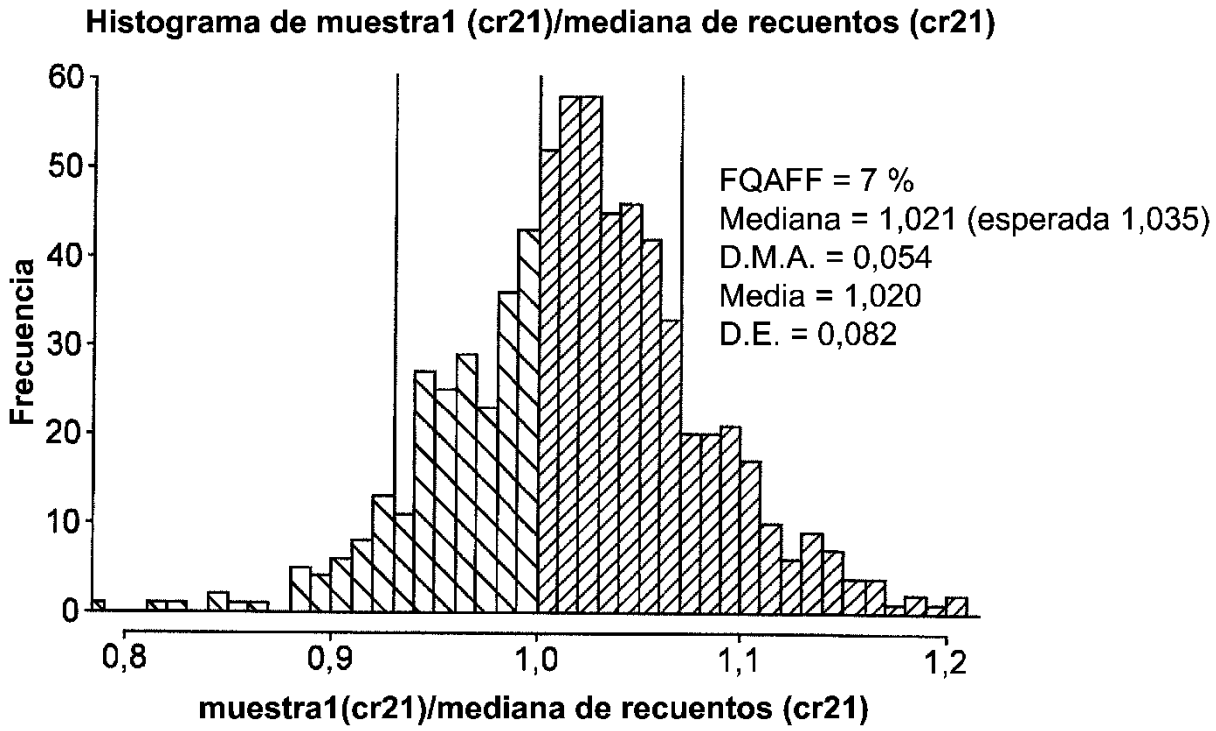

Histograma de muestra1 (cr21)/mediana de recuentos (cr21)
Frecuencia
60
50
40
30
20
10
0
0,8
0,9
1,0
1,1
1,2
muestra1(cr21)/mediana de recuentos (cr21)
FQAFF = 7 %
Mediana = 1,021 (esperada 1,035)
D.M.A. = 0,054
Media = 1,020
D.E. = 0,082

FIG. 27

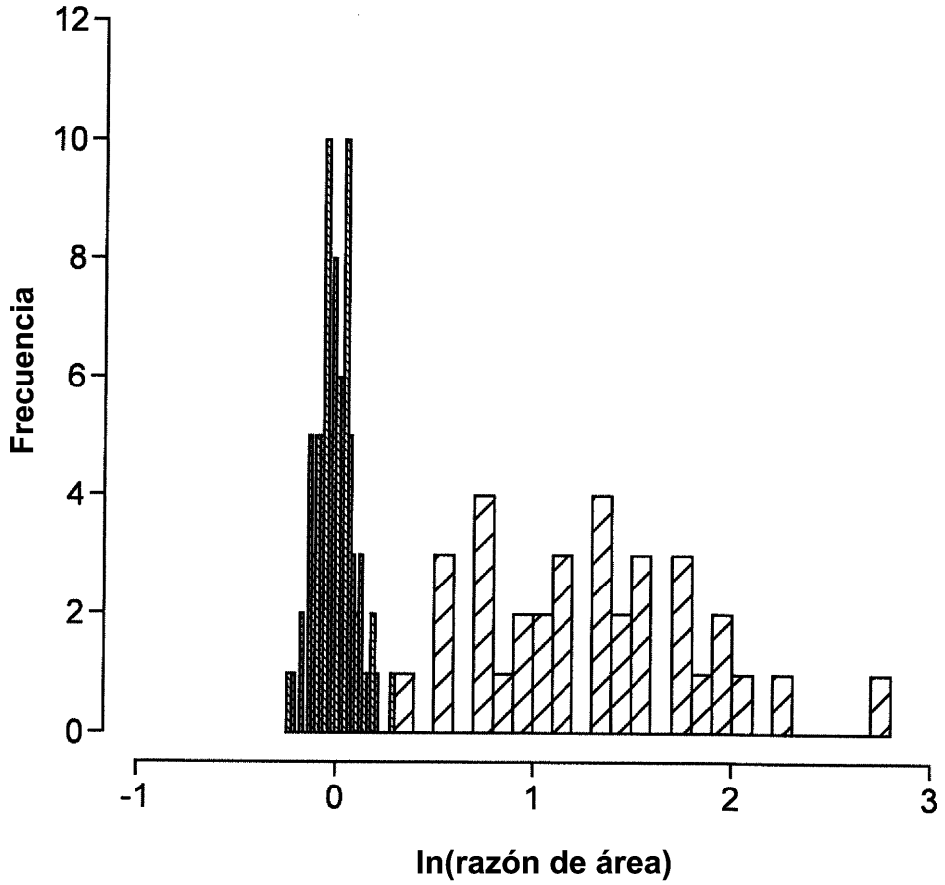


FIG. 28

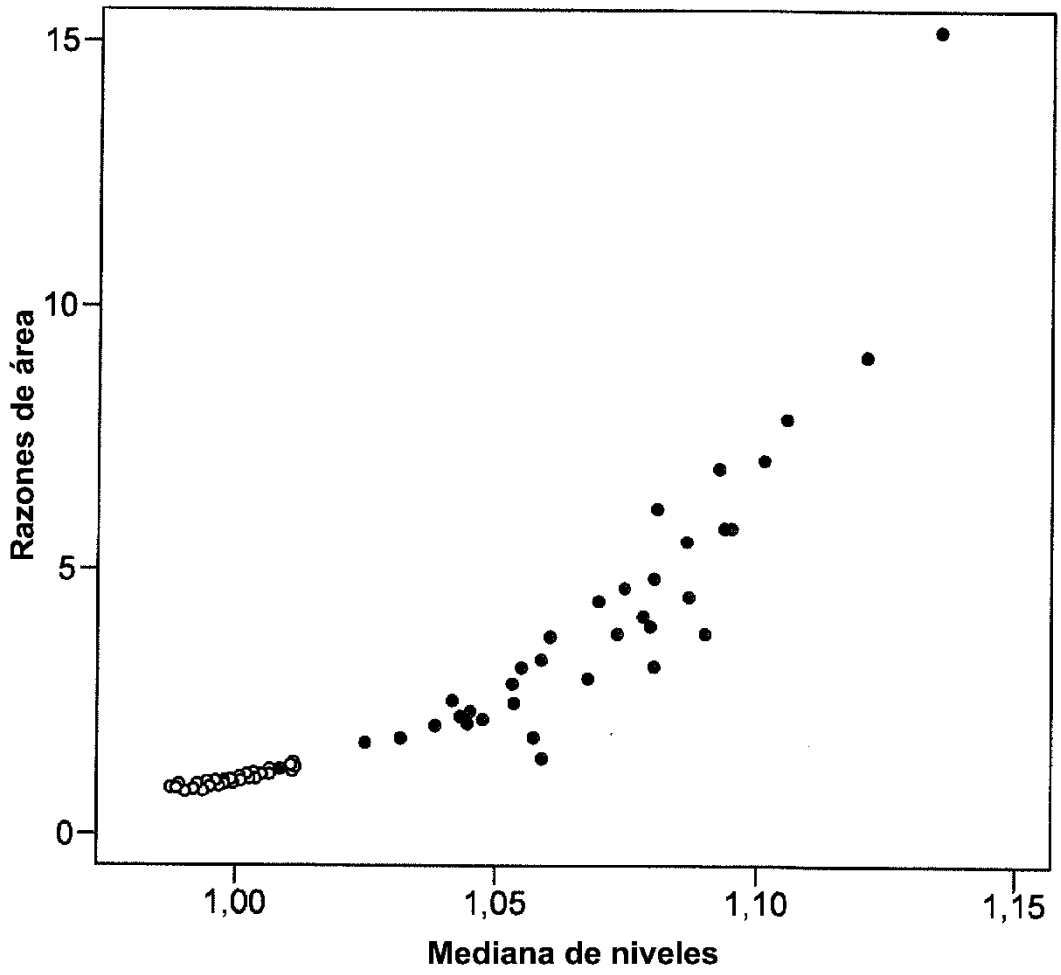


FIG. 29

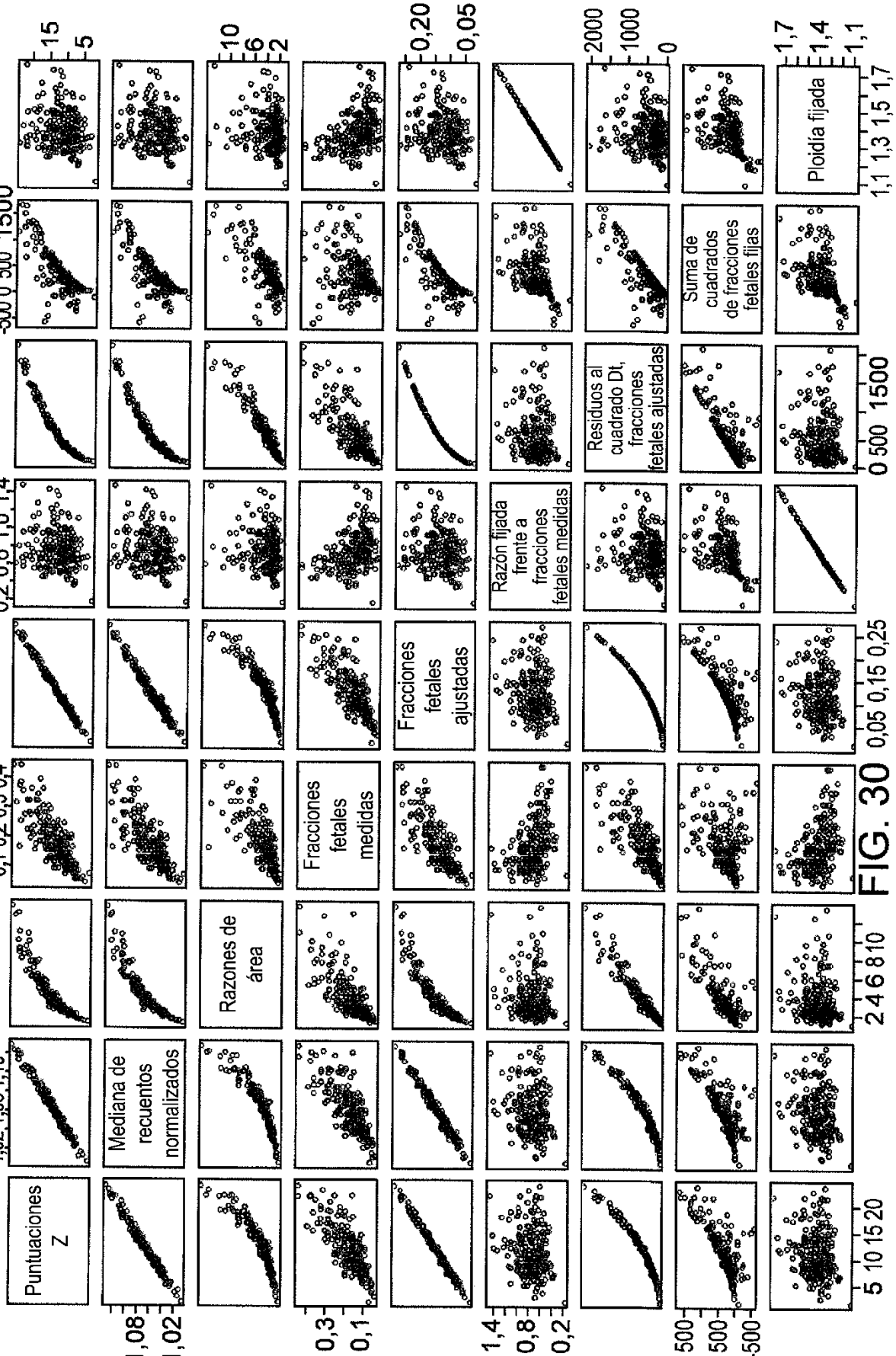

Puntuaciones Z
Mediana de recuentos normalizados
Razones de área
Fracciones fetales medidas
Fracciones fetales ajustadas
Razón fijada frente a fracciones fetales medidas
Residuos al cuadrado Dt, fracciones fetales ajustadas
Suma de cuadrados de fracciones fetales fijas
Ploidía fijada

FIG. 30

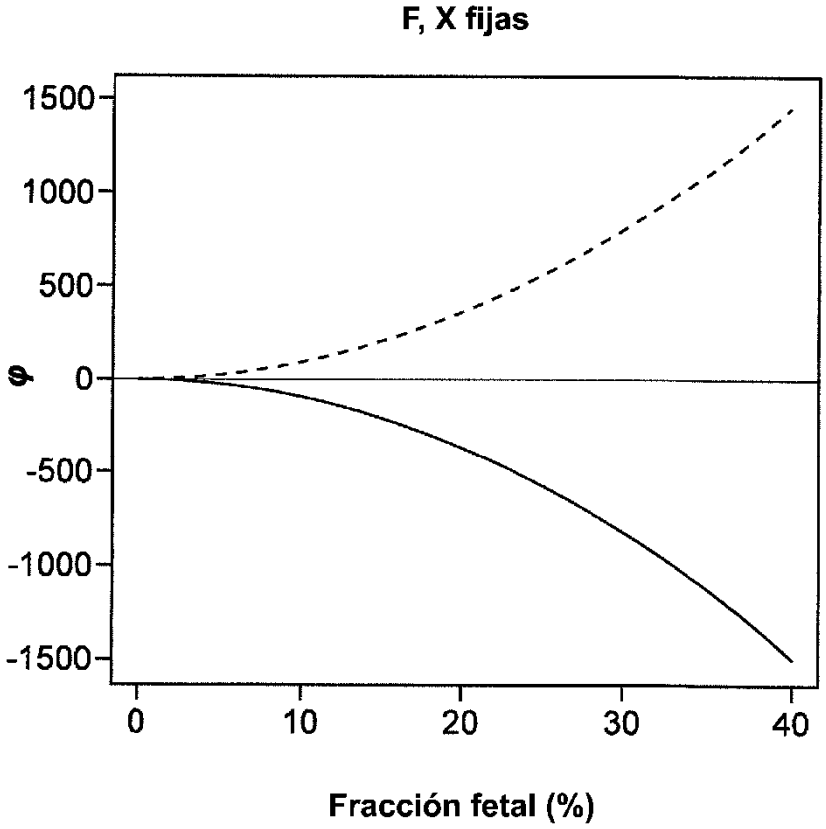

F, X fijas
1500
1000
500
S 0
-500
-1000
-1500
0
10
20
30
40
Fracción fetal (%)

FIG. 31

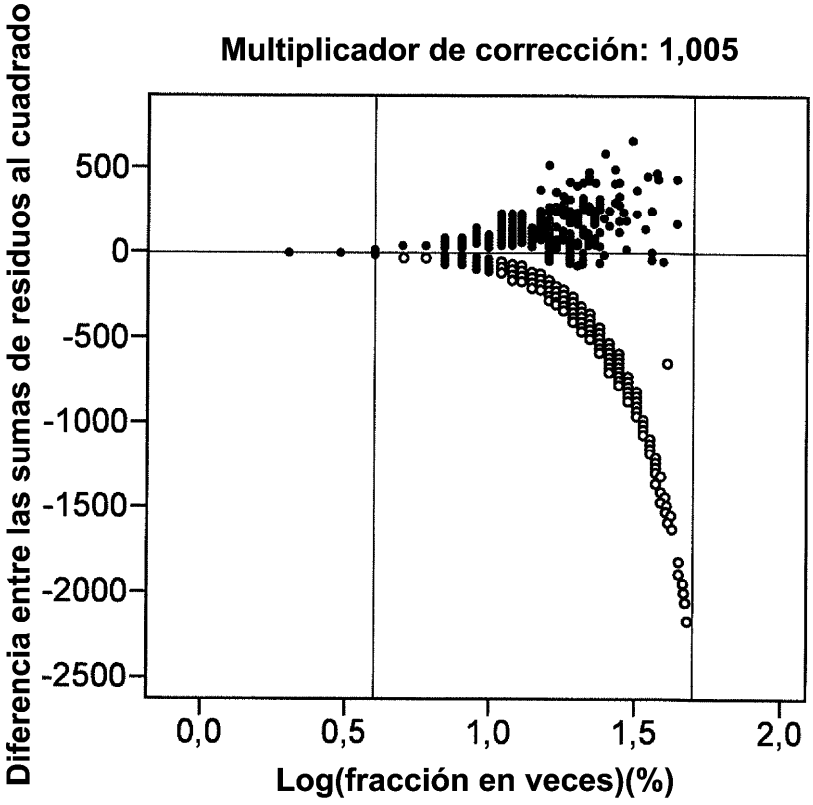

Multiplicador de corrección: 1,005
Diferencia entre las sumas de residuos al cuadrado
500
0
-500
-1000
-1500
-2000
-2500
0,0
0,5
1,0
1,5
2,0
Log(fracción en veces)(%)

FIG. 32

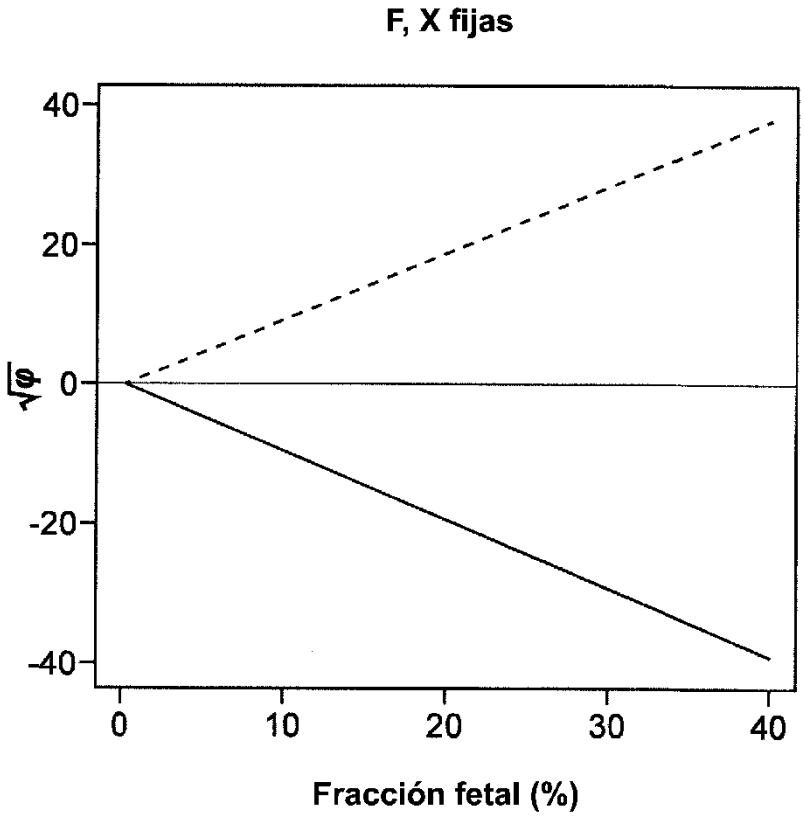


FIG. 33

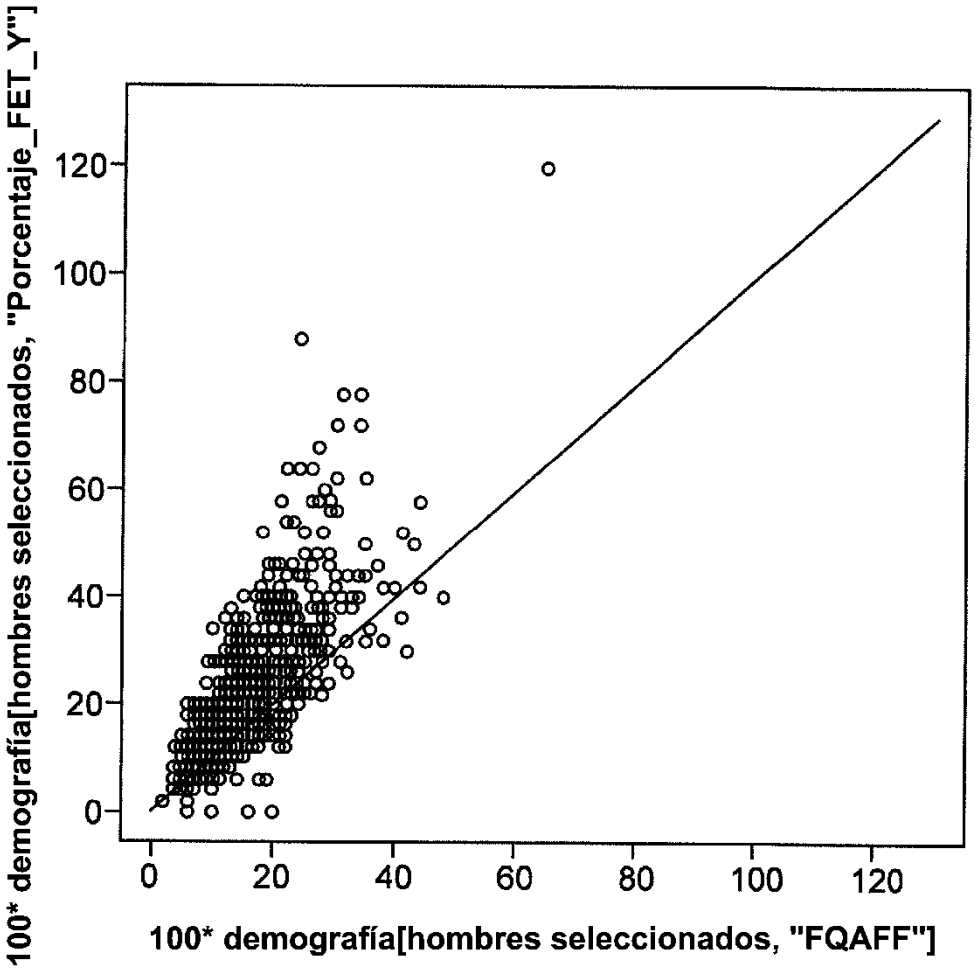

100* demografía[hombres seleccionados, "Porcentaje_FET_Y"]
120
100
80
60
40
20
0
0
20
40
60
80
100
120
100* demografía[hombres seleccionados, "FQAFF"]

FIG. 34

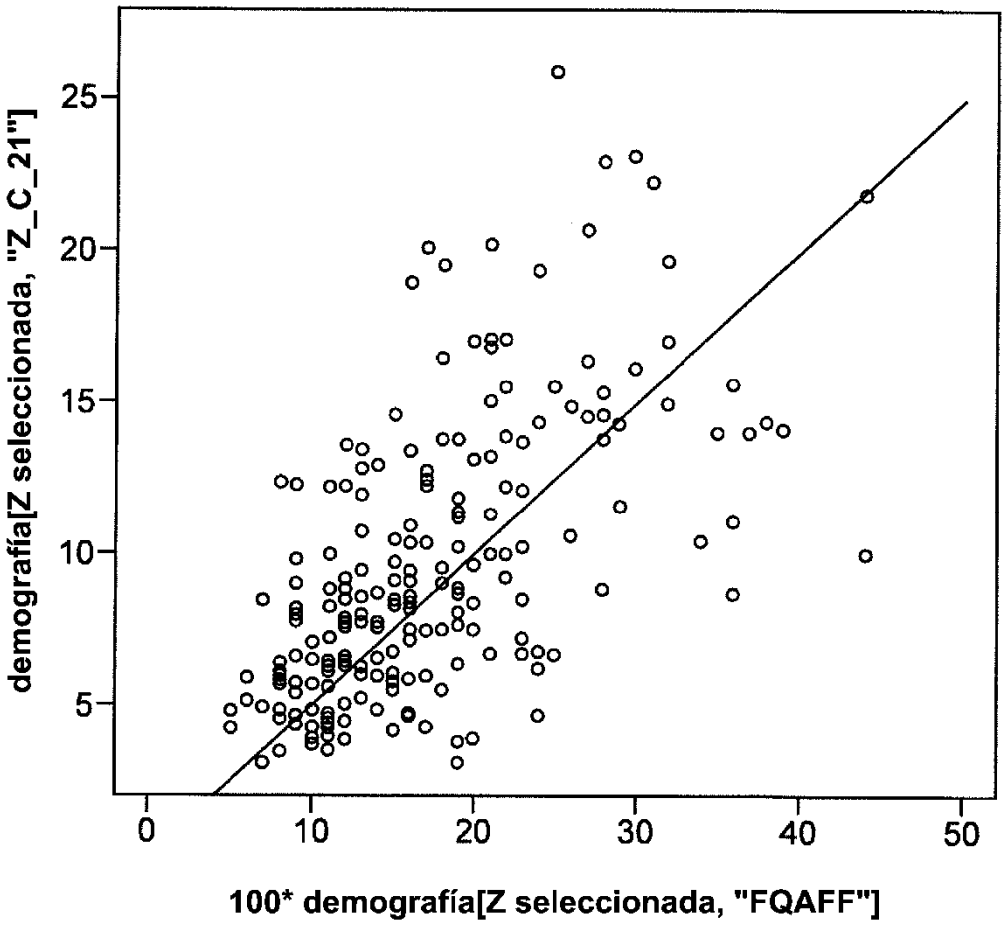

demografía[Z seleccionada, "Z_C_21"]
25
20
15
10
5
0
10
20
30
40
50
100* demografía[Z seleccionada, "FQAFF"]

FIG. 35

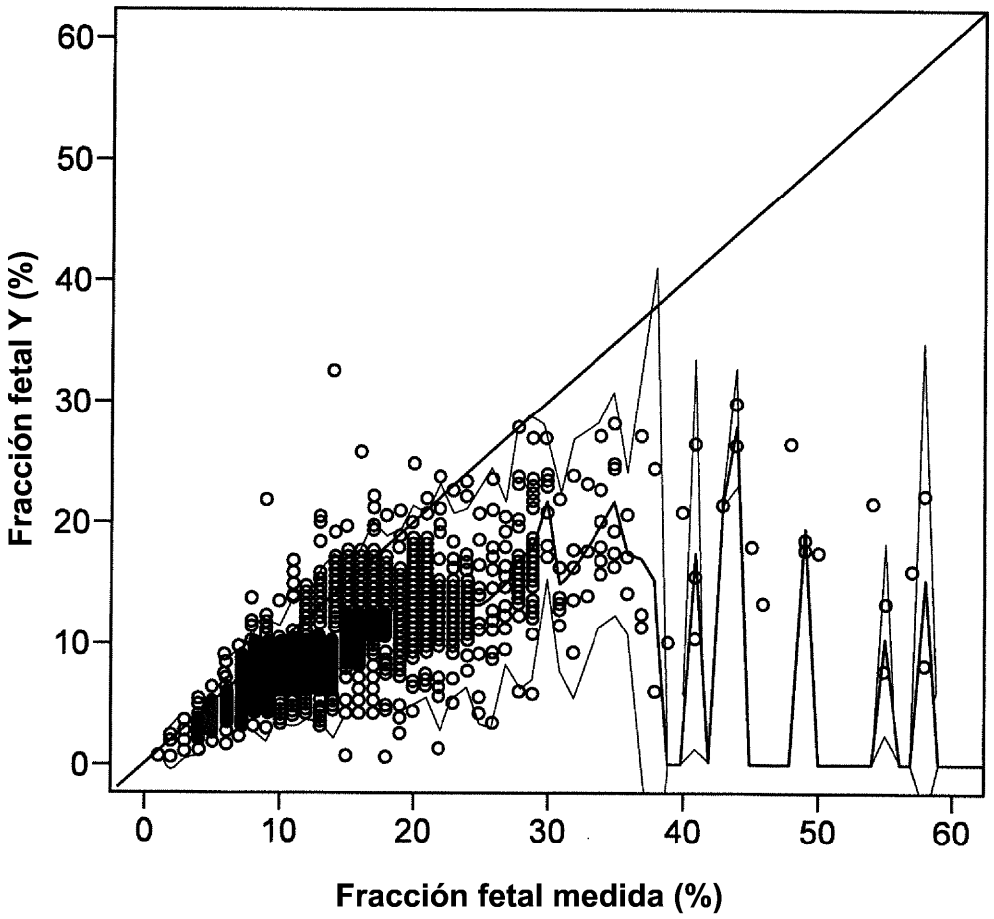
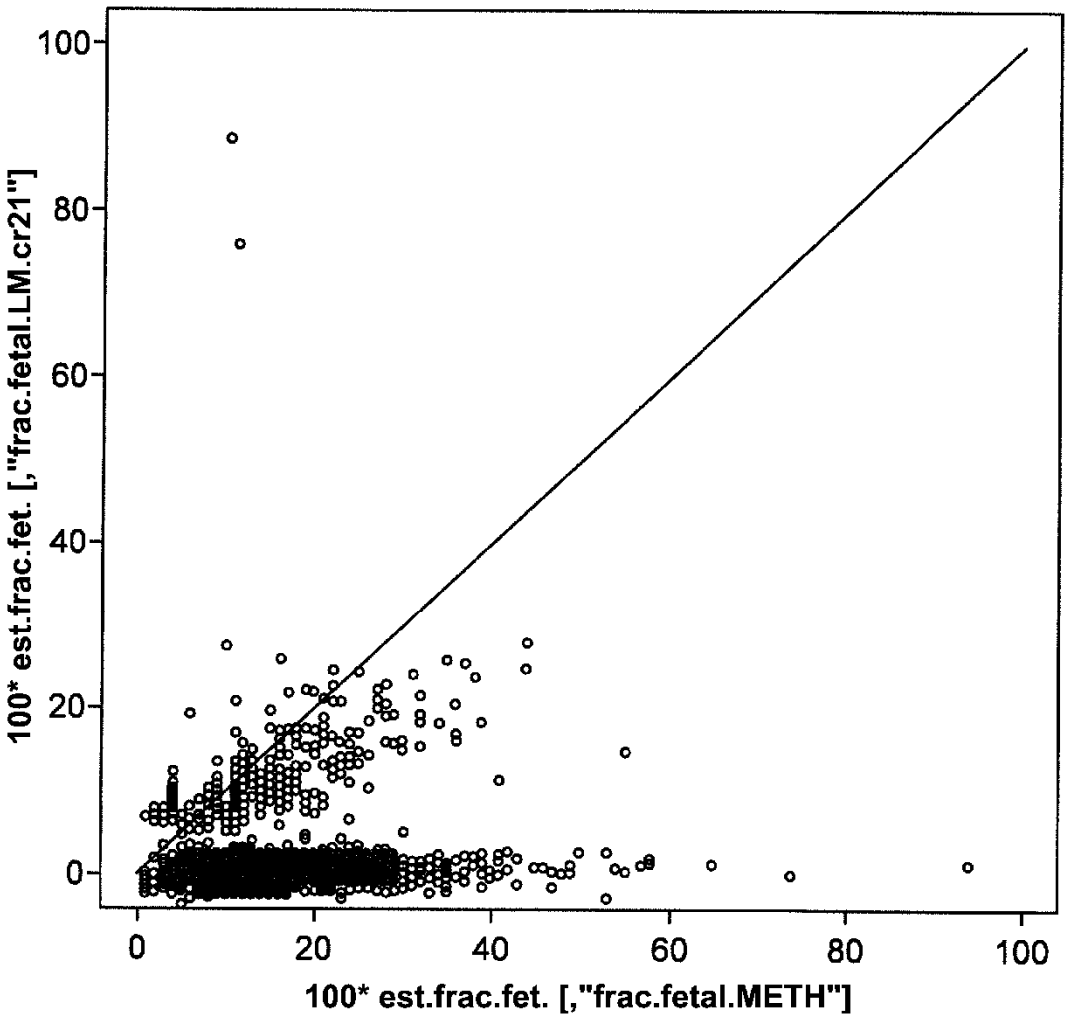

60
50
40
30
20
10
0
Fracción fetal Y (%)
0
10
20
30
40
50
60
Fracción fetal medida (%)

FIG. 36



Estimaciones de fracción fetal basadas en cr. 21 frente a fracciones fetales medidas.

FIG. 37

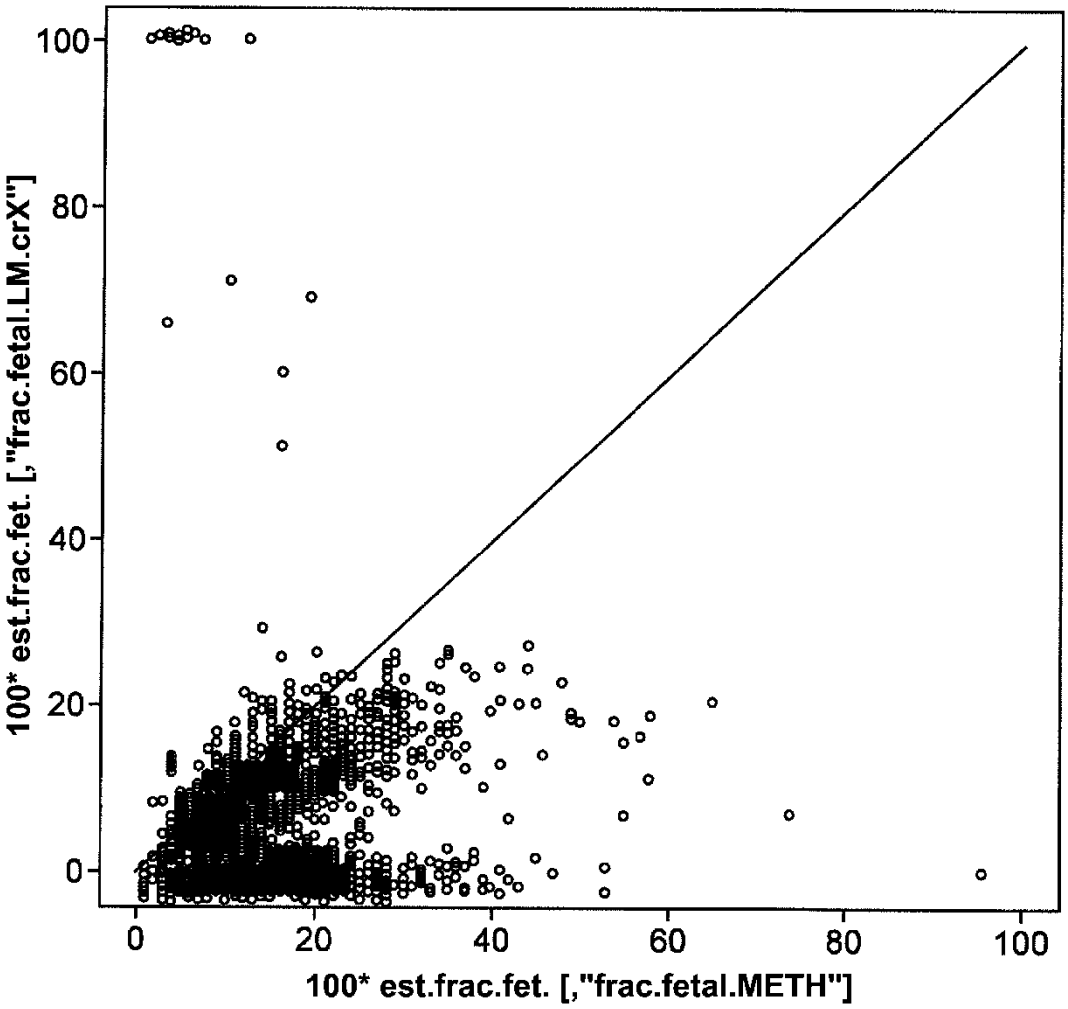

100
80
60
40
20
0
100* est.frac.fet. [,"frac.fetal.LM.crX"]
0
20
40
60
80
100
100* est.frac.fet. [,"frac.fetal.METH"]

FIG. 38

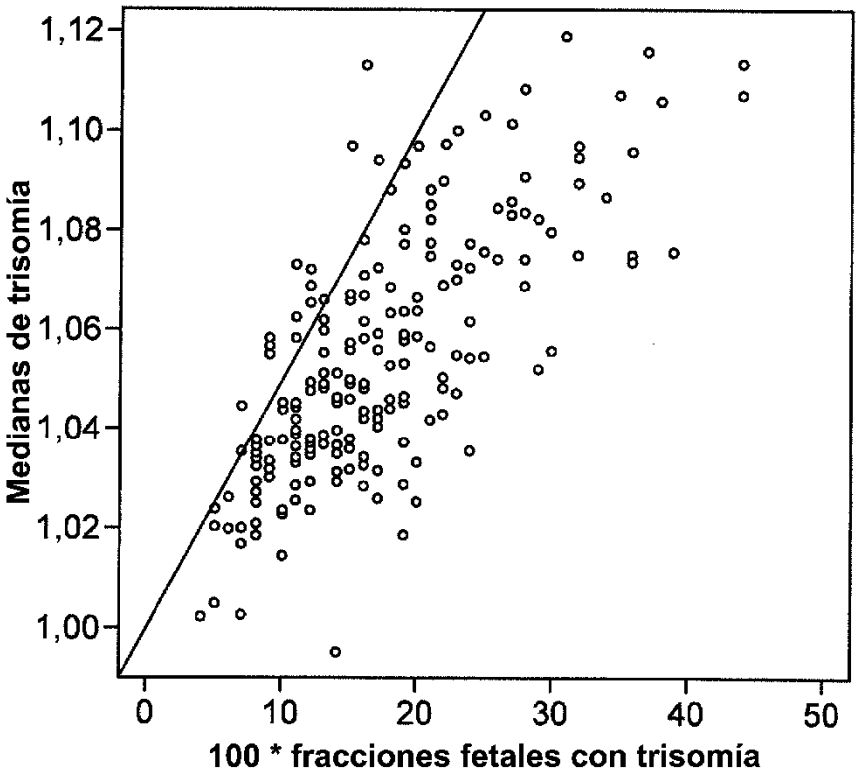

Medianas de trisomía
1,12
1,10
1,08
1,06
1,04
1,02
1,00
0
10
20
30
40
50
100 * fracciones fetales con trisomía

FIG. 39

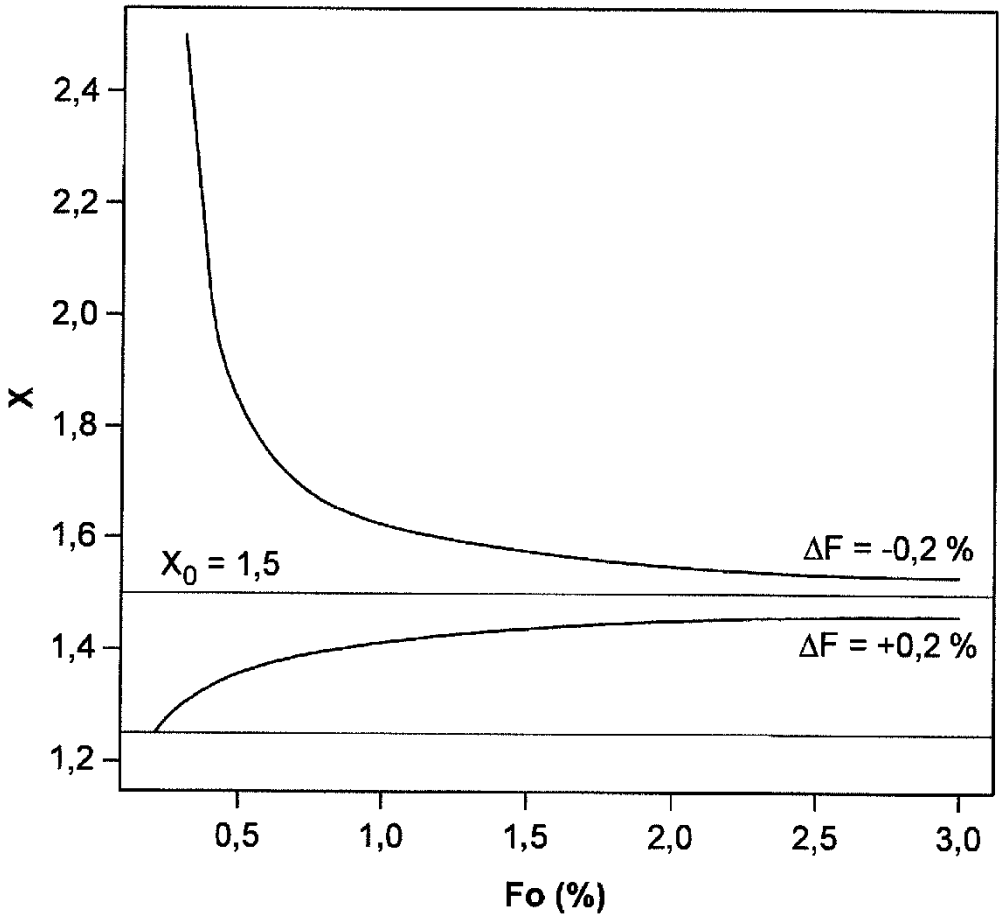


FIG. 40

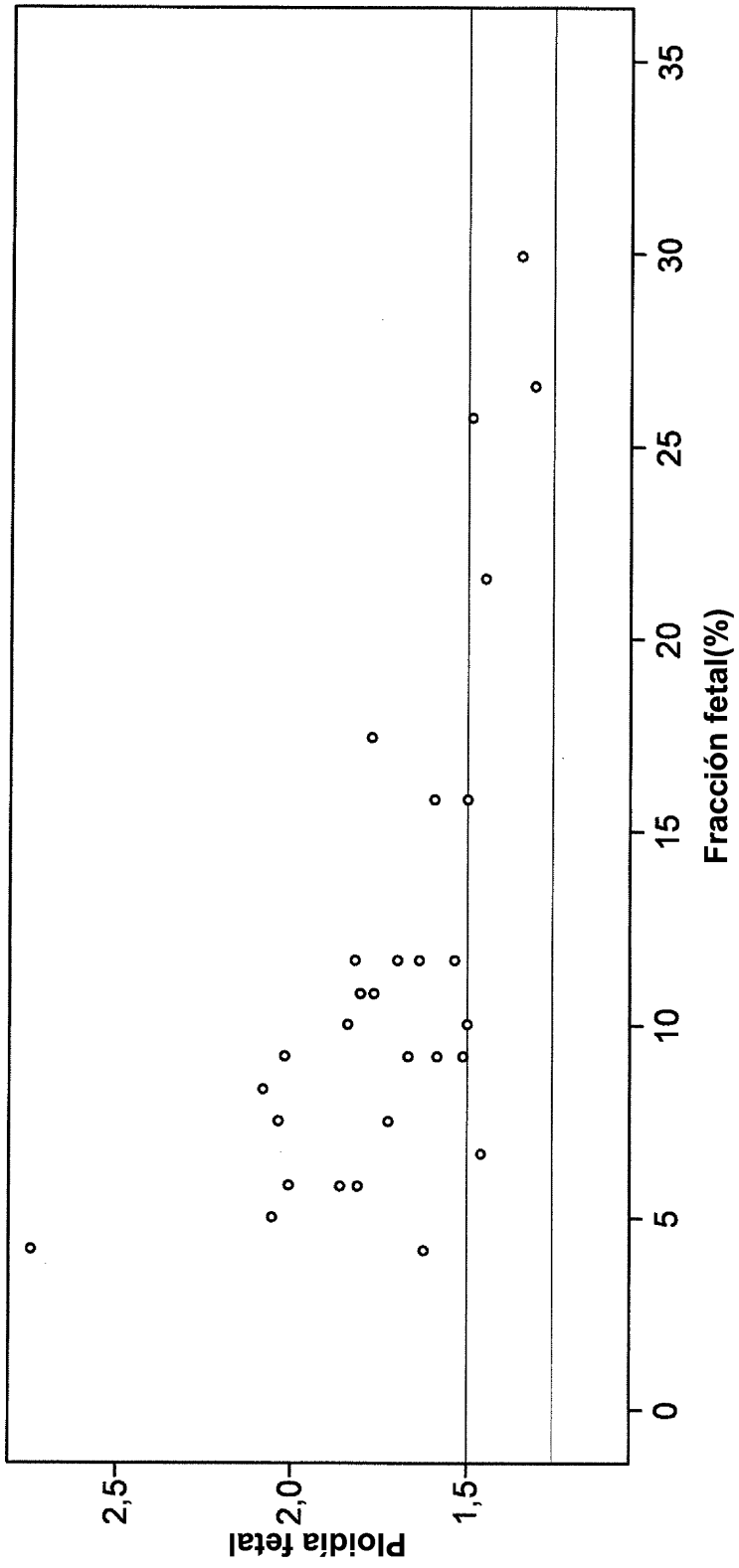


FIG. 41

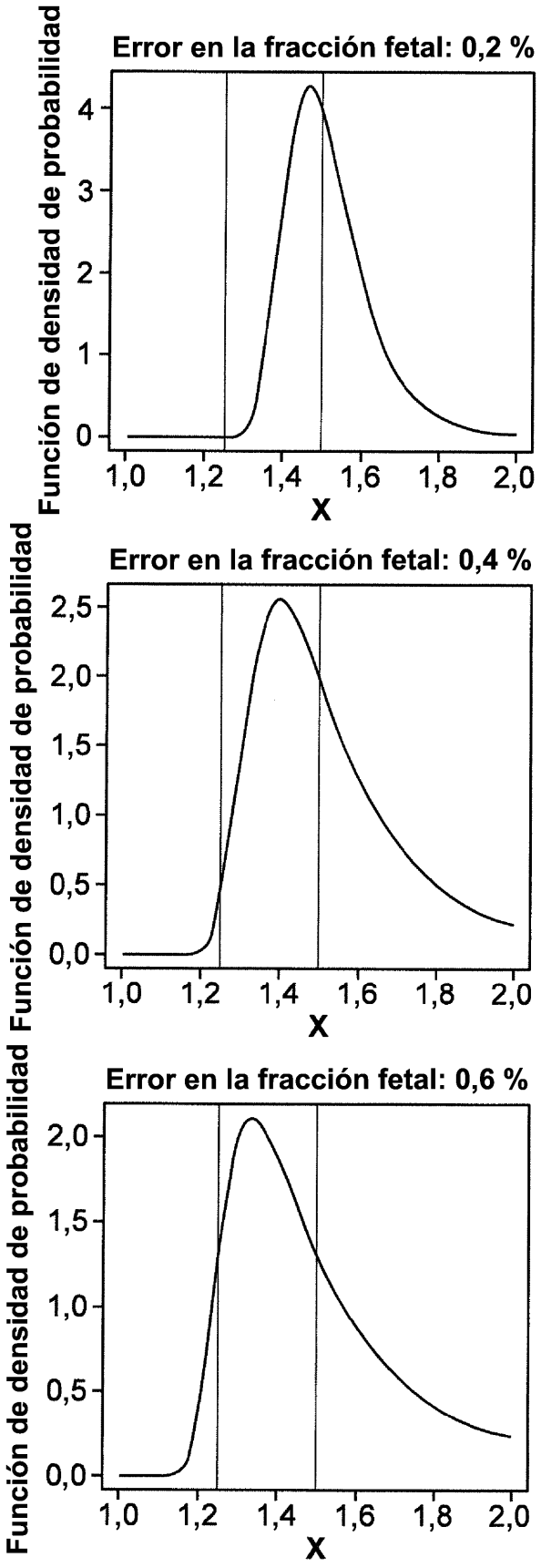

Error en la fracción fetal: 0,2 %
Función de densidad de probabilidad
0
1
2
3
4
1,0
1,2
1,4
1,6
1,8
2,0
X
Error en la fracción fetal: 0,4 %
Función de densidad de probabilidad
0,0
0,5
1,0
1,5
2,0
2,5
1,0
1,2
1,4
1,6
1,8
2,0
X
Error en la fracción fetal: 0,6 %
Función de densidad de probabilidad
0,0
0,5
1,0
1,5
2,0
1,0
1,2
1,4
1,6
1,8
2,0
X

FIG. 42

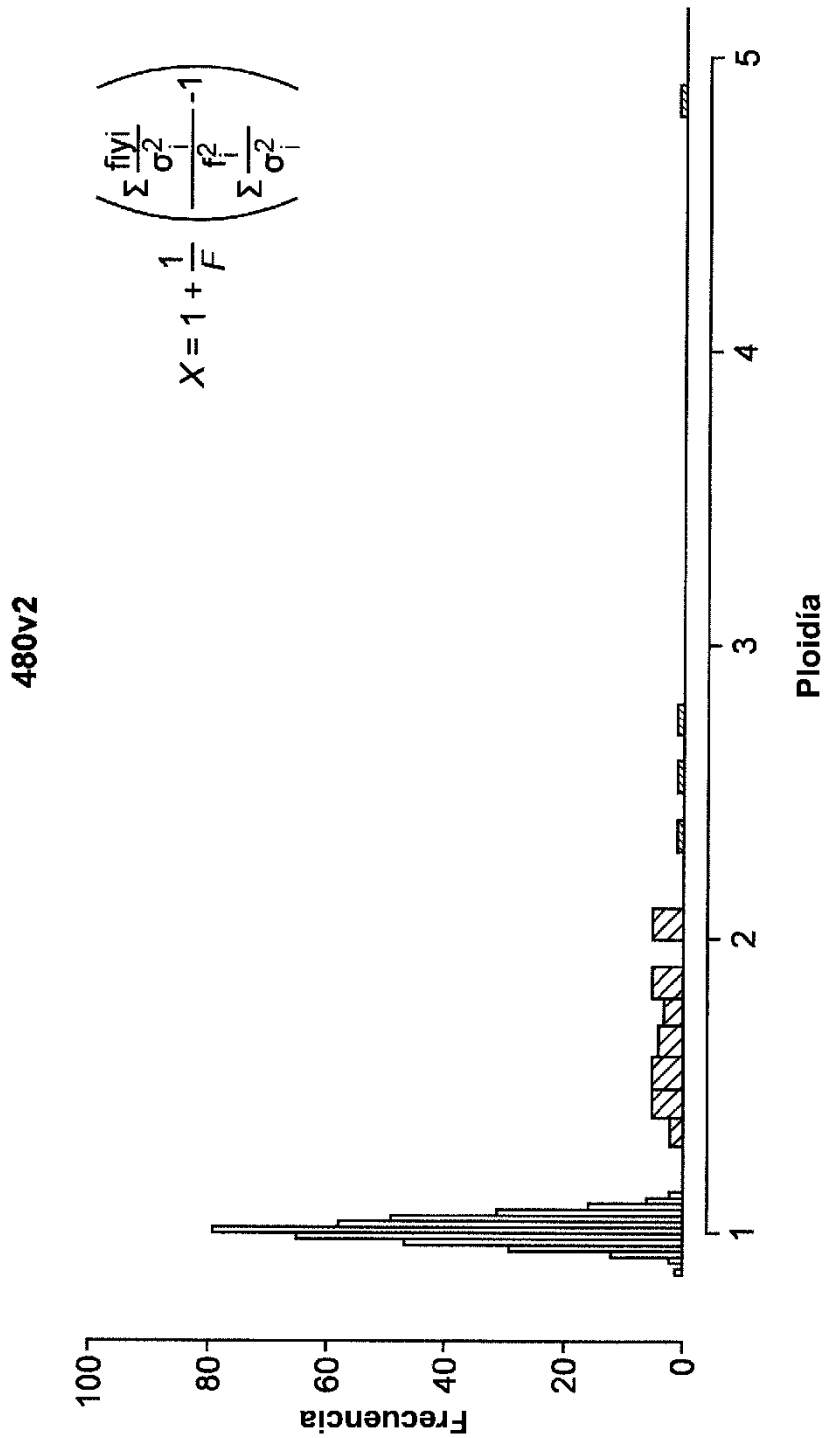

480v2
$X = 1 + \frac{1}{F}\left(\sum \frac{fiyi}{\sigma_i^2} \frac{f_i^2}{\sum \frac{1}{\sigma_i^2}}\right)^{-1}$
Frecuencia
0
20
40
60
80
100
1
2
3
4
5
Ploidía

FIG. 43

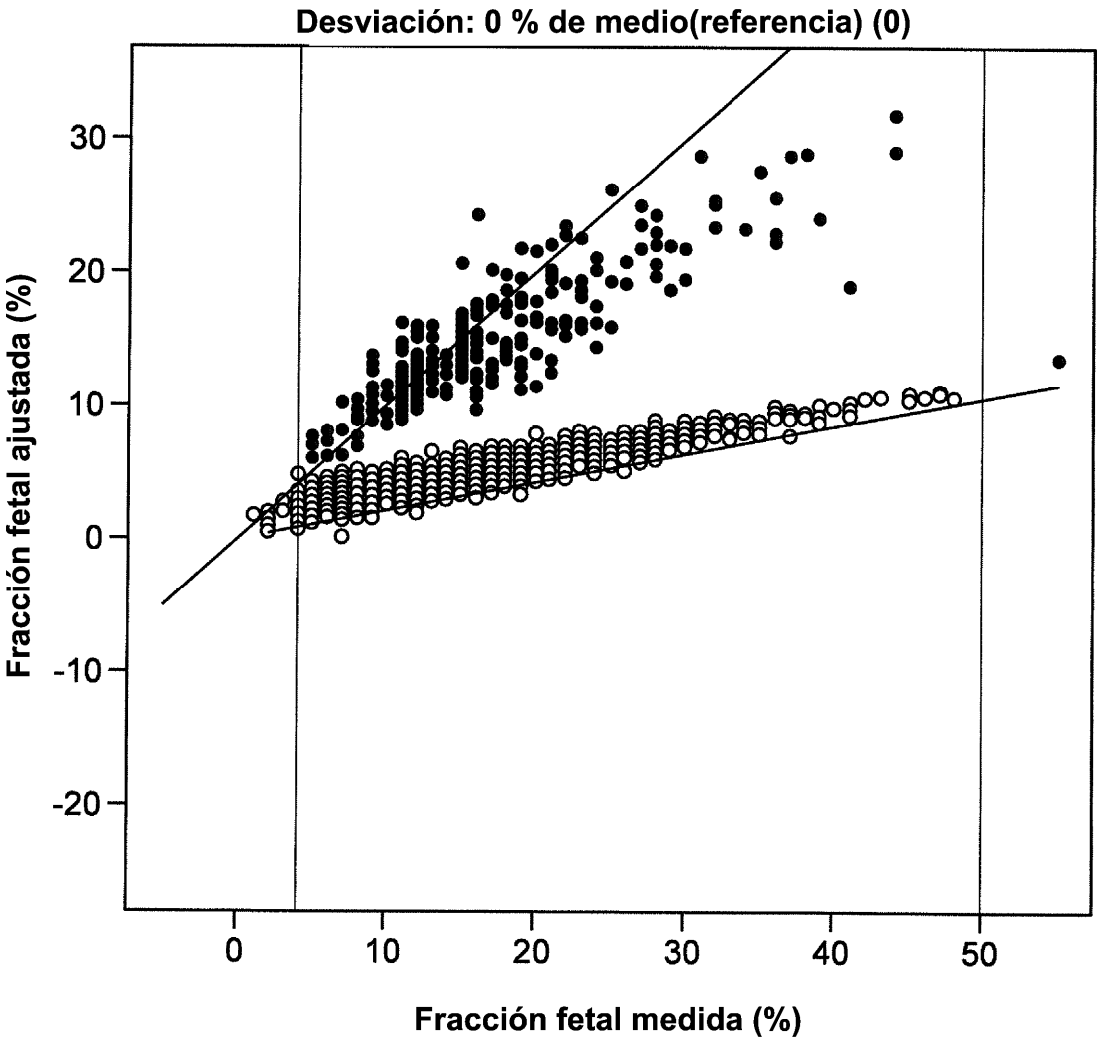

Desviación: 0 % de medio(referencia) (0)
Fracción fetal ajustada (%)
30
20
10
0
-10
-20
0
10
20
30
40
50
Fracción fetal medida (%)

FIG. 44

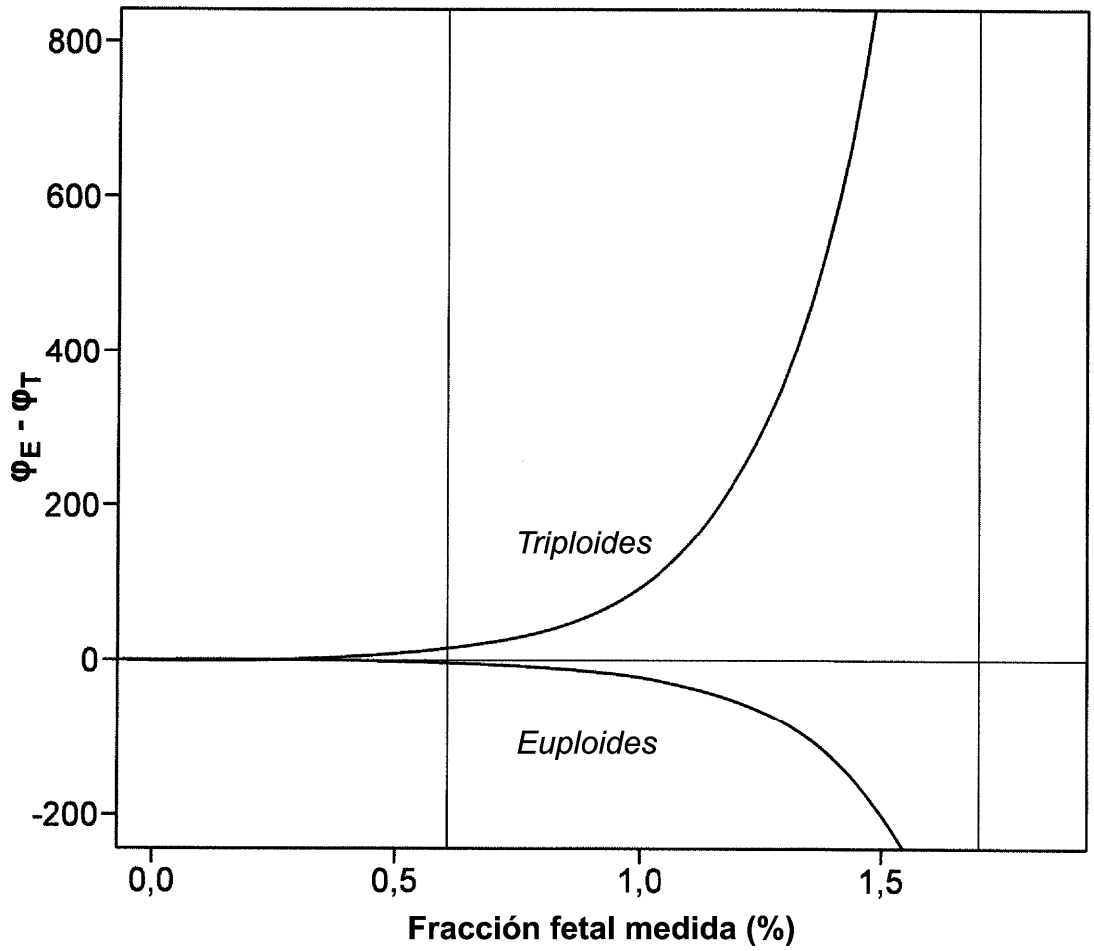

800
600
400
200
0
-200
$\varphi_E - \varphi_T$
Triploides
Euploides
0,0
0,5
1,0
1,5
Fracción fetal medida (%)

FIG. 45

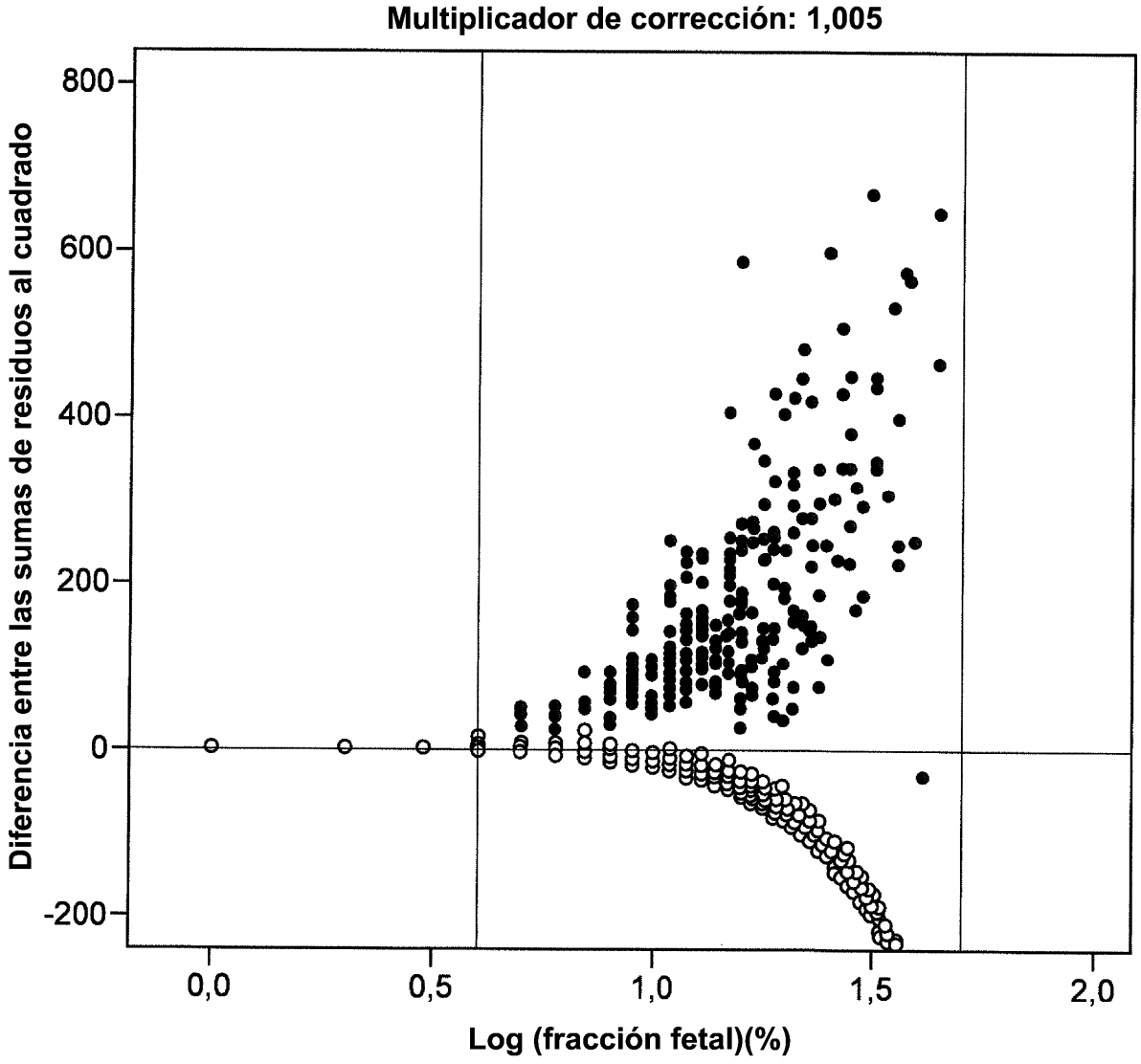

Multiplicador de corrección: 1,005
Diferencia entre las sumas de residuos al cuadrado
800
600
400
200
0
-200
0,0
0,5
1,0
1,5
2,0
Log (fracción fetal)(%)

FIG. 46

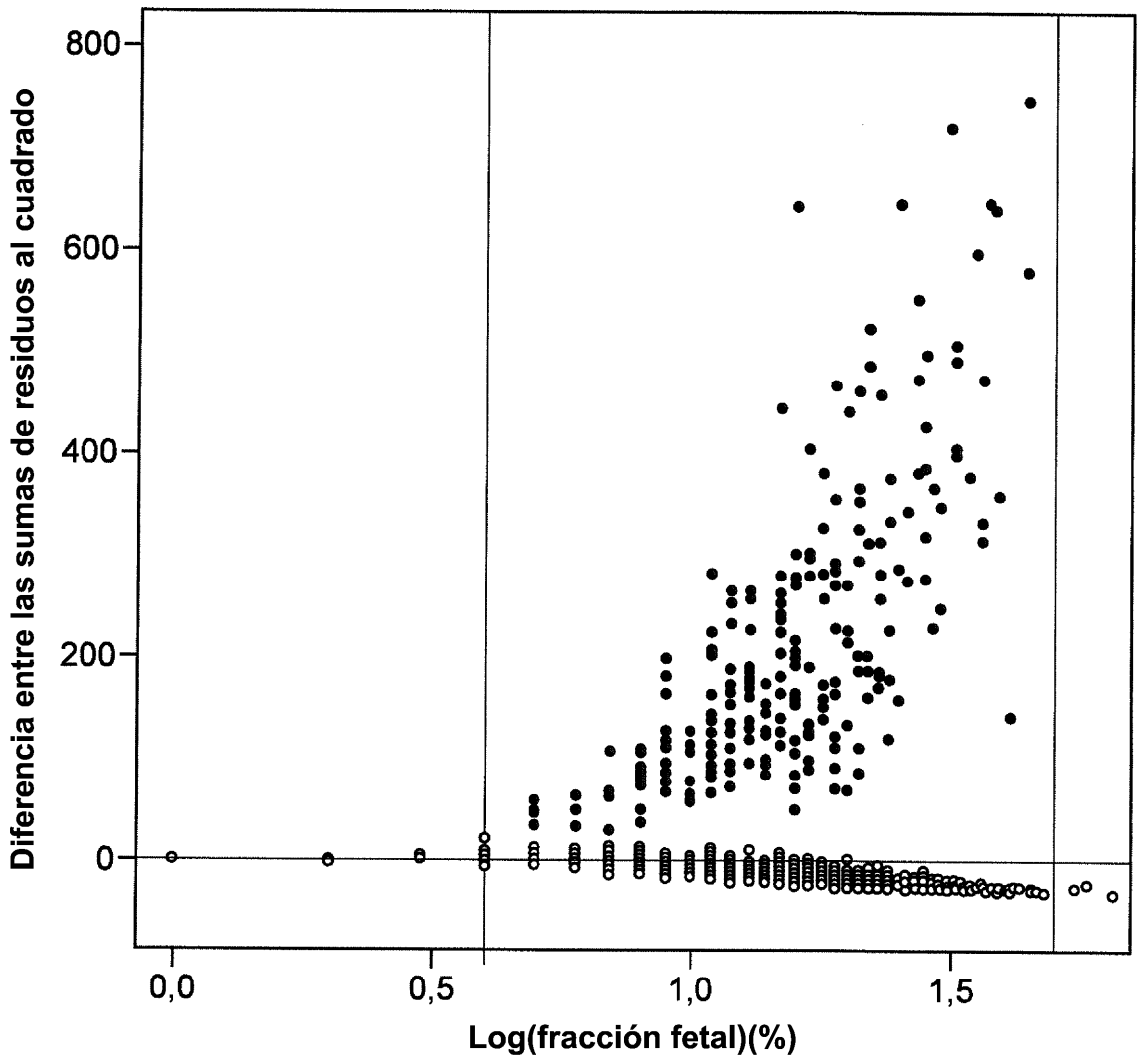

800
600
400
200
0
Diferencia entre las sumas de residuos al cuadrado
0,0
0,5
1,0
1,5
Log(fracción fetal)(%)

FIG. 47

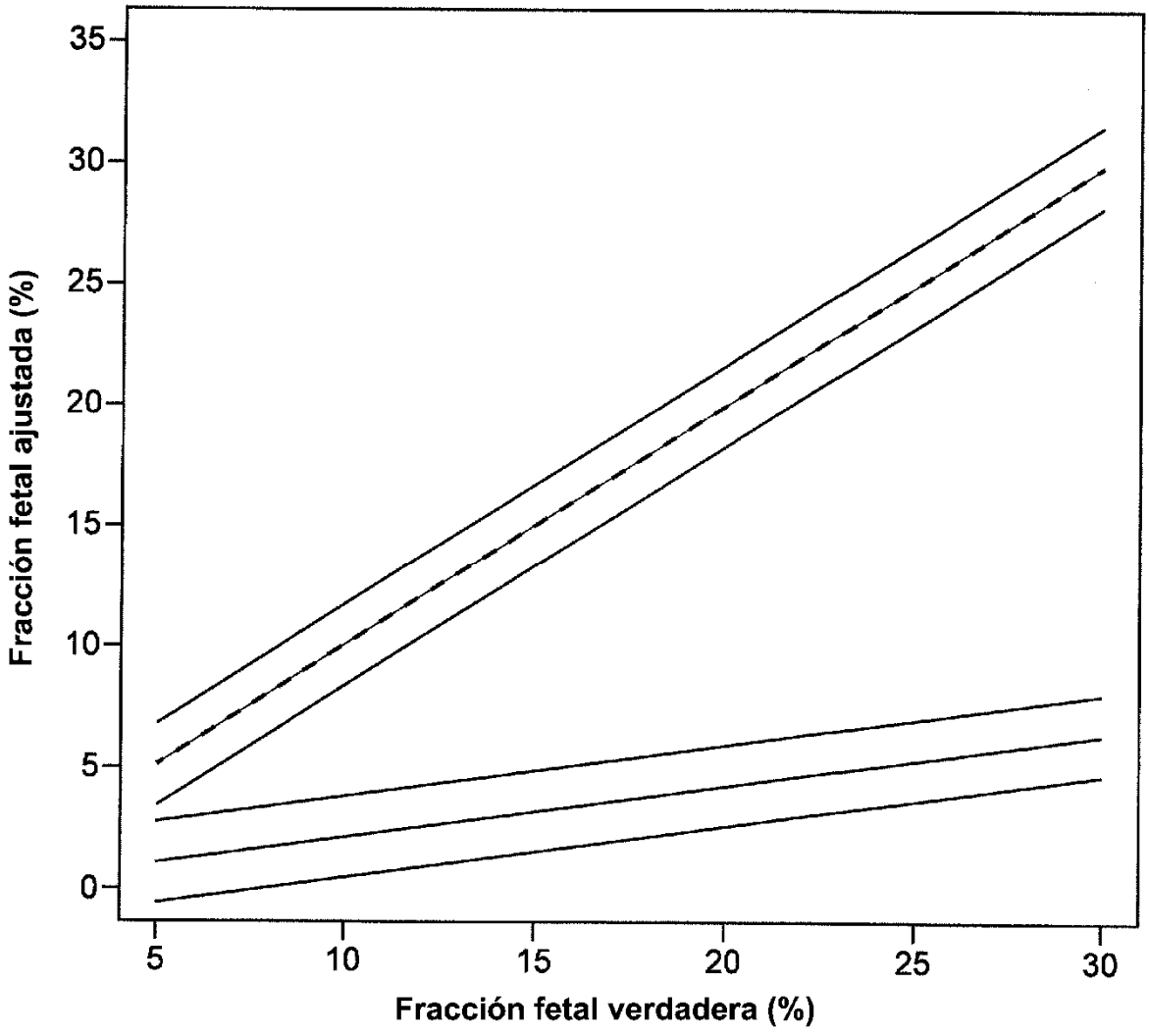

35
30
25
20
15
10
5
0
Fracción fetal ajustada (%)
5
10
15
20
25
30
Fracción fetal verdadera (%)

FIG. 48

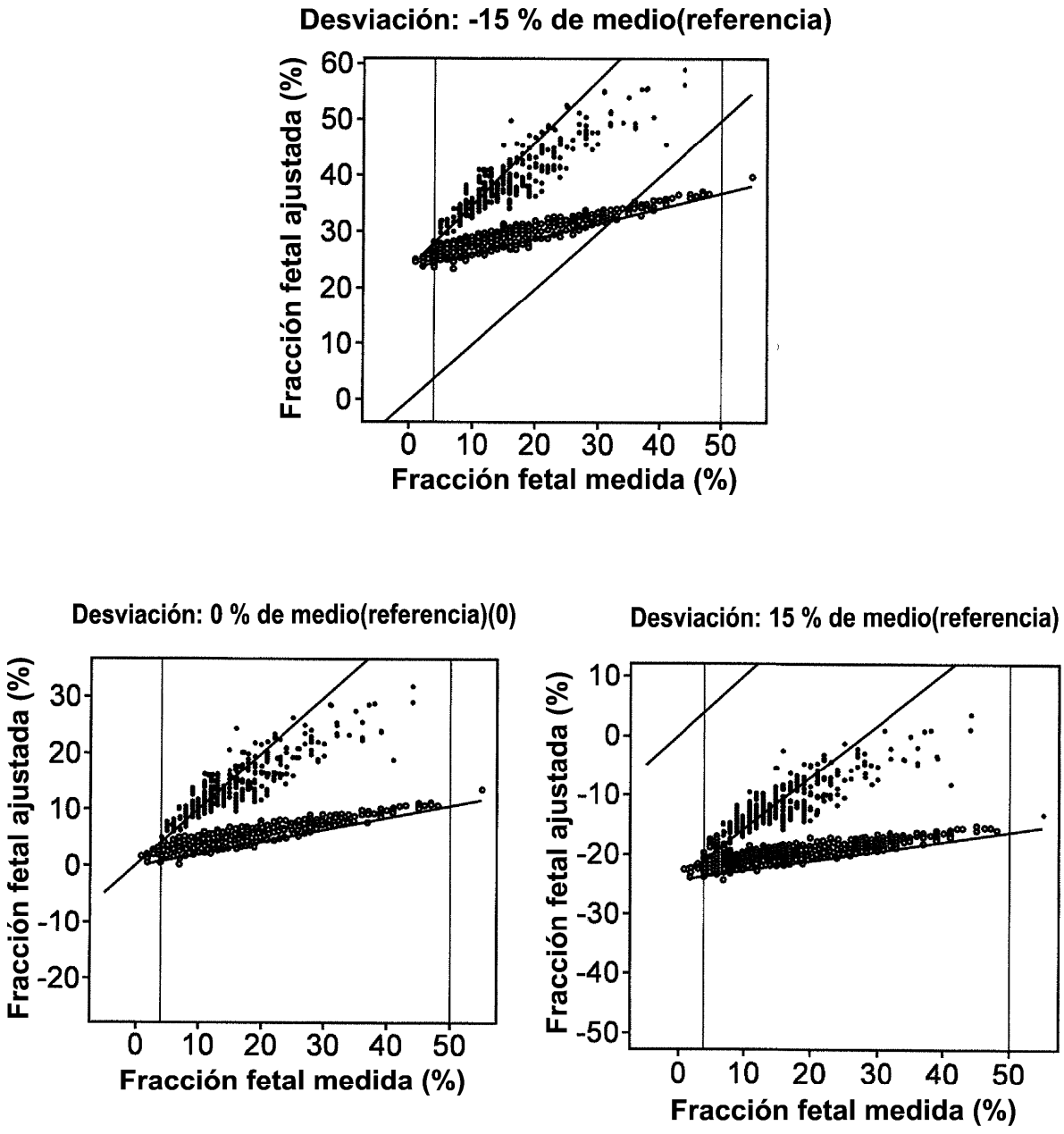

Desviación: -15 % de medio(referencia)
Fracción fetal ajustada (%)
60
50
40
30
20
10
0
0 10 20 30 40 50
Fracción fetal medida (%)
Desviación: 0 % de medio(referencia)(0)
Fracción fetal ajustada (%)
30
20
10
0
-10
-20
0 10 20 30 40 50
Fracción fetal medida (%)
Desviación: 15 % de medio(referencia)
Fracción fetal ajustada (%)
10
0
-10
-20
-30
-40
-50
0 10 20 30 40 50
Fracción fetal medida (%)

FIG. 49

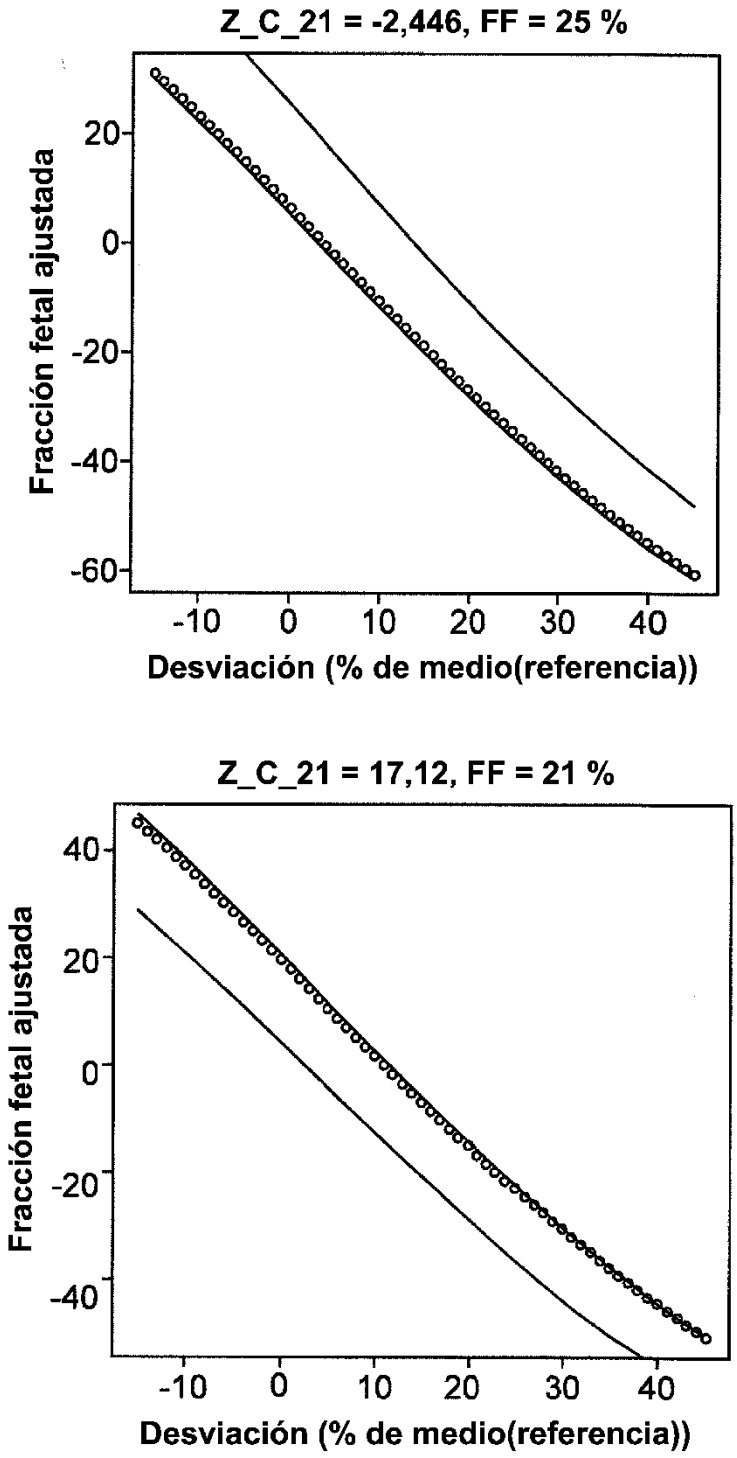

Z_C_21 = -2,446, FF = 25 %
Fracción fetal ajustada
20
0
-20
-40
-60
-10
0
10
20
30
40
Desviación (% de medio(referencia))
Z_C_21 = 17,12, FF = 21 %
Fracción fetal ajustada
40
20
0
-20
-40
-10
0
10
20
30
40
Desviación (% de medio(referencia))

FIG. 50

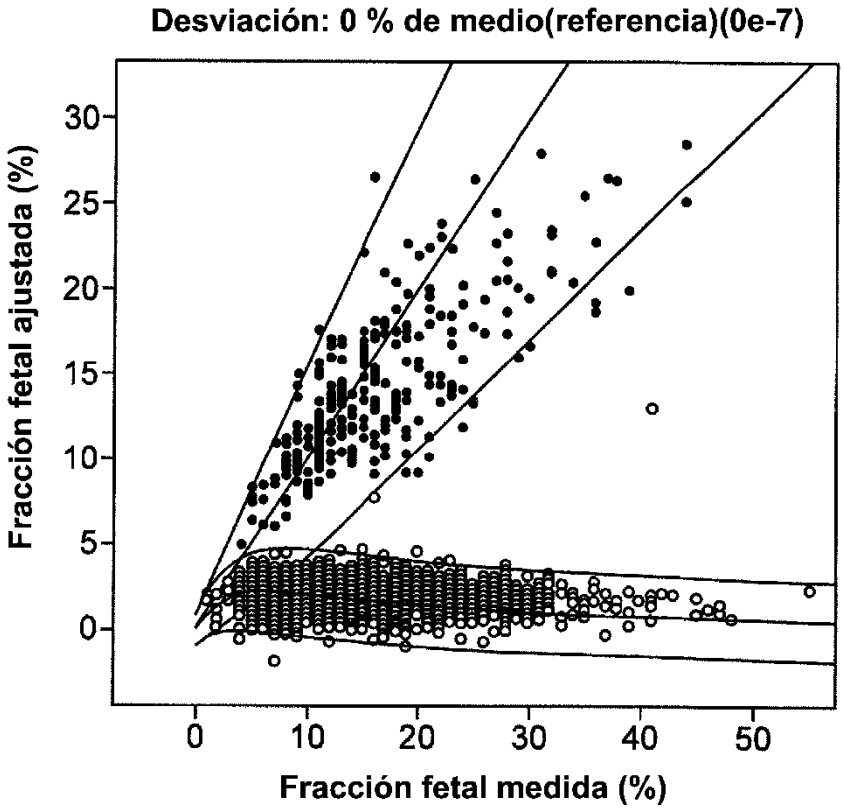

Desviación: 0 % de medio(referencia)(0e-7)
Fracción fetal ajustada (%)
30
25
20
15
10
5
0
0
10
20
30
40
50
Fracción fetal medida (%)

FIG. 51

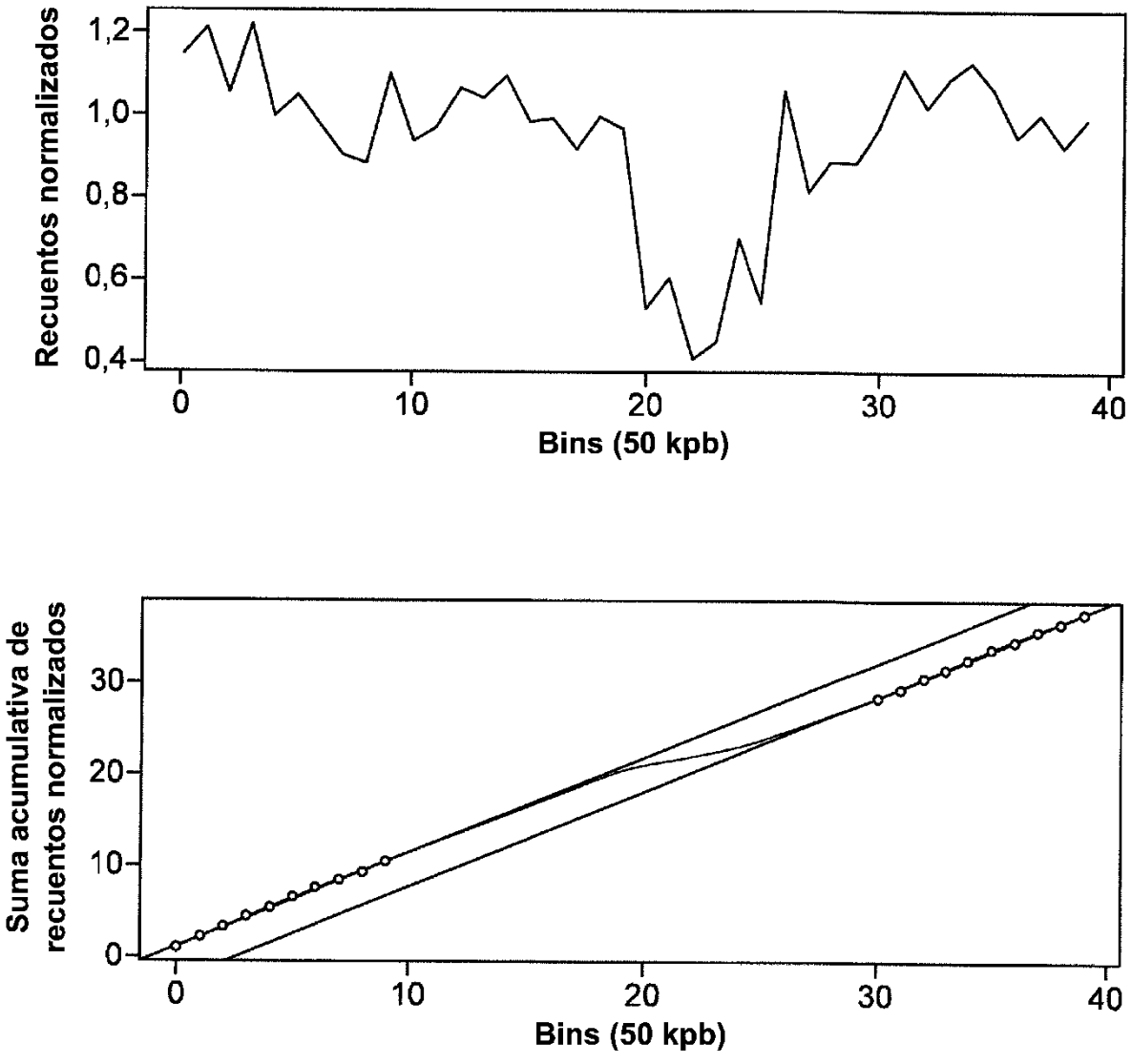

Recuentos normalizados
1,2
1,0
0,8
0,6
0,4
0
10
20
30
40
Bins (50 kpb)
Suma acumulativa de recuentos normalizados
30
20
10
0
0
10
20
30
40
Bins (50 kpb)

FIG. 52

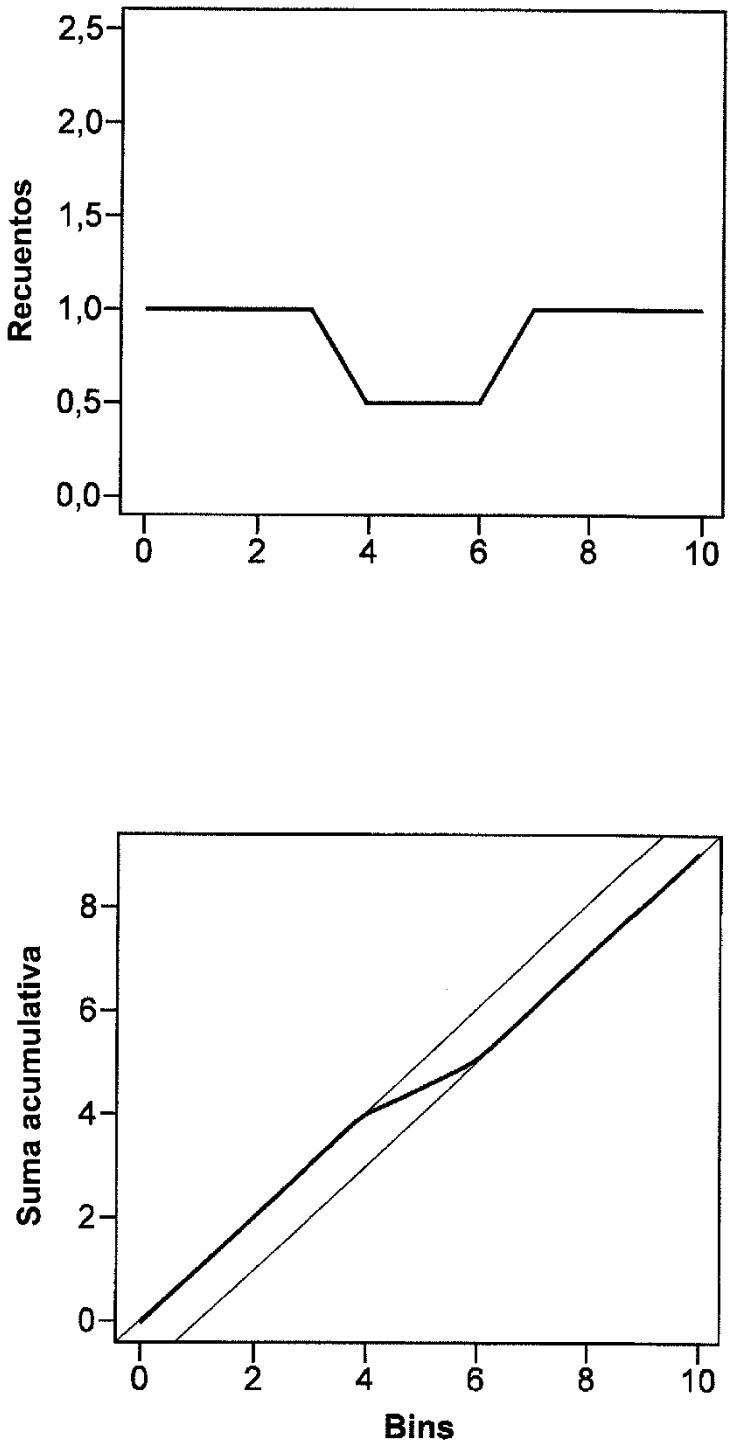

Recuentos
2,5
2,0
1,5
1,0
0,5
0,0
0
2
4
6
8
10
Suma acumulativa
8
6
4
2
0
0
2
4
6
8
10
Bins

FIG. 53

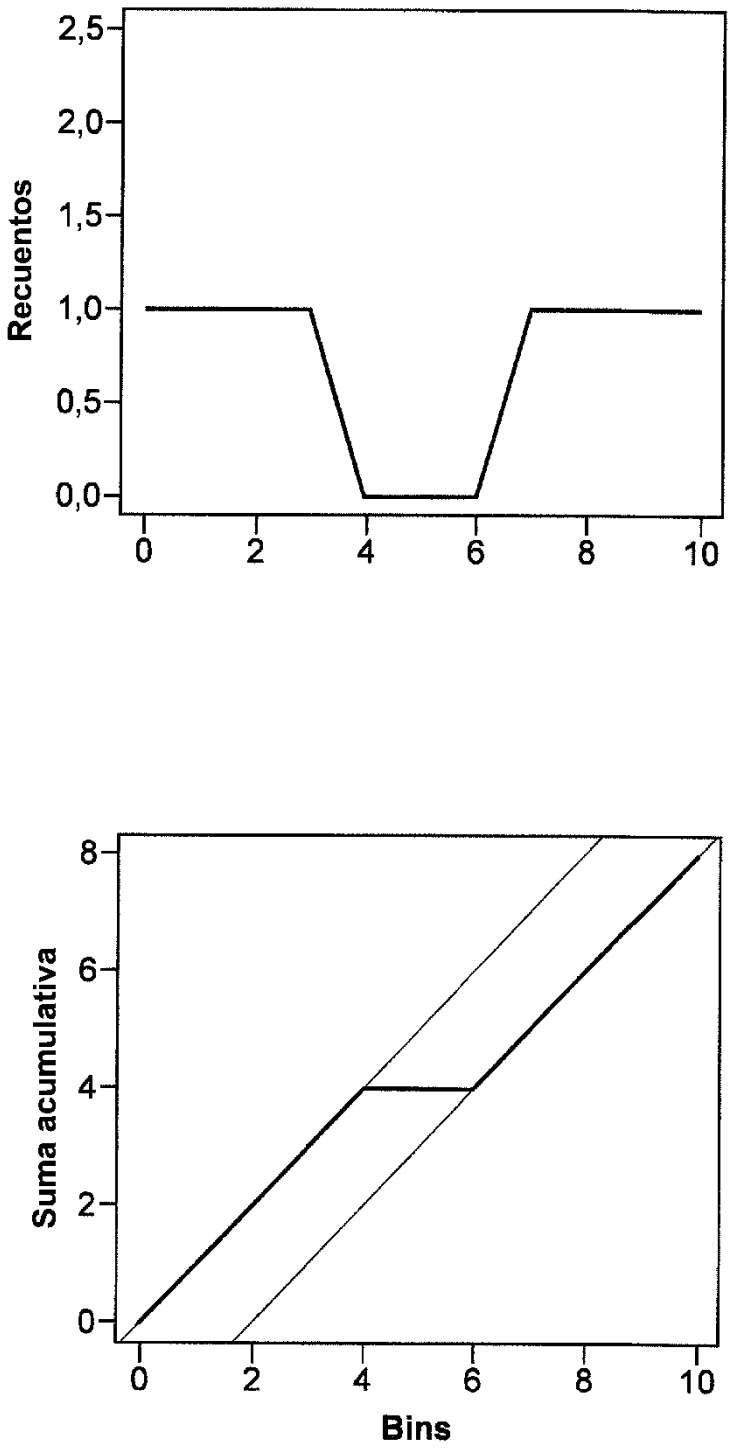

Recuentos
2,5
2,0
1,5
1,0
0,5
0,0
0
2
4
6
8
10
Suma acumulativa
8
6
4
2
0
0
2
4
6
8
10
Bins

FIG. 54

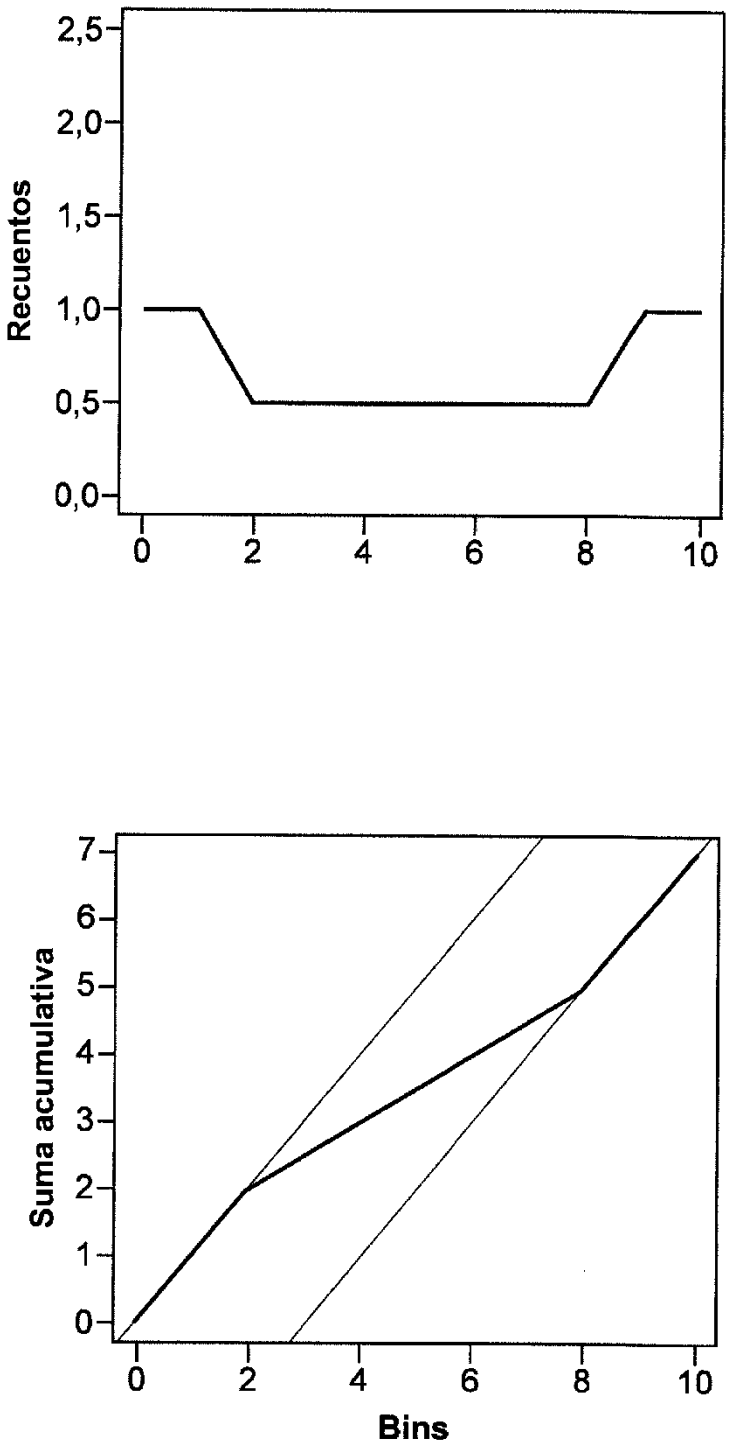

Recuentos
2,5
2,0
1,5
1,0
0,5
0,0
0
2
4
6
8
10
Suma acumulativa
7
6
5
4
3
2
1
0
0
2
4
6
8
10
Bins

FIG. 55

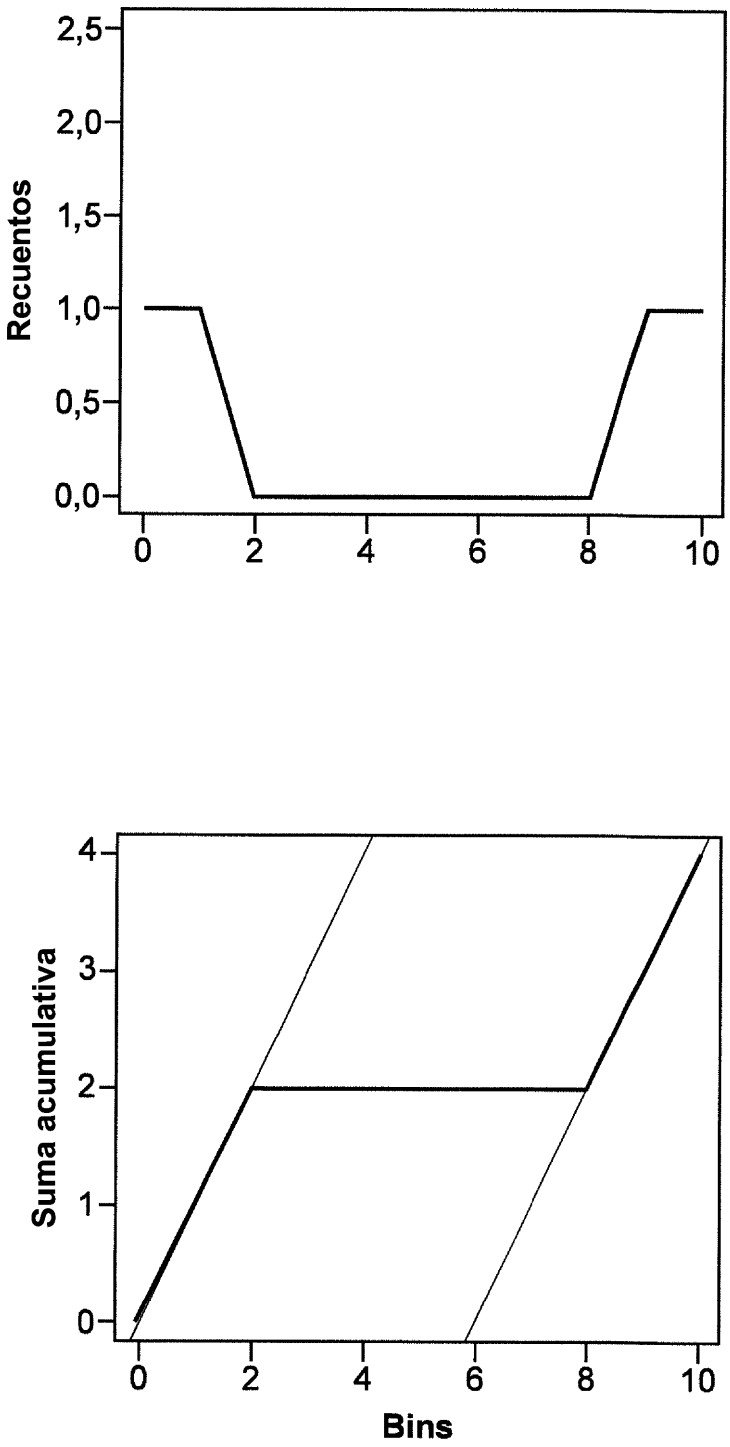

Recuentos
2,5
2,0
1,5
1,0
0,5
0,0
0
2
4
6
8
10
Suma acumulativa
4
3
2
1
0
0
2
4
6
8
10
Bins

FIG. 56

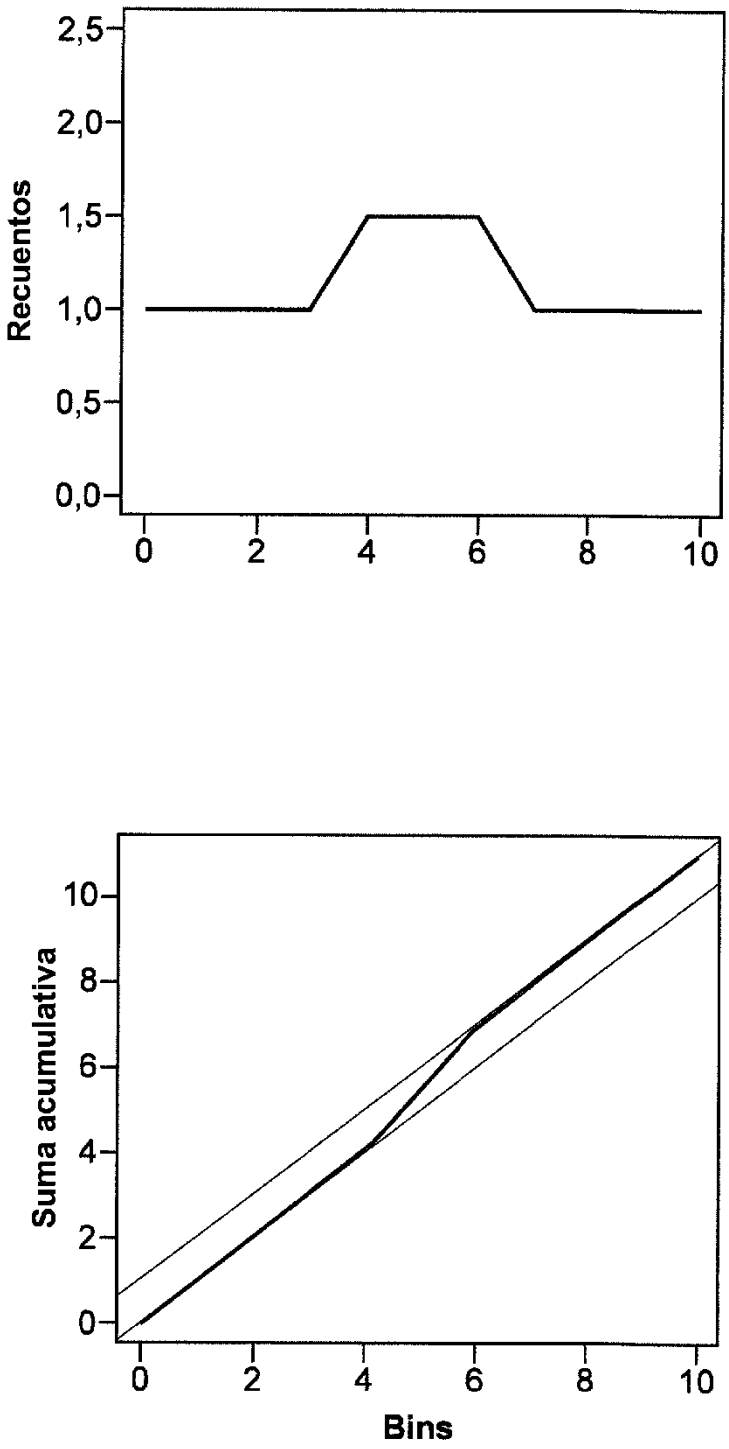

Recuentos
2,5
2,0
1,5
1,0
0,5
0,0
0
2
4
6
8
10
Suma acumulativa
10
8
6
4
2
0
0
2
4
6
8
10
Bins

FIG. 57

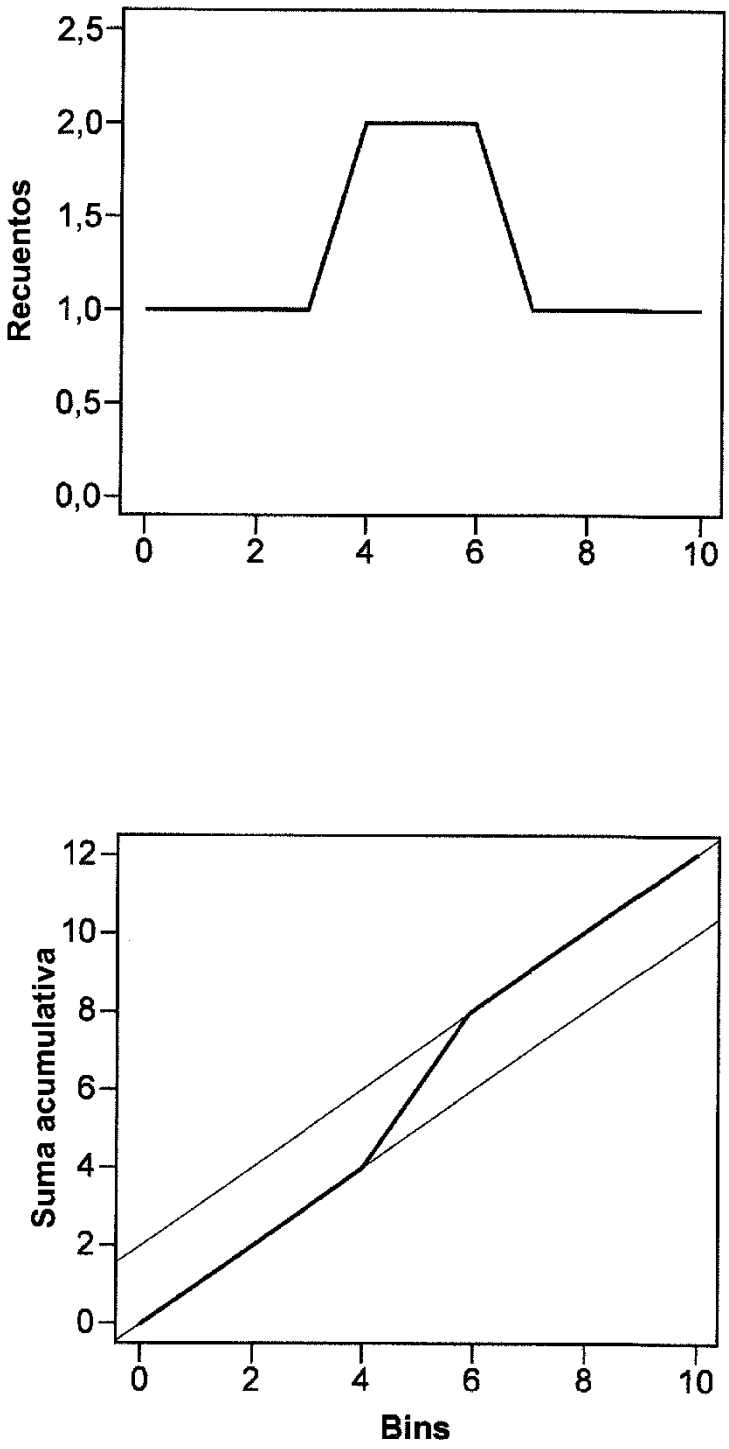

Recuentos
2,5
2,0
1,5
1,0
0,5
0,0
0
2
4
6
8
10
Suma acumulativa
12
10
8
6
4
2
0
0
2
4
6
8
10
Bins

FIG. 58

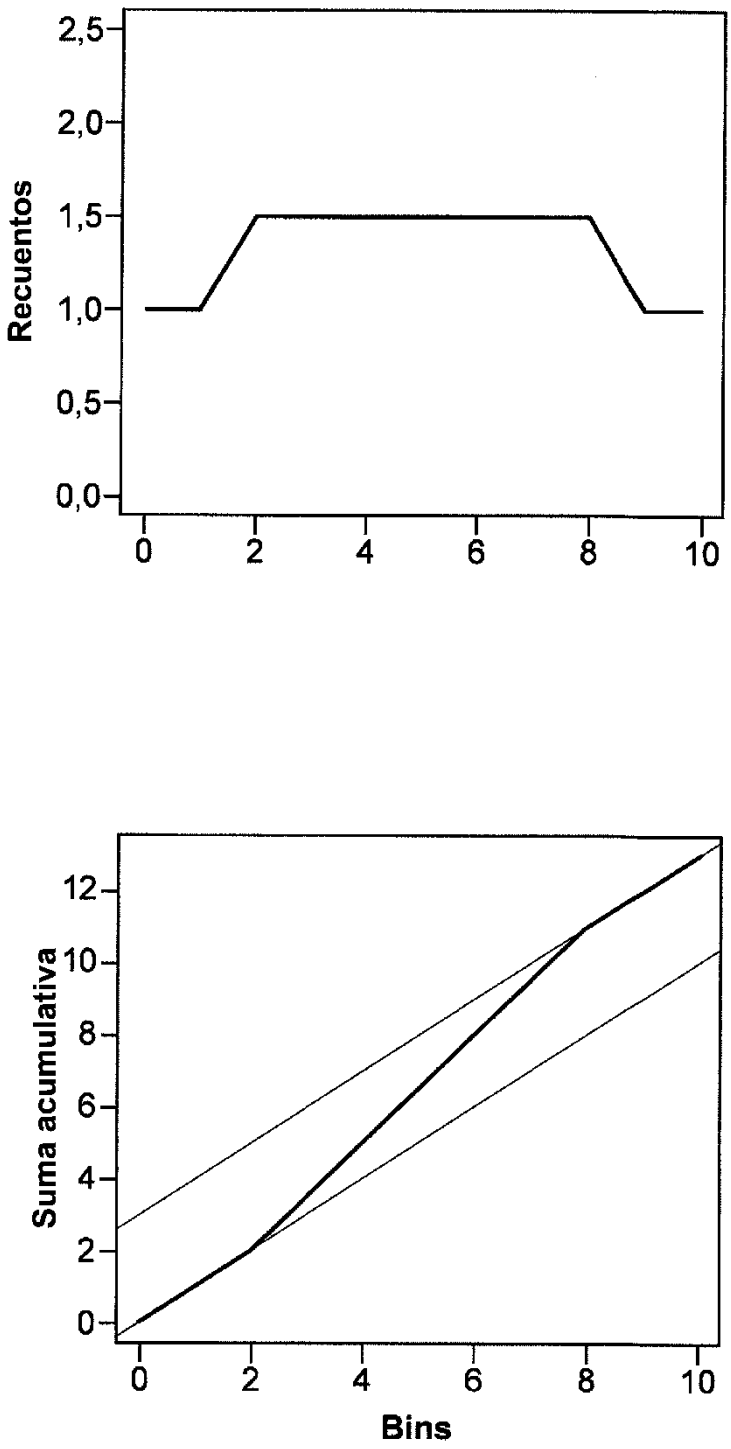

Recuentos
2,5
2,0
1,5
1,0
0,5
0,0
0
2
4
6
8
10
Suma acumulativa
12
10
8
6
4
2
0
0
2
4
6
8
10
Bins

FIG. 59

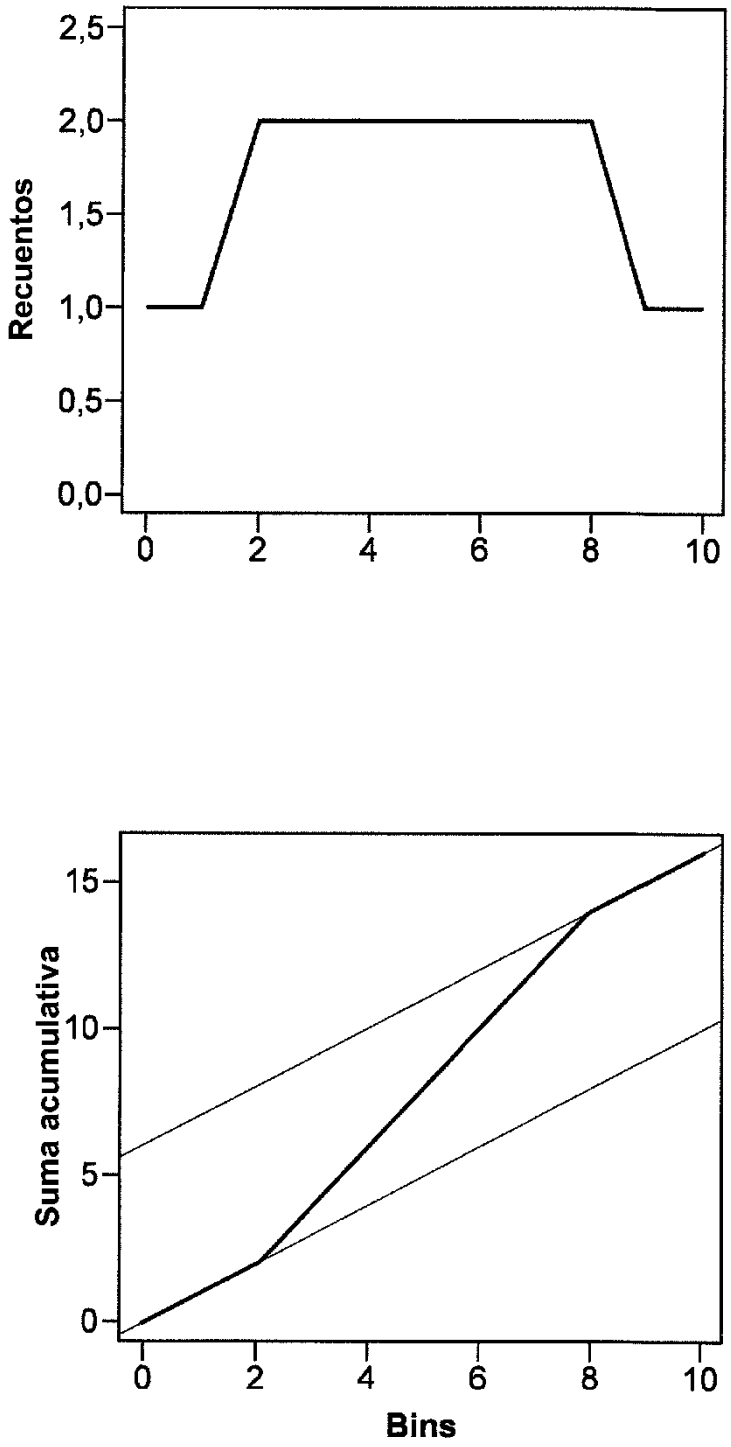

Recuentos
2,5
2,0
1,5
1,0
0,5
0,0
0
2
4
6
8
10
Suma acumulativa
15
10
5
0
0
2
4
6
8
10
Bins

FIG. 60

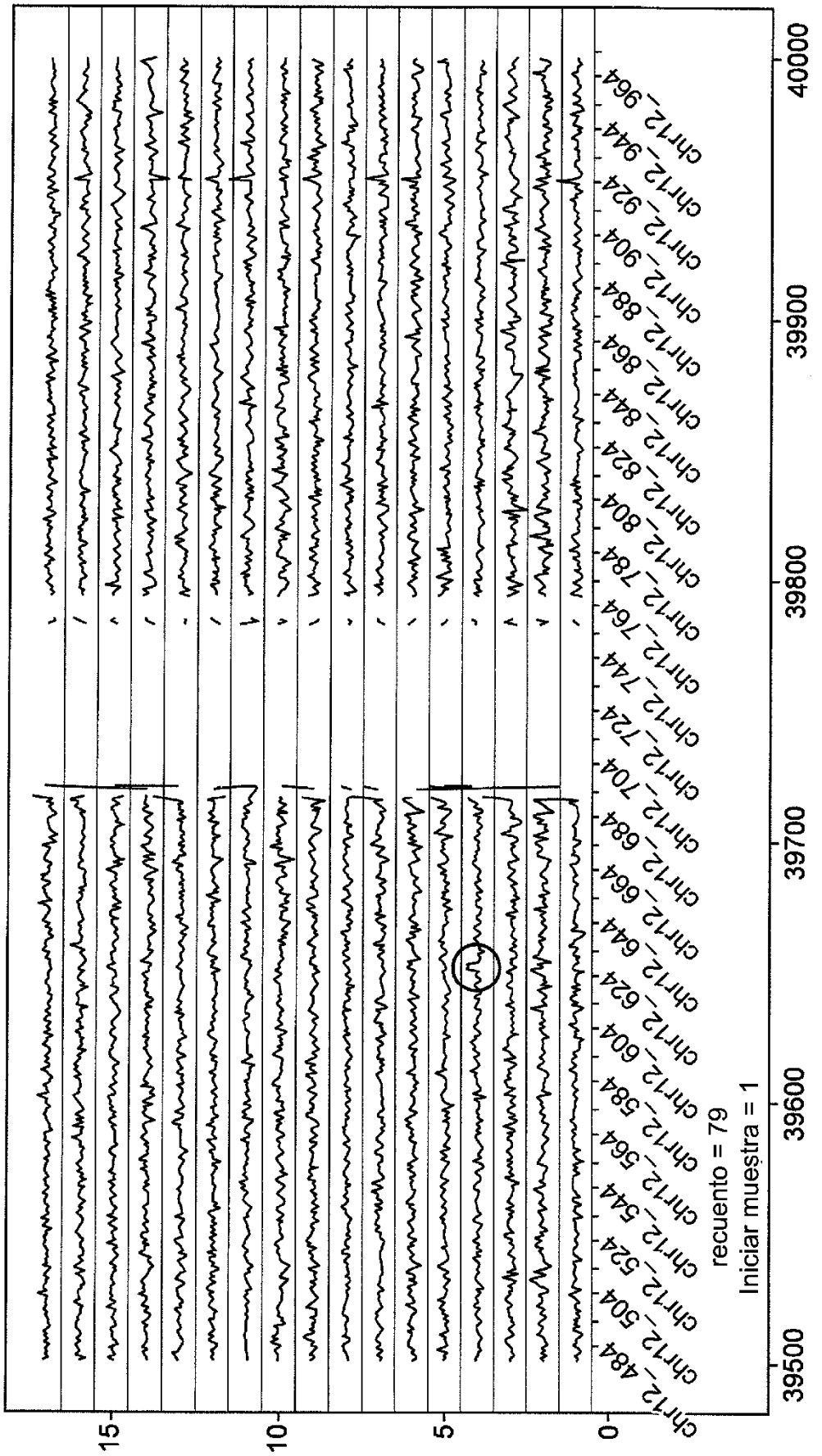

15
10
5
0
39500
39600
39700
39800
39900
40000
chr12_48a
chr12_50a
chr12_52a
chr12_54a
chr12_56a
chr12_58a
chr12_60a
chr12_62a
chr12_64a
chr12_66a
chr12_68a
chr12_70a
chr12_72a
chr12_74a
chr12_76a
chr12_78a
chr12_80a
chr12_82a
chr12_84a
chr12_86a
chr12_88a
chr12_90a
chr12_92a
chr12_94a
chr12_96a
recuento = 79
Iniciar muestra = 1

FIG. 61A

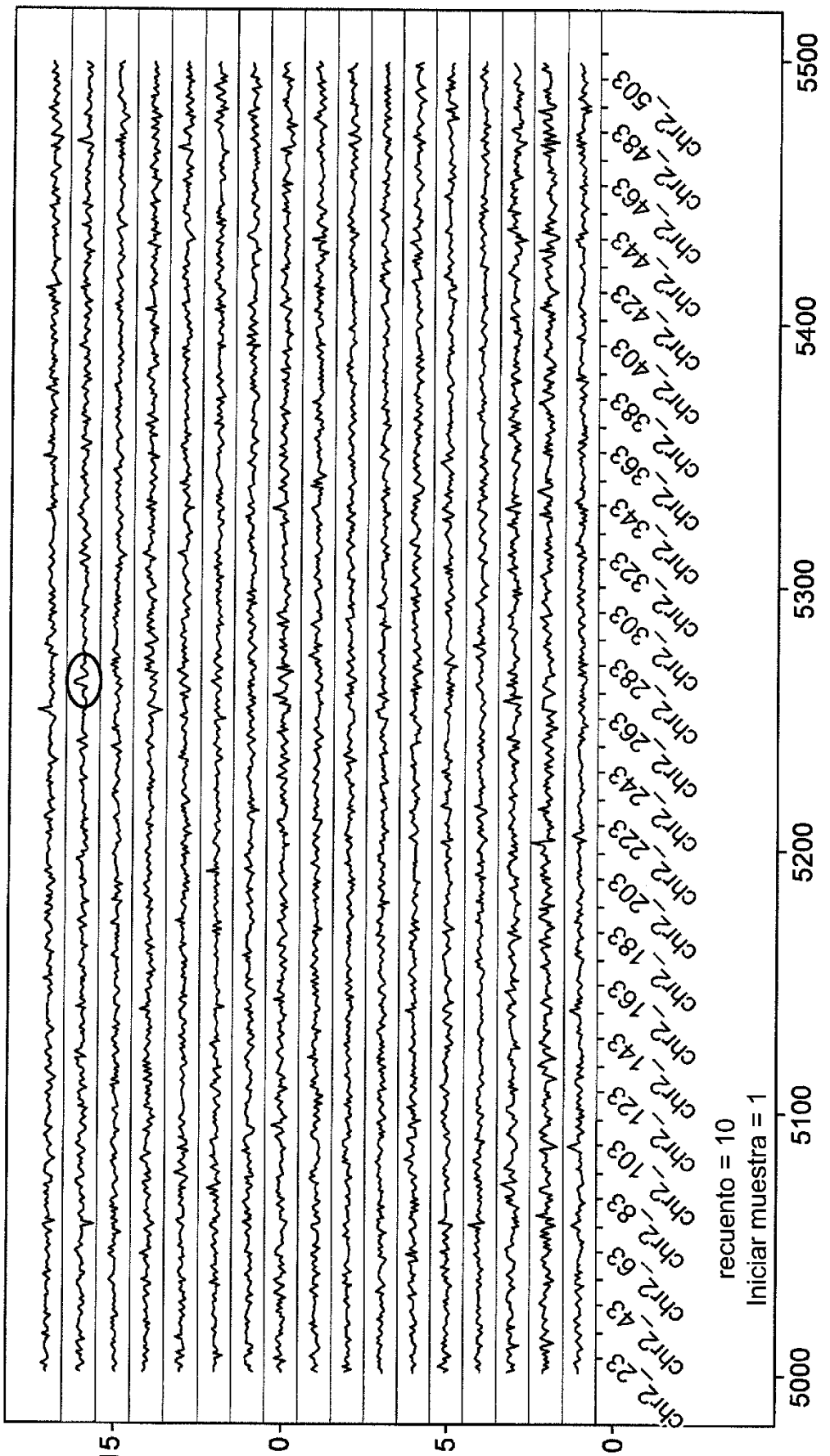

15
10
5
0
5000
5100
5200
5300
5400
5500
recuento = 10
Iniciar muestra = 1
chr2_23
chr2_43
chr2_63
chr2_83
chr2_103
chr2_123
chr2_143
chr2_163
chr2_183
chr2_203
chr2_223
chr2_243
chr2_263
chr2_283
chr2_303
chr2_323
chr2_343
chr2_363
chr2_383
chr2_403
chr2_423
chr2_443
chr2_463
chr2_483
chr2_503

FIG. 61B

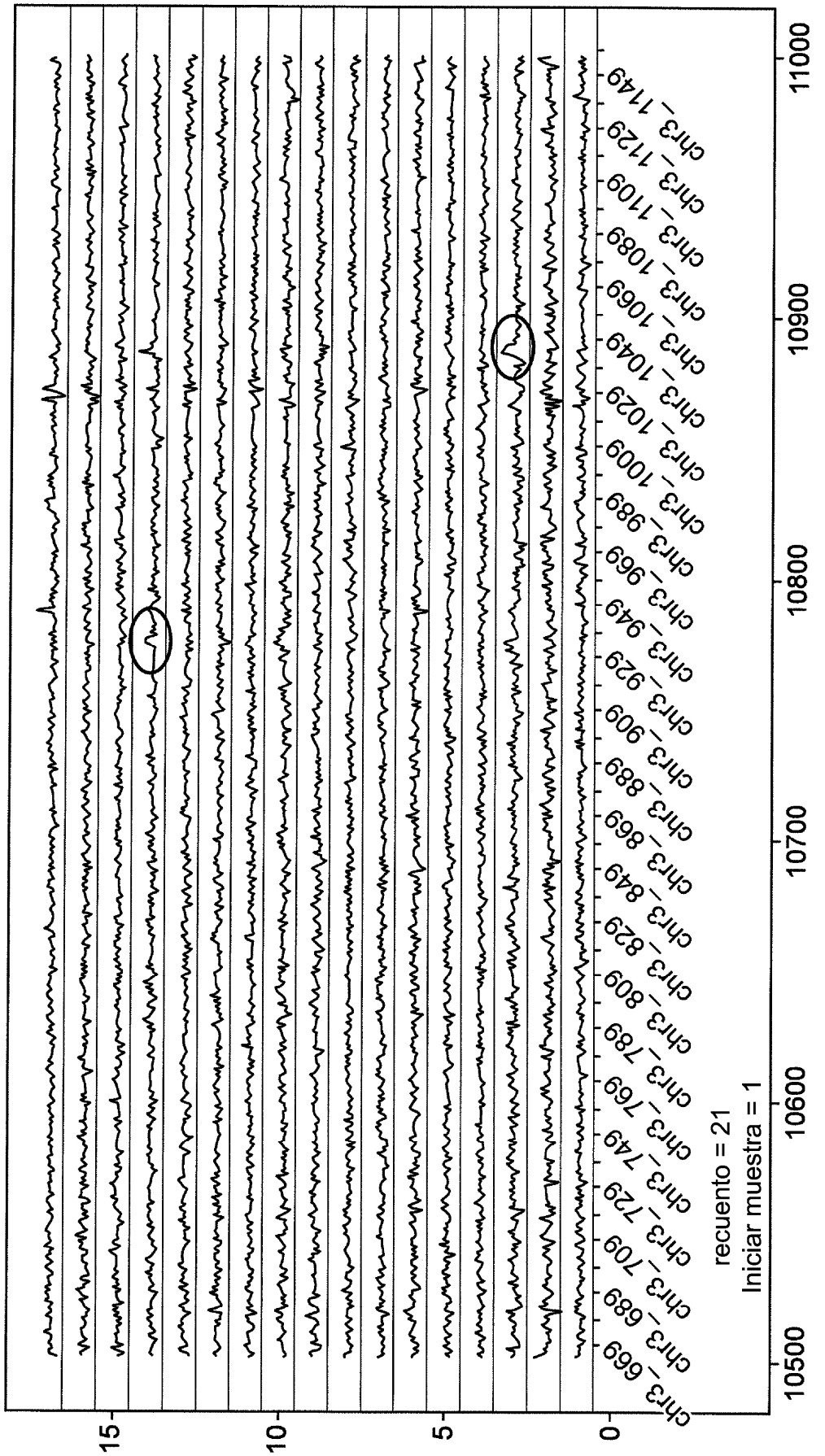

15
10
5
0
10500
10600
10700
10800
10900
11000
recuento = 21
Iniciar muestra = 1
chr3_669
chr3_689
chr3_709
chr3_729
chr3_749
chr3_769
chr3_789
chr3_809
chr3_829
chr3_849
chr3_869
chr3_889
chr3_909
chr3_929
chr3_949
chr3_969
chr3_989
chr3_1009
chr3_1029
chr3_1049
chr3_1069
chr3_1089
chr3_1109
chr3_1129
chr3_1149

FIG. 61C

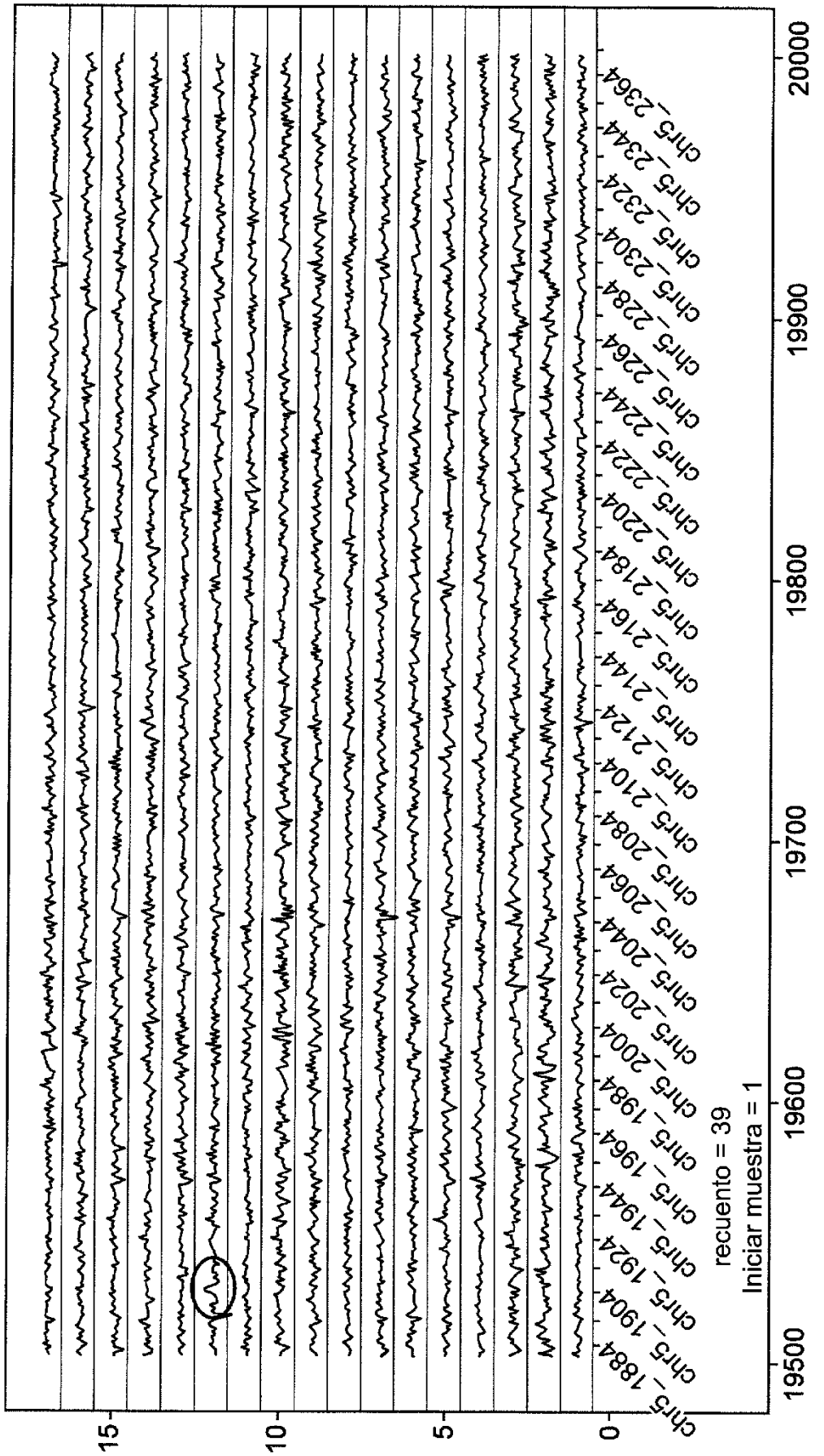

15
10
5
0
19500
19600
19700
19800
19900
20000
Iniciar muestra = 1
recuento = 39
chr5_1884
chr5_1904
chr5_1924
chr5_1944
chr5_1964
chr5_1984
chr5_2004
chr5_2024
chr5_2044
chr5_2064
chr5_2084
chr5_2104
chr5_2124
chr5_2144
chr5_2164
chr5_2184
chr5_2204
chr5_2224
chr5_2244
chr5_2264
chr5_2284
chr5_2304
chr5_2324
chr5_2344
chr5_2364

FIG. 61D

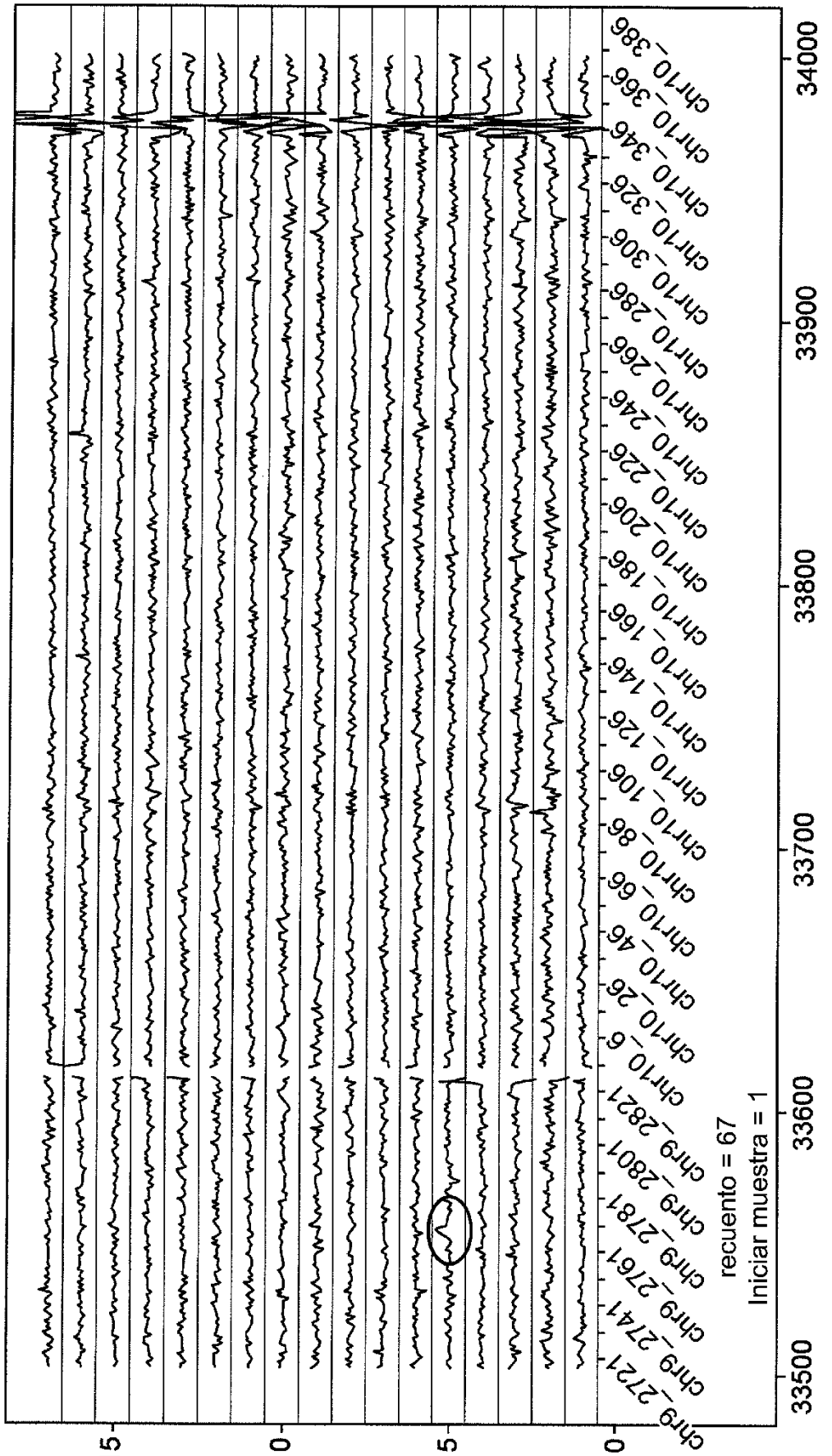

chr9_2721
chr9_2741
chr9_2761
chr9_2781
chr9_2801
chr9_2821
chr10_6
chr10_26
chr10_46
chr10_66
chr10_86
chr10_106
chr10_126
chr10_146
chr10_166
chr10_186
chr10_206
chr10_226
chr10_246
chr10_266
chr10_286
chr10_306
chr10_326
chr10_346
chr10_366
chr10_386
recuento = 67
Iniciar muestra = 1
0
5
10
15
33500
33600
33700
33800
33900
34000

FIG. 61E

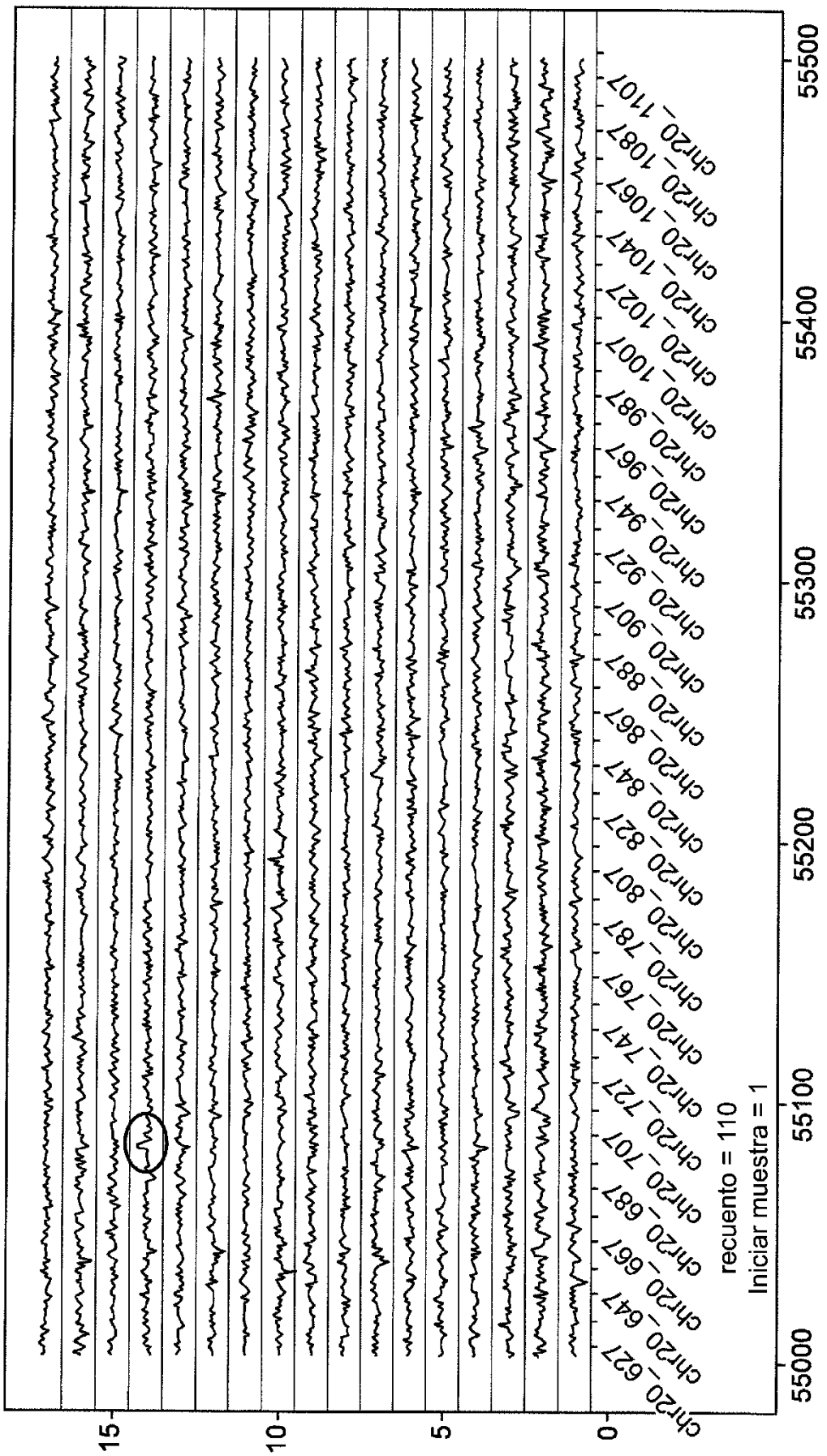

15
10
5
0
55000
55100
55200
55300
55400
55500
recuento = 110
Iniciar muestra = 1
chr20_627
chr20_647
chr20_667
chr20_687
chr20_707
chr20_727
chr20_747
chr20_767
chr20_787
chr20_807
chr20_827
chr20_847
chr20_867
chr20_887
chr20_907
chr20_927
chr20_947
chr20_967
chr20_987
chr20_1007
chr20_1027
chr20_1047
chr20_1067
chr20_1087
chr20_1107

FIG. 61F

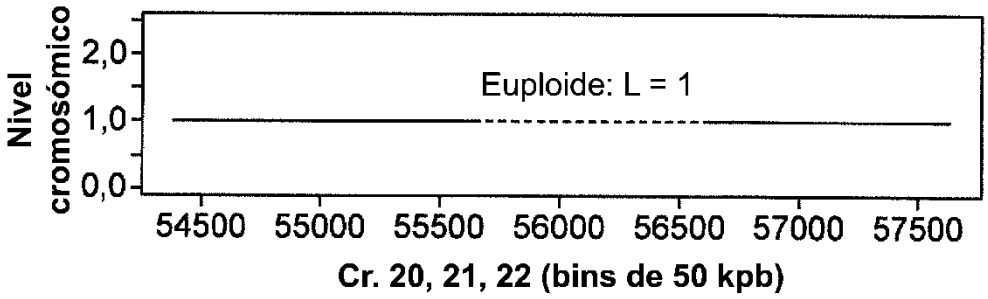

Nivel cromosómico
2,0
1,0
0,0
Euploide: L = 1
54500
55000
55500
56000
56500
57000
57500
Cr. 20, 21, 22 (bins de 50 kpb)

FIG. 62

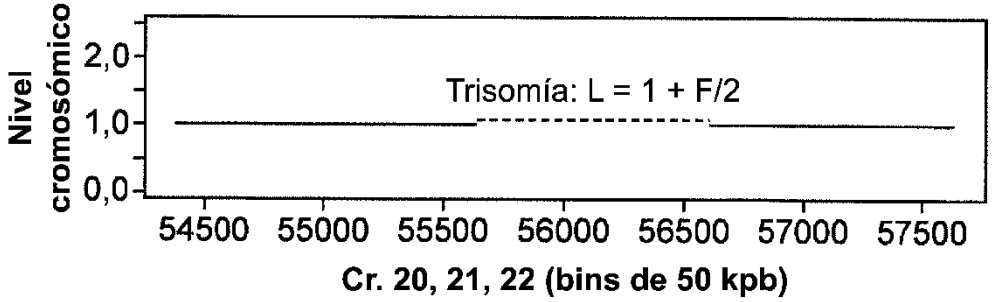

Nivel cromosómico
2,0
1,0
0,0
Trisomía: L = 1 + F/2
54500
55000
55500
56000
56500
57000
57500
Cr. 20, 21, 22 (bins de 50 kpb)

FIG. 63

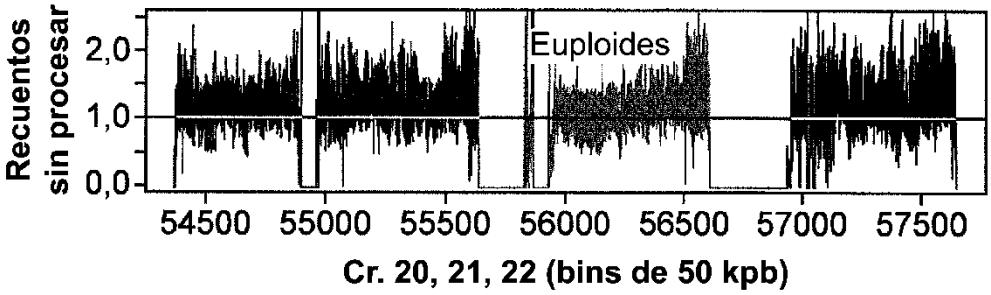

Recuentos sin procesar
2,0
1,0
0,0
Euploides
54500
55000
55500
56000
56500
57000
57500
Cr. 20, 21, 22 (bins de 50 kpb)

FIG. 64

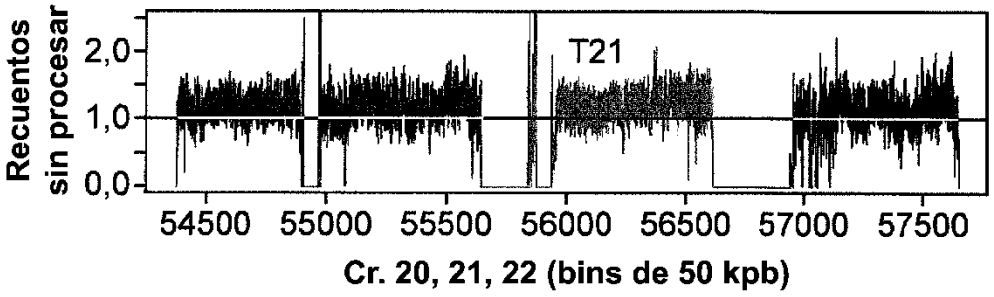

Recuentos sin procesar
2,0
1,0
0,0
T21
54500
55000
55500
56000
56500
57000
57500
Cr. 20, 21, 22 (bins de 50 kpb)

FIG. 65

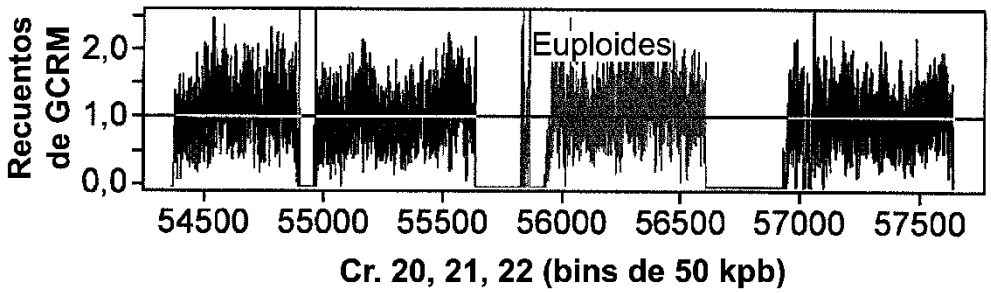

Recuentos de GCRM
2,0
1,0
0,0
Euploides
54500
55000
55500
56000
56500
57000
57500
Cr. 20, 21, 22 (bins de 50 kpb)

FIG. 66

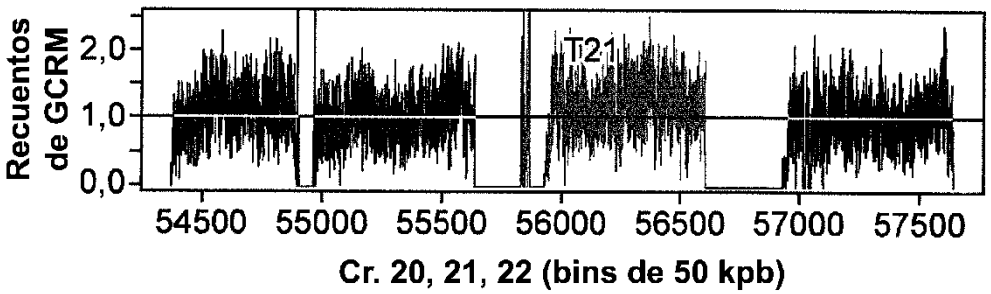

Recuentos de GCRM
2,0
1,0
0,0
T21
54500
55000
55500
56000
56500
57000
57500
Cr. 20, 21, 22 (bins de 50 kpb)

FIG. 67

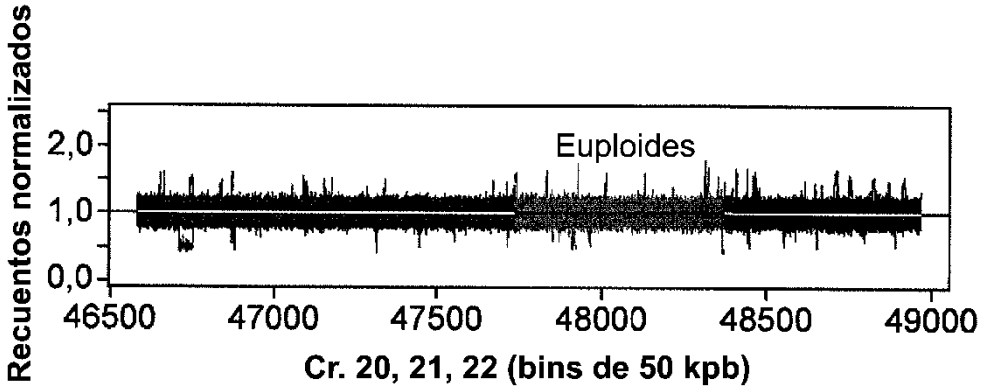

Recuentos normalizados
2,0
1,0
0,0
Euploides
46500
47000
47500
48000
48500
49000
Cr. 20, 21, 22 (bins de 50 kpb)

FIG. 68

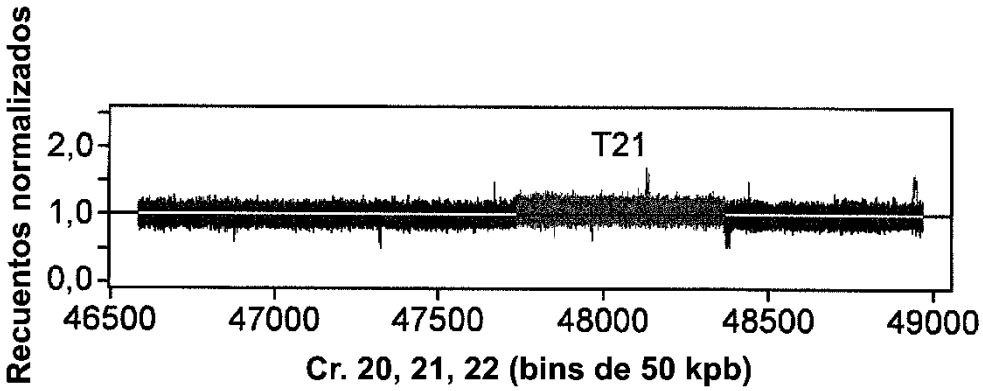

Recuentos normalizados
2,0
1,0
0,0
T21
46500
47000
47500
48000
48500
49000
Cr. 20, 21, 22 (bins de 50 kpb)

FIG. 69

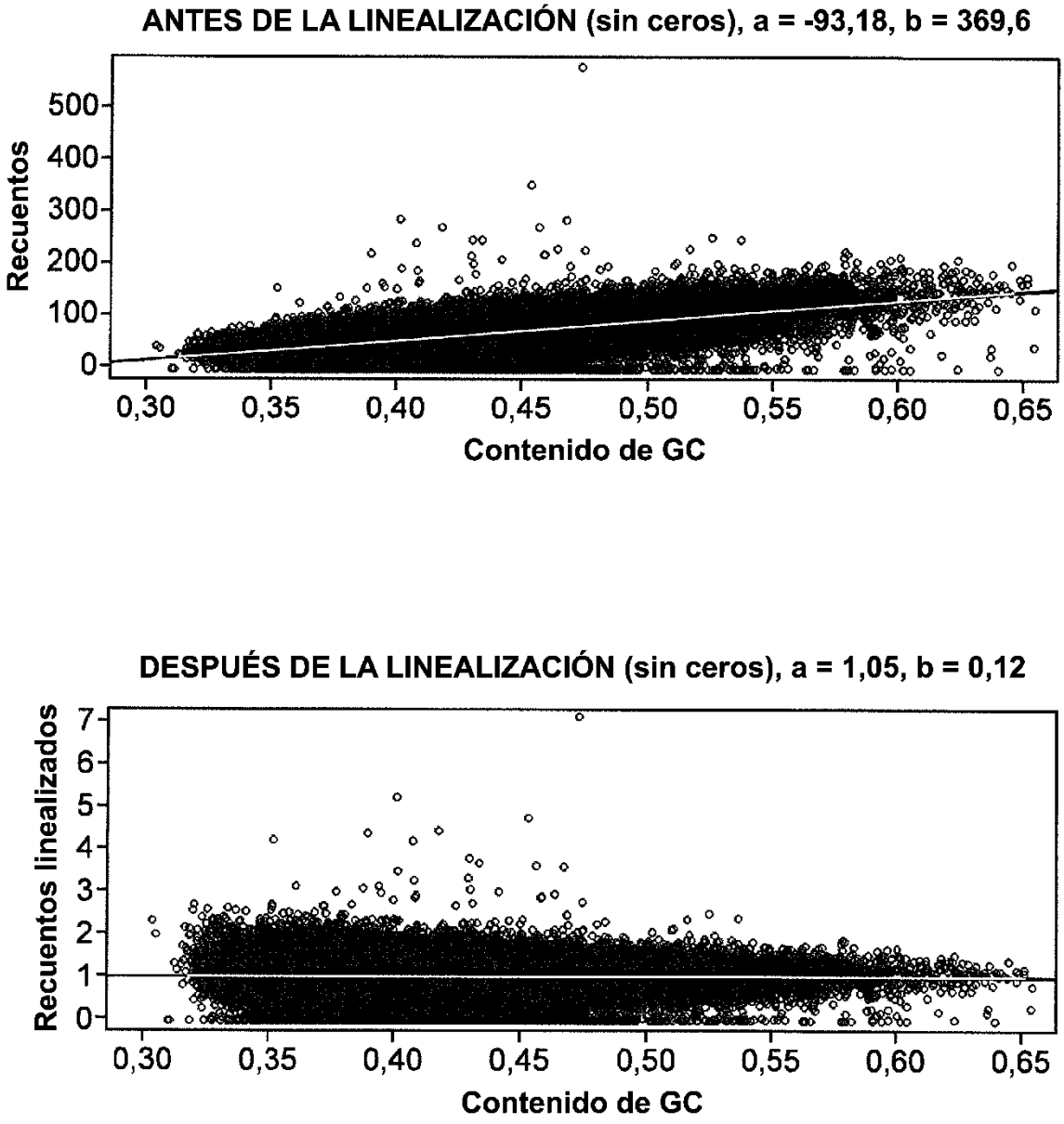

ANTES DE LA LINEALIZACIÓN (sin ceros), a = -93,18, b = 369,6
Recuentos
500
400
300
200
100
0
0,30
0,35
0,40
0,45
0,50
0,55
0,60
0,65
Contenido de GC
DESPUÉS DE LA LINEALIZACIÓN (sin ceros), a = 1,05, b = 0,12
Recuentos linealizados
7
6
5
4
3
2
1
0
0,30
0,35
0,40
0,45
0,50
0,55
0,60
0,65
Contenido de GC

FIG. 70

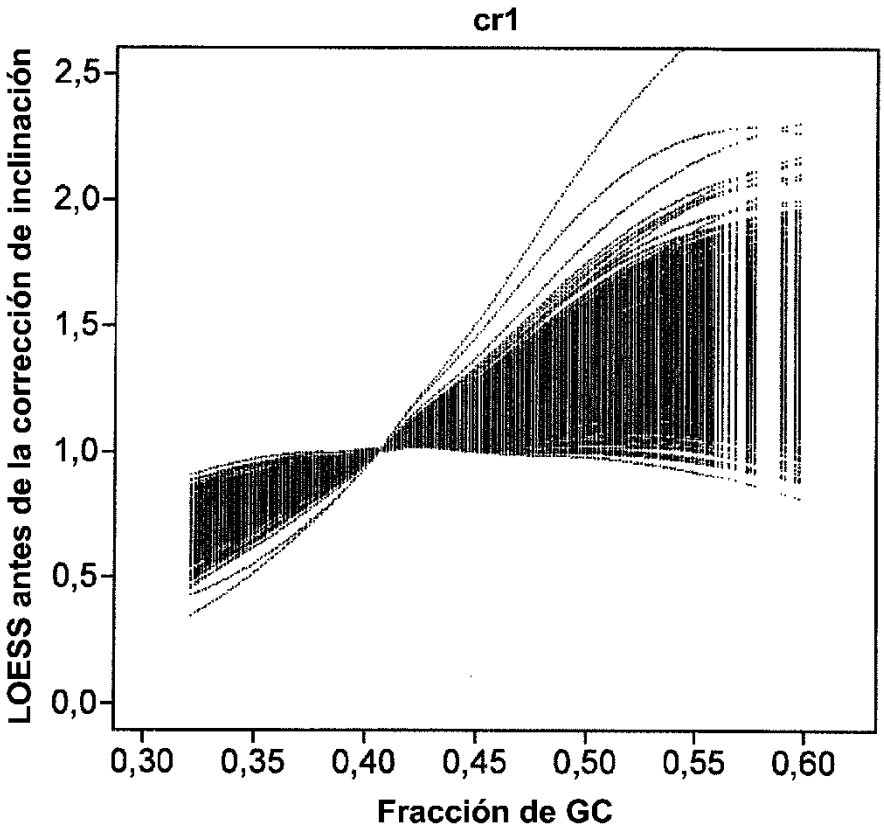

cr1
LOESS antes de la corrección de inclinación
2,5
2,0
1,5
1,0
0,5
0,0
0,30
0,35
0,40
0,45
0,50
0,55
0,60
Fracción de GC

FIG. 71

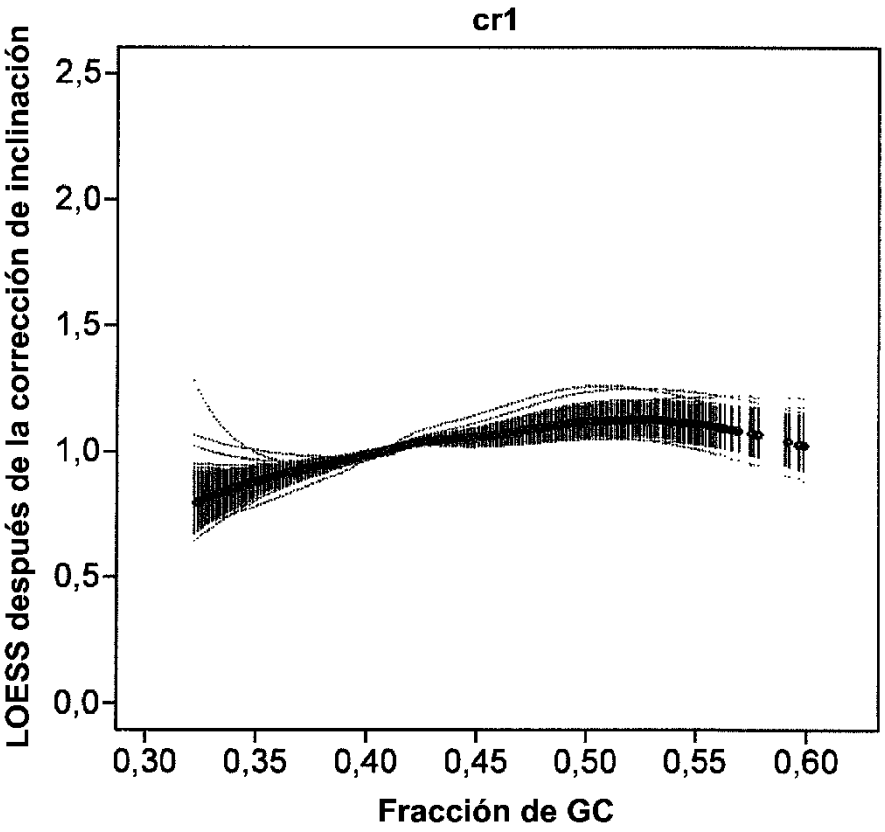

cr1
LOESS después de la corrección de inclinación
2,5
2,0
1,5
1,0
0,5
0,0
0,30
0,35
0,40
0,45
0,50
0,55
0,60
Fracción de GC

FIG. 72

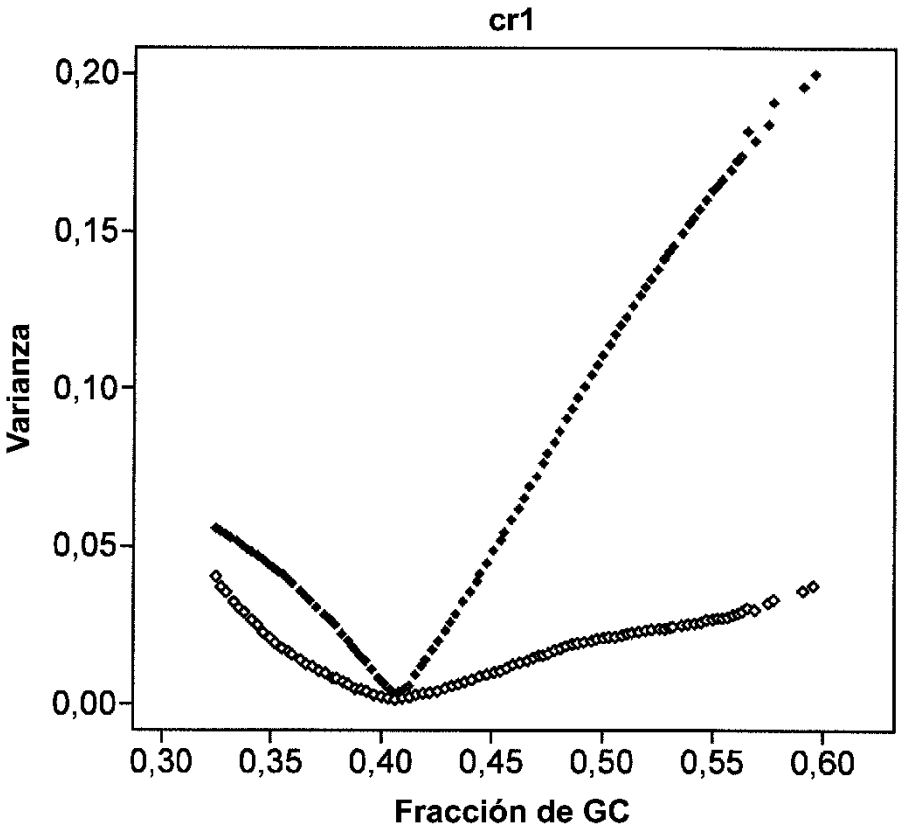

cr1
0,20
0,15
0,10
0,05
0,00
Varianza
0,30
0,35
0,40
0,45
0,50
0,55
0,60
Fracción de GC

FIG. 73

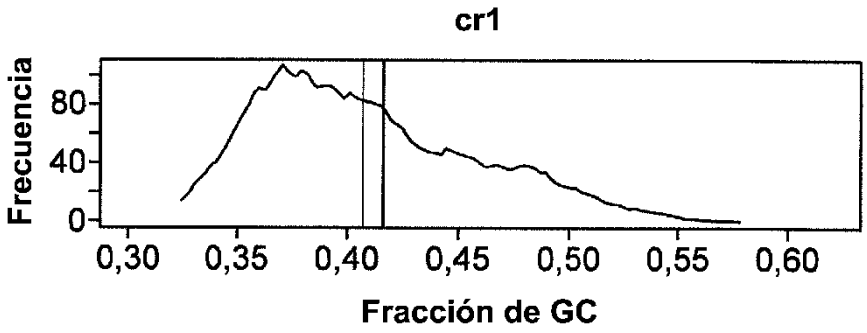

cr1
Frecuencia
80
40
0
0,30
0,35
0,40
0,45
0,50
0,55
0,60
Fracción de GC

FIG. 74

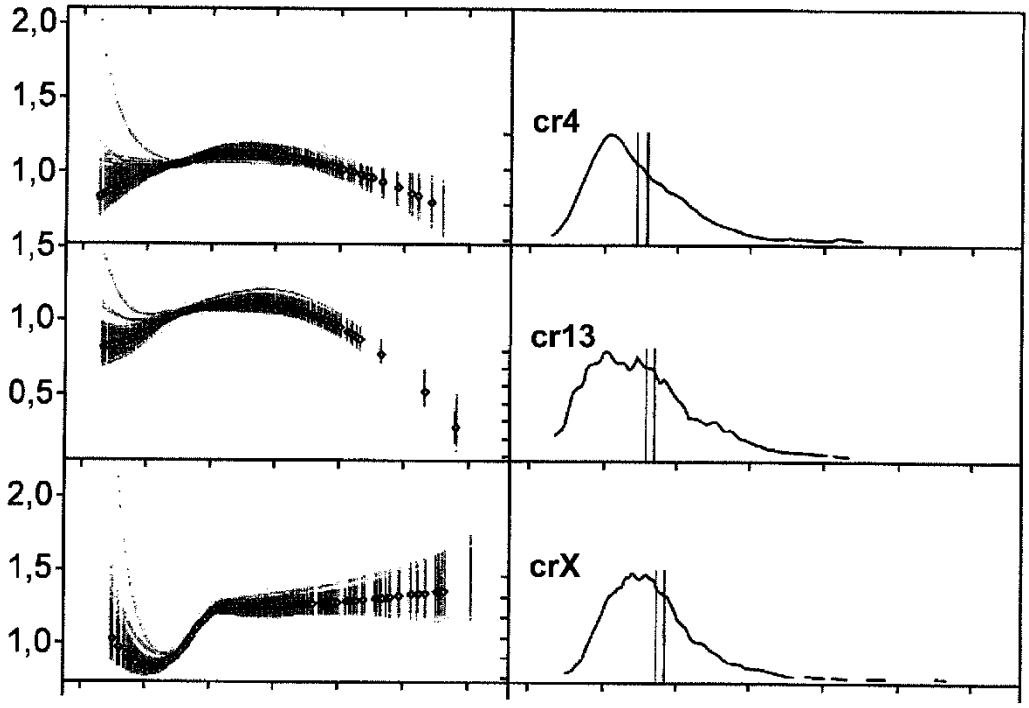


FIG. 75A

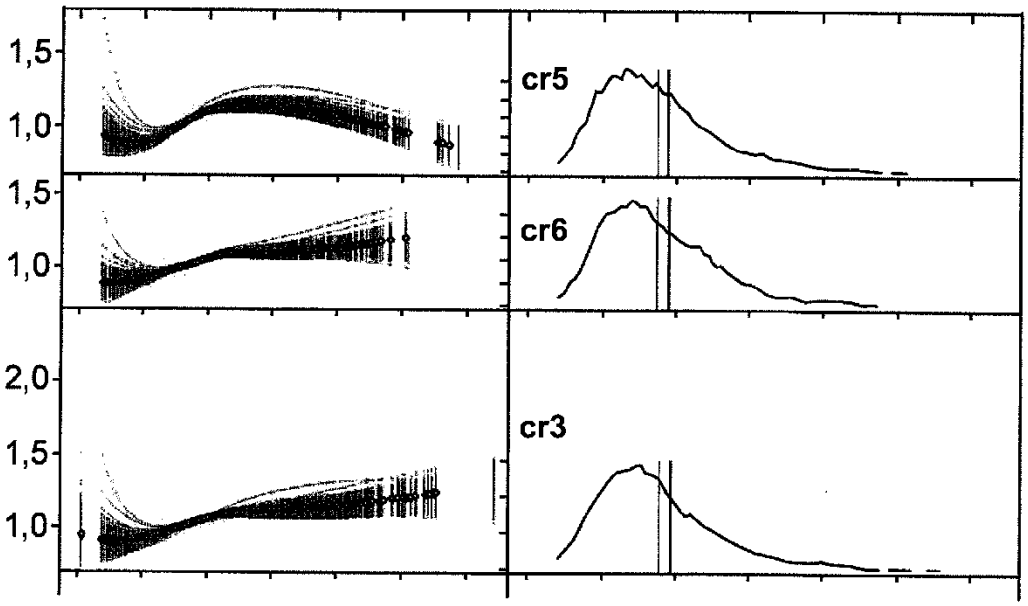


FIG. 75B

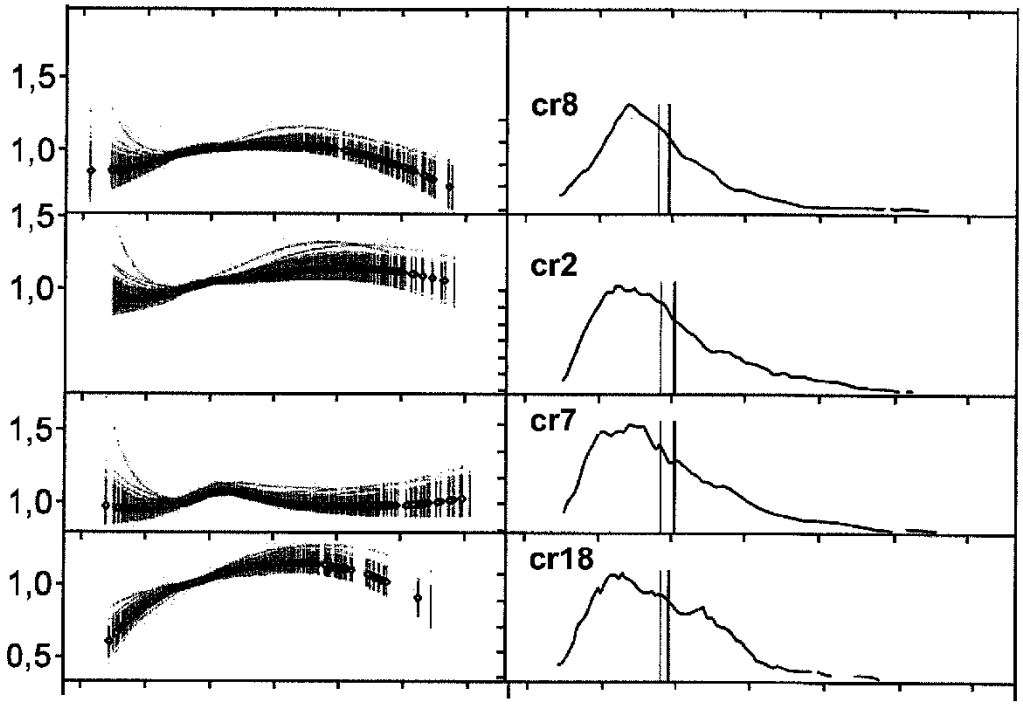


FIG. 75C

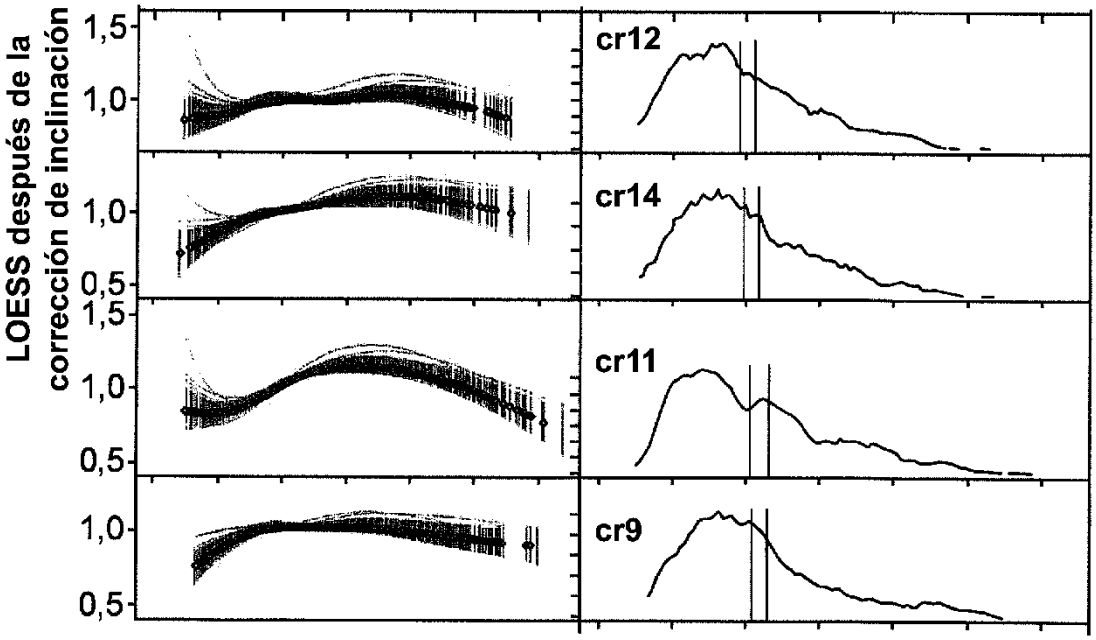


FIG. 75D

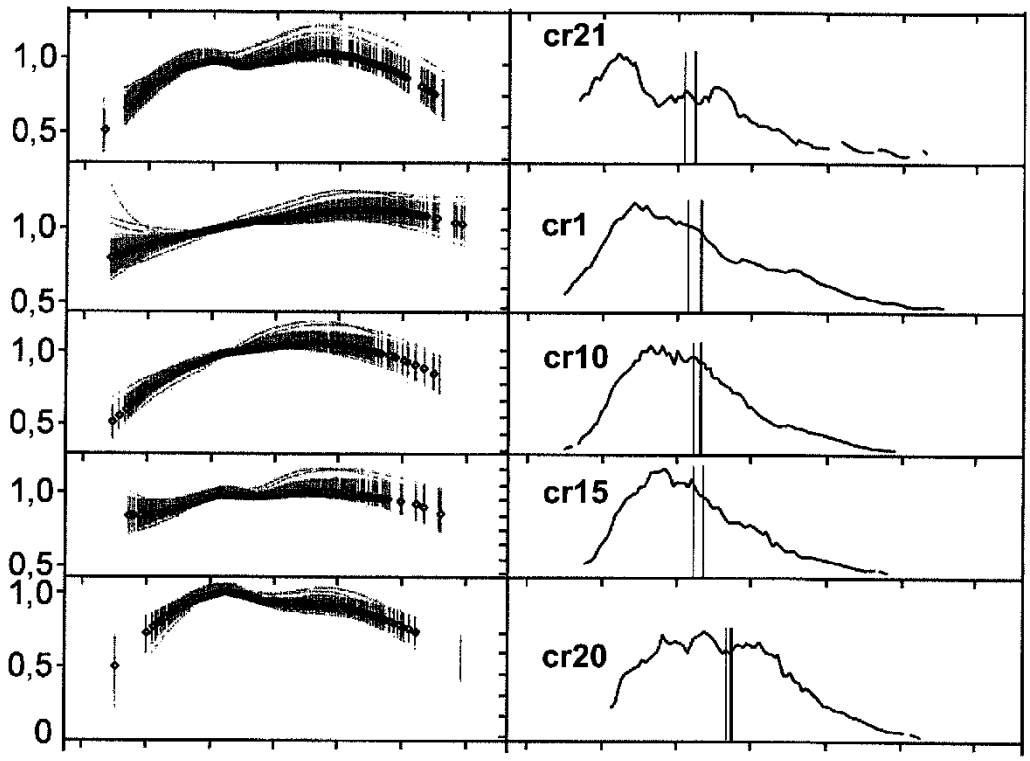


FIG. 75E

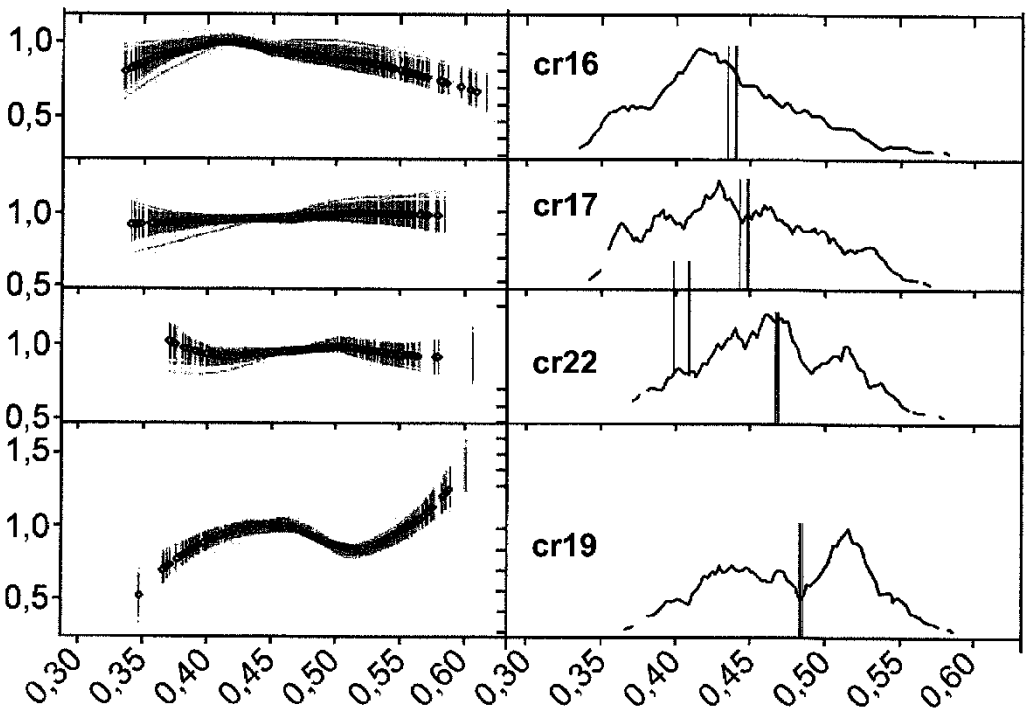


FIG. 75F

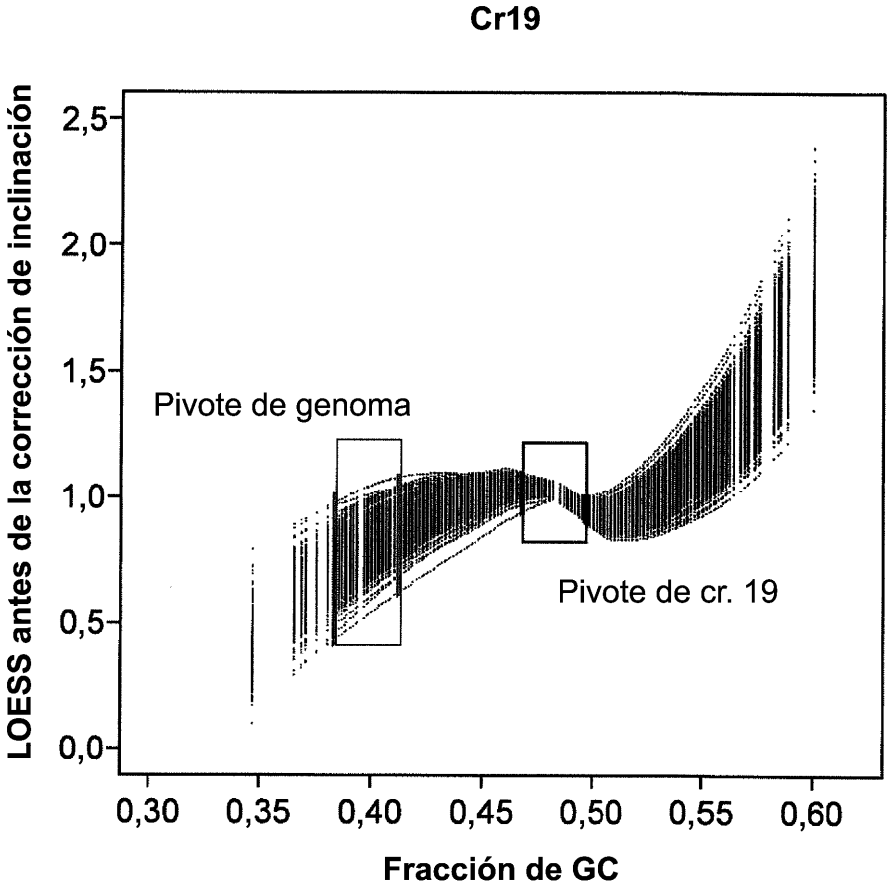

Cr19
LOESS antes de la corrección de inclinación
2,5
2,0
1,5
1,0
0,5
0,0
Pivote de genoma
Pivote de cr. 19
0,30
0,35
0,40
0,45
0,50
0,55
0,60
Fracción de GC

FIG. 76

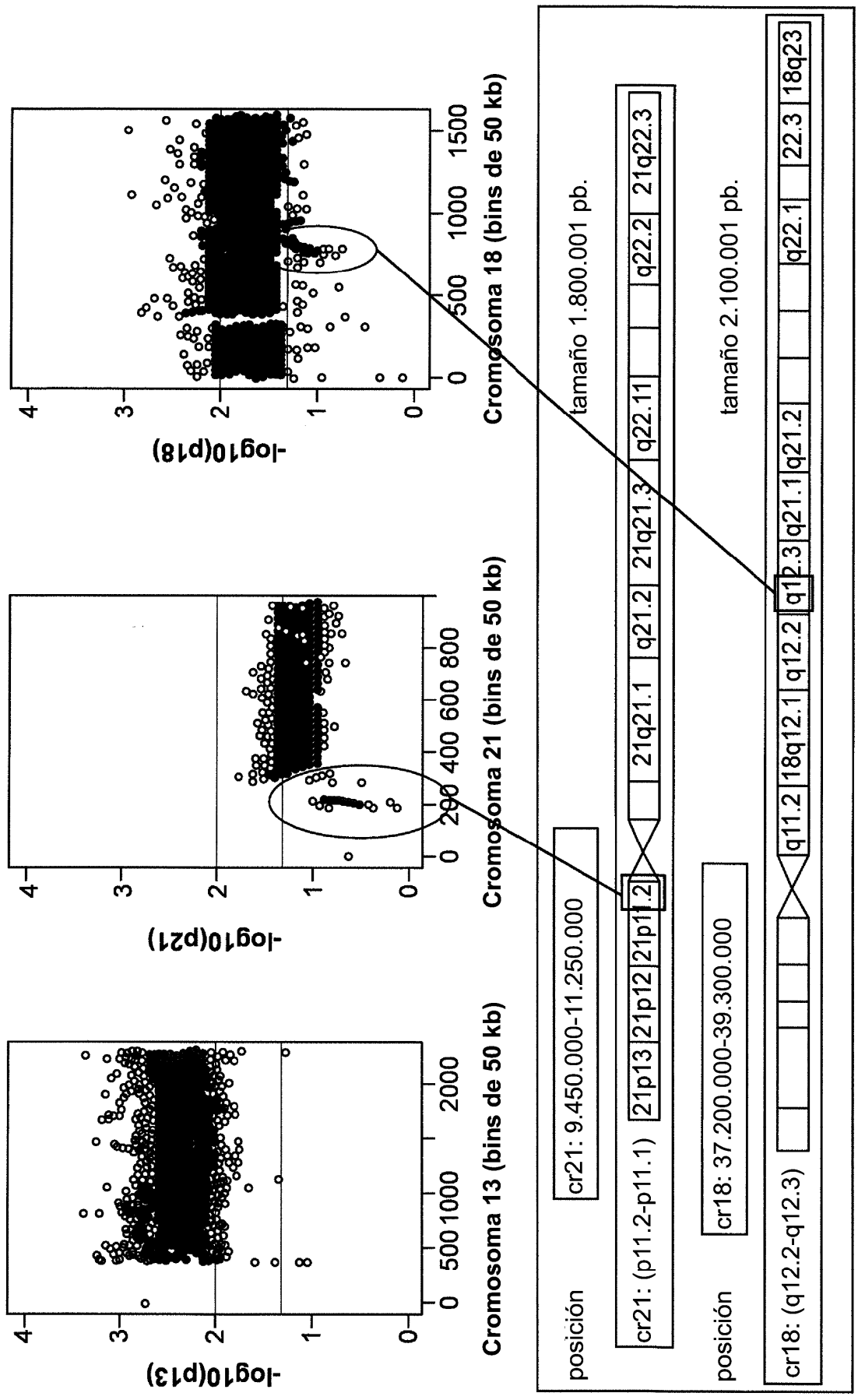

-log10(p13)
Cromosoma 13 (bins de 50 kb)
-log10(p21)
Cromosoma 21 (bins de 50 kb)
-log10(p18)
Cromosoma 18 (bins de 50 kb)
posición
cr21: 9.450.000-11.250.000
tamaño 1.800.001 pb.
cr21: (p11.2-p11.1)
21p13
21p12
21p11.2
21q21.1
q21.2
21q21.3
q22.11
q22.2
21q22.3
posición
cr18: 37.200.000-39.300.000
tamaño 2.100.001 pb.
cr18: (q12.2-q12.3)
q11.2
18q12.1
q12.2
q12.3
q21.1
q21.2
q22.1
22.3
18q23

FIG. 77

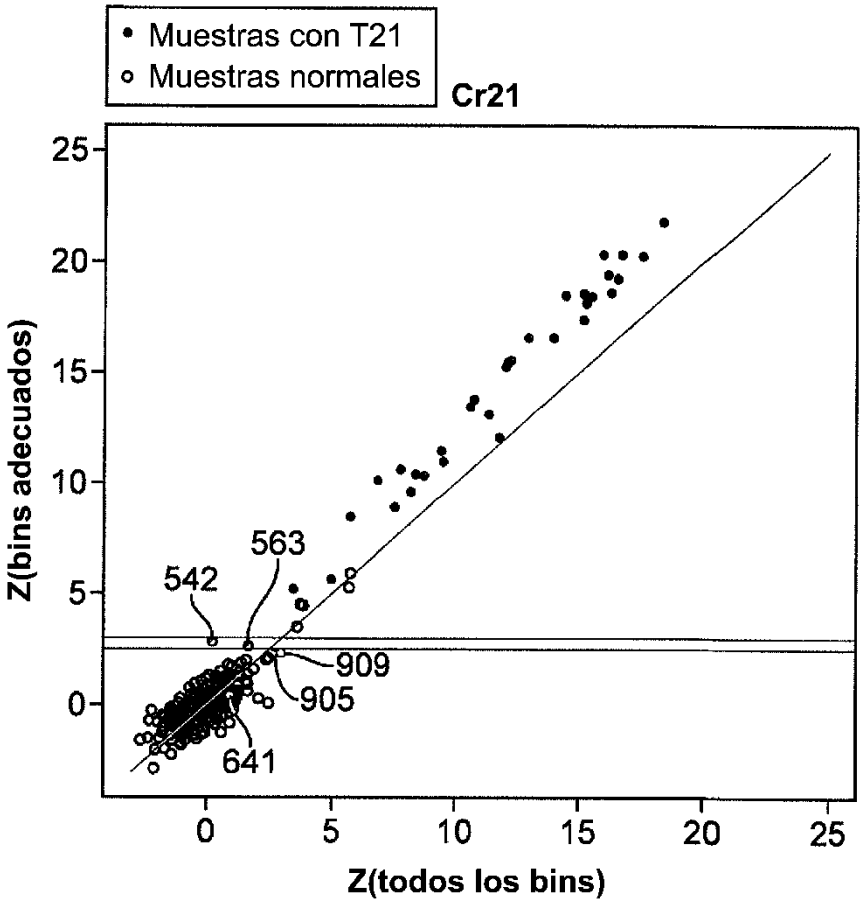

Muestras con T21
Muestras normales
Cr21
Z(bins adecuados)
Z(todos los bins)
25
20
15
10
5
0
0
5
10
15
20
25
563
542
909
905
641

FIG. 78

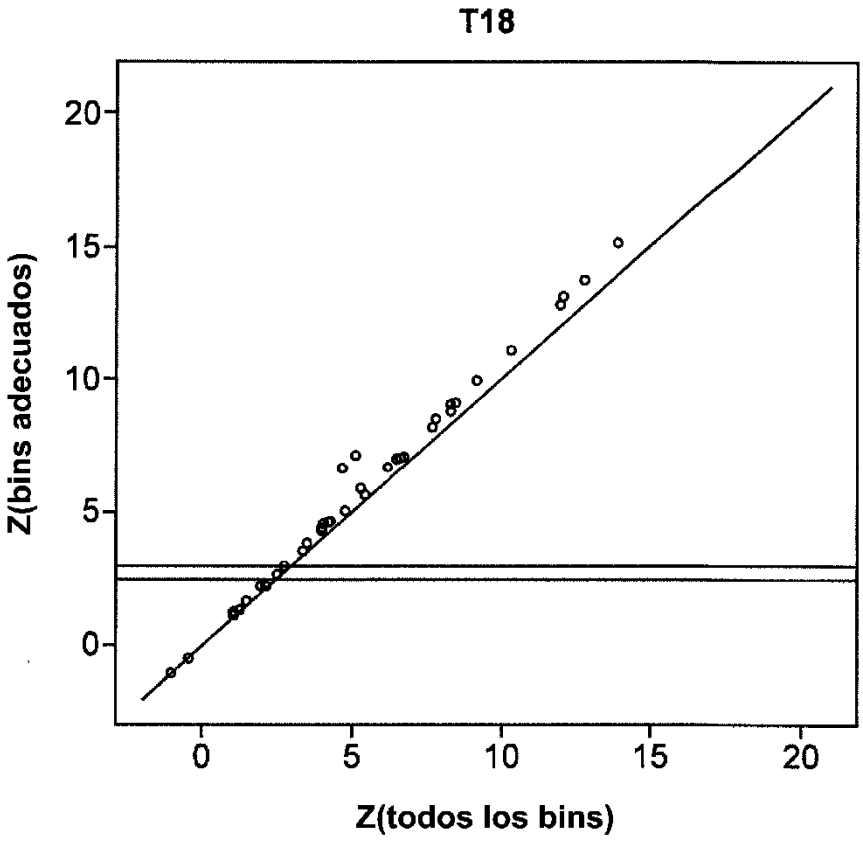


FIG. 79

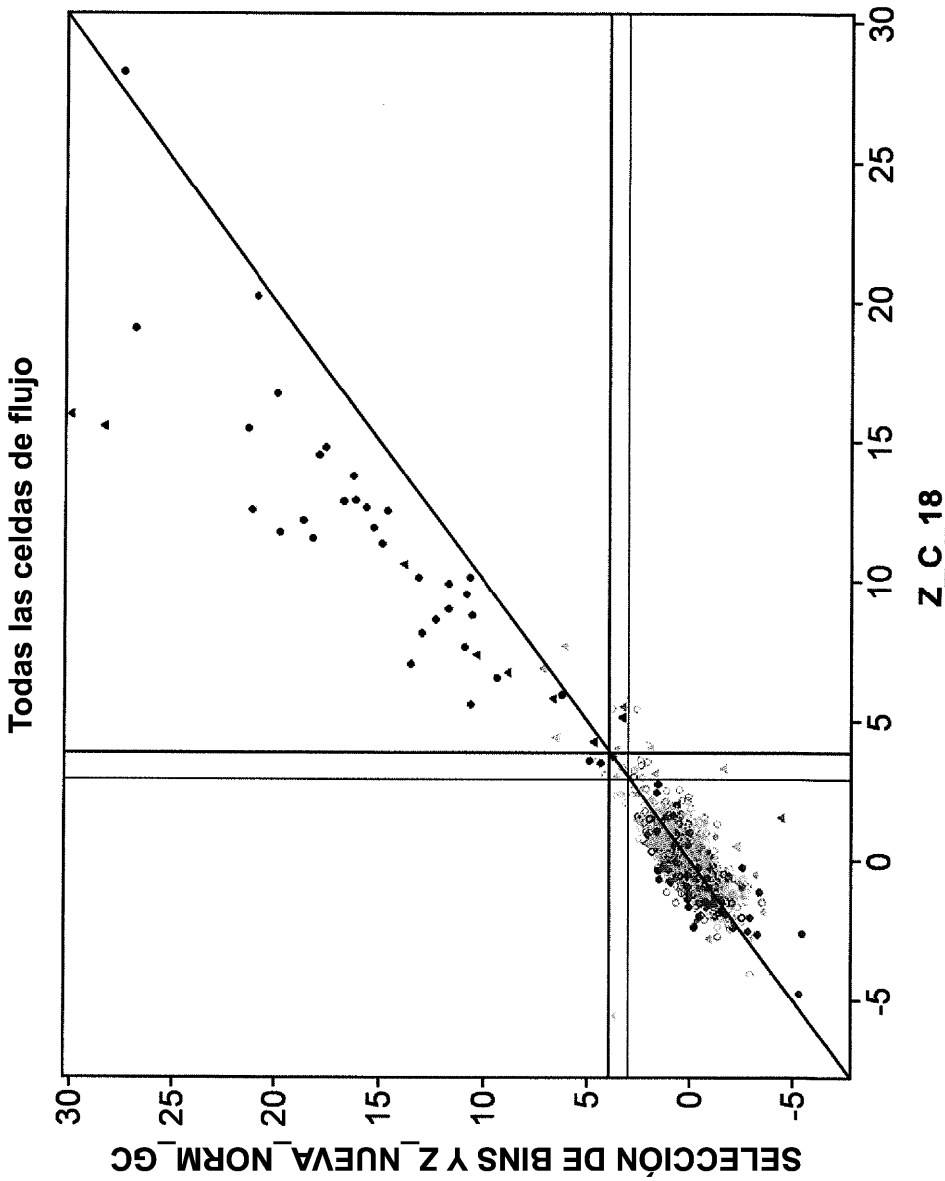

Todas las celdas de flujo
SELECCIÓN DE BINS Y Z_NUEVA_NORM_GC
Z_C_18
30
25
20
15
10
5
0
-5

FIG. 80

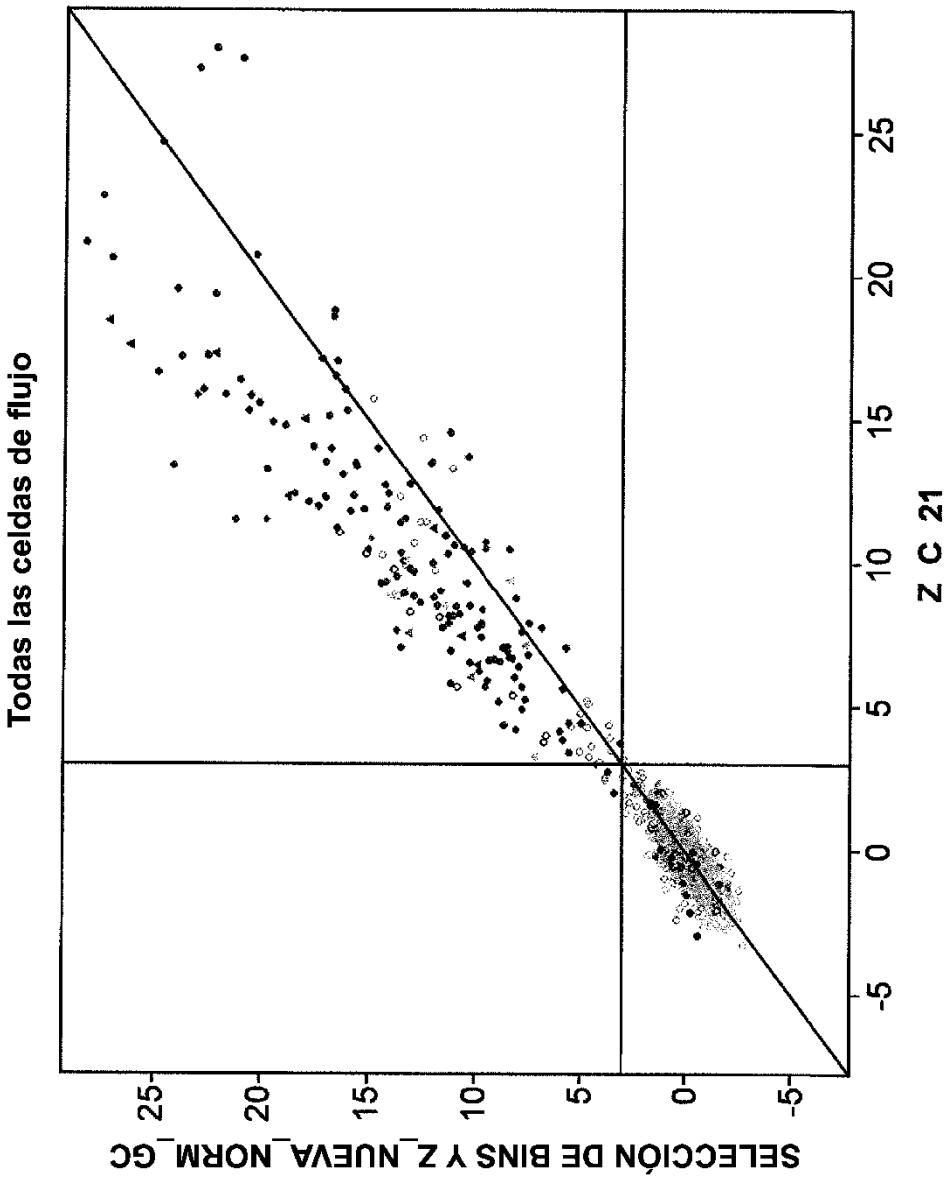

Todas las celdas de flujo
SELECCIÓN DE BINS Y Z_NUEVA_NORM_GC
Z_C_21
25
20
15
10
5
0
-5

FIG. 81

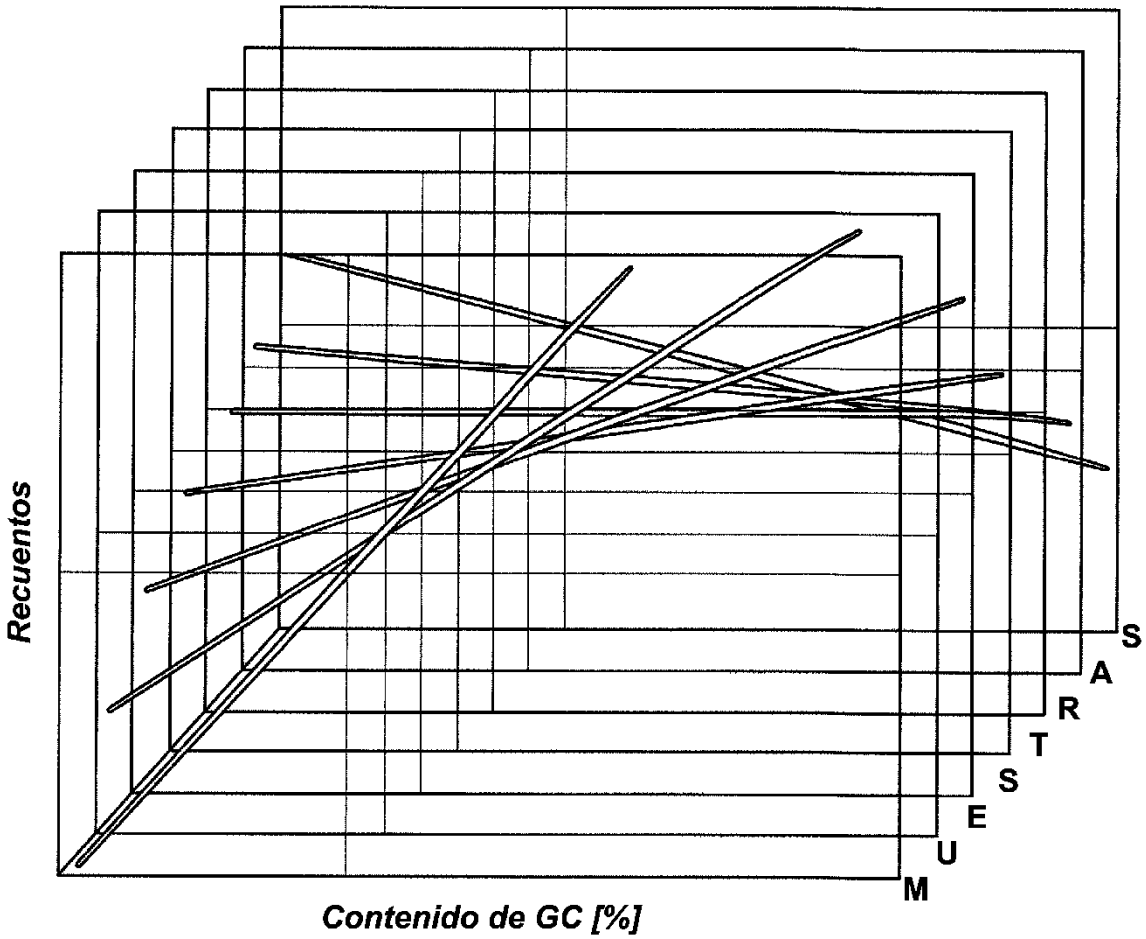


FIG. 82

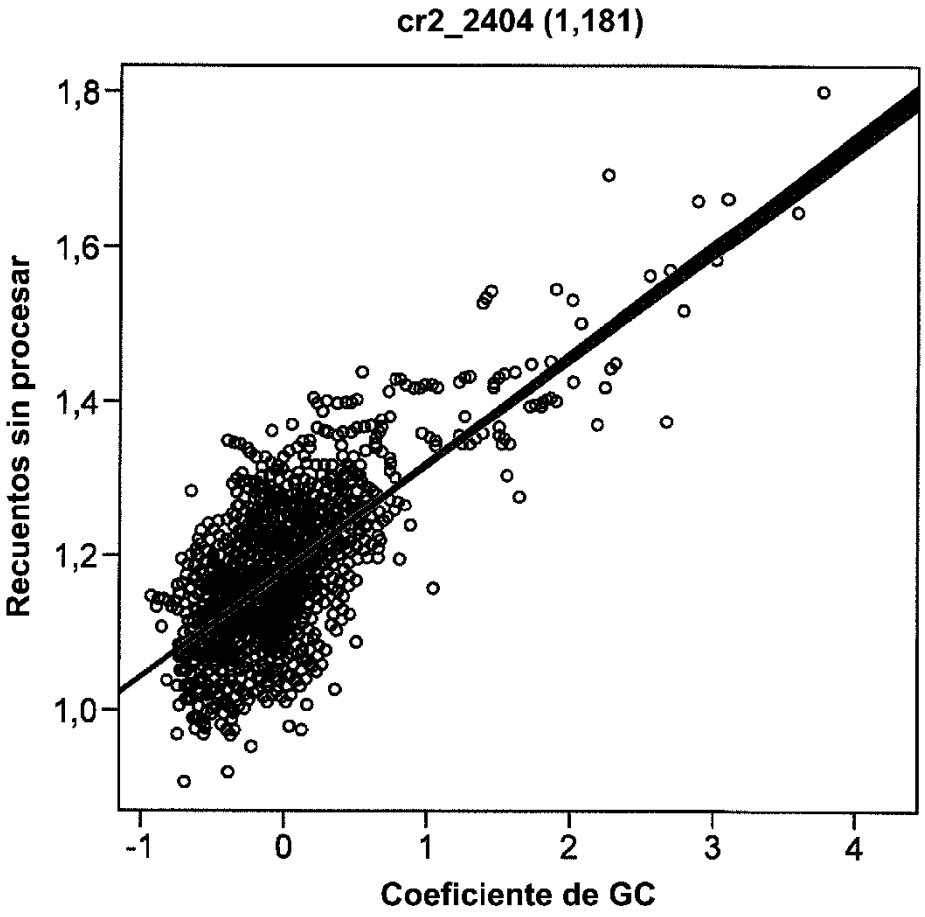

cr2_2404 (1,181)
Recuentos sin procesar
1,8
1,6
1,4
1,2
1,0
-1
0
1
2
3
4
Coeficiente de GC

FIG. 83

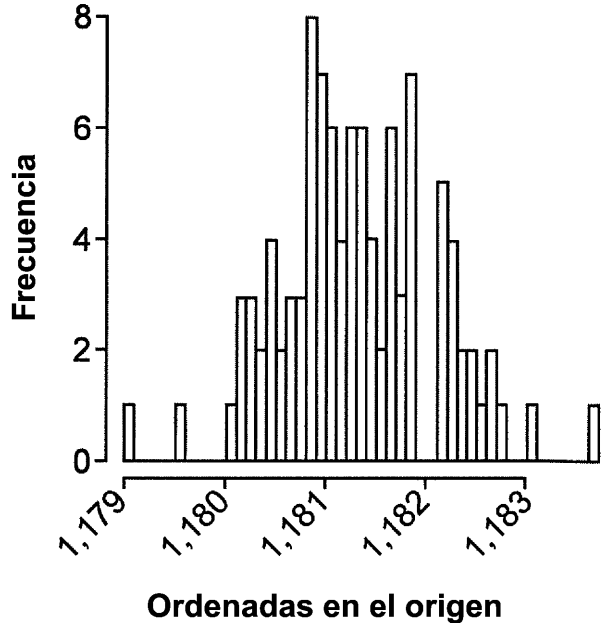

Frecuencia
8
6
4
2
0
1,179
1,180
1,181
1,182
1,183
Ordenadas en el origen

FIG. 84

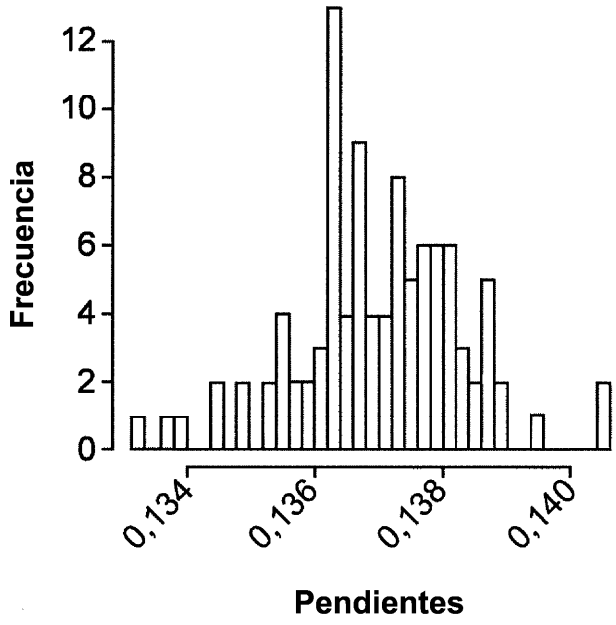

Frecuencia
12
10
8
6
4
2
0
0,134
0,136
0,138
0,140
Pendientes

FIG. 85

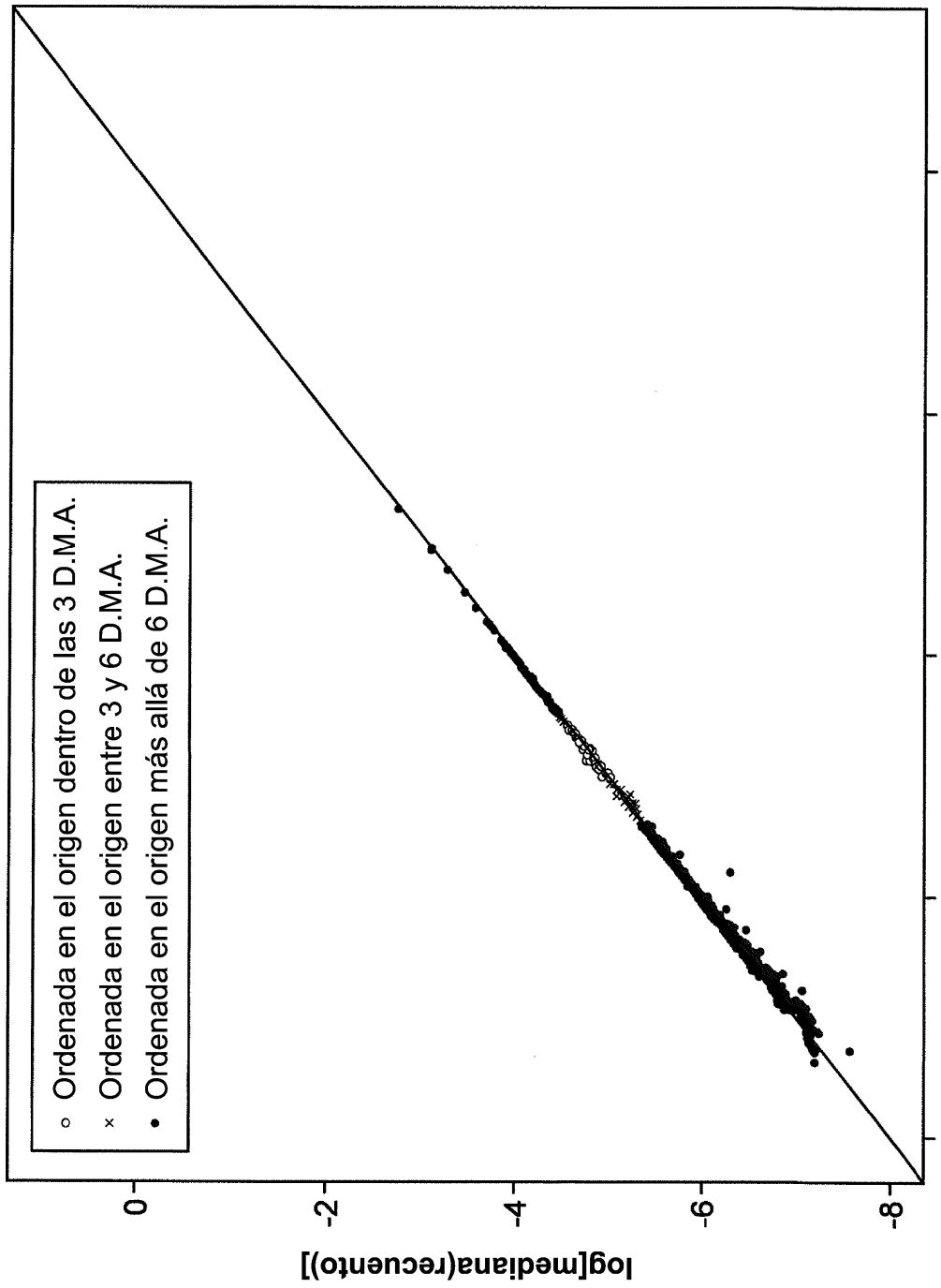

log[mediana(recuento)]
0
-2
-4
-6
-8
log(ordenada en el origen)
0
-2
-4
-6
-8
○ Ordenada en el origen dentro de las 3 D.M.A.
× Ordenada en el origen entre 3 y 6 D.M.A.
• Ordenada en el origen más allá de 6 D.M.A.

FIG. 86

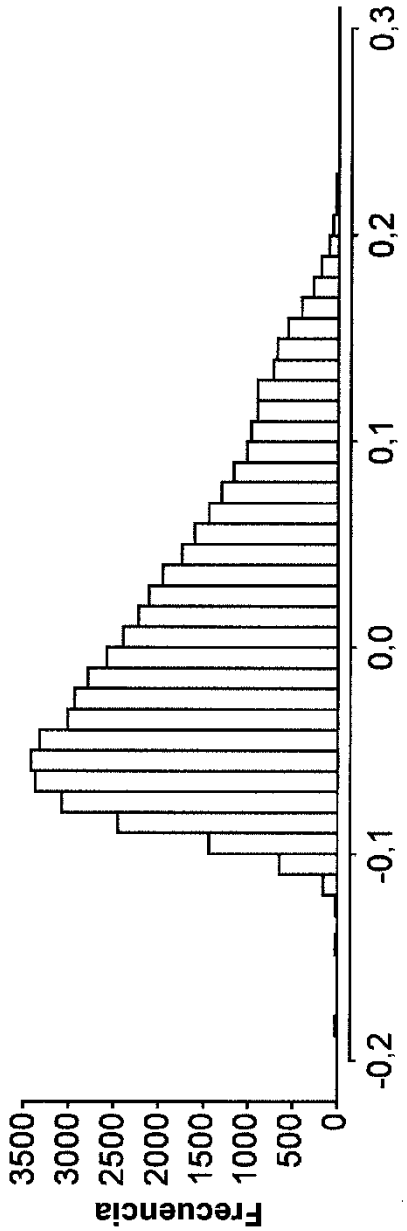

Frecuencia
3500
3000
2500
2000
1500
1000
500
0
-0,2
-0,1
0,0
0,1
0,2
0,3
Pendientes

FIG. 87

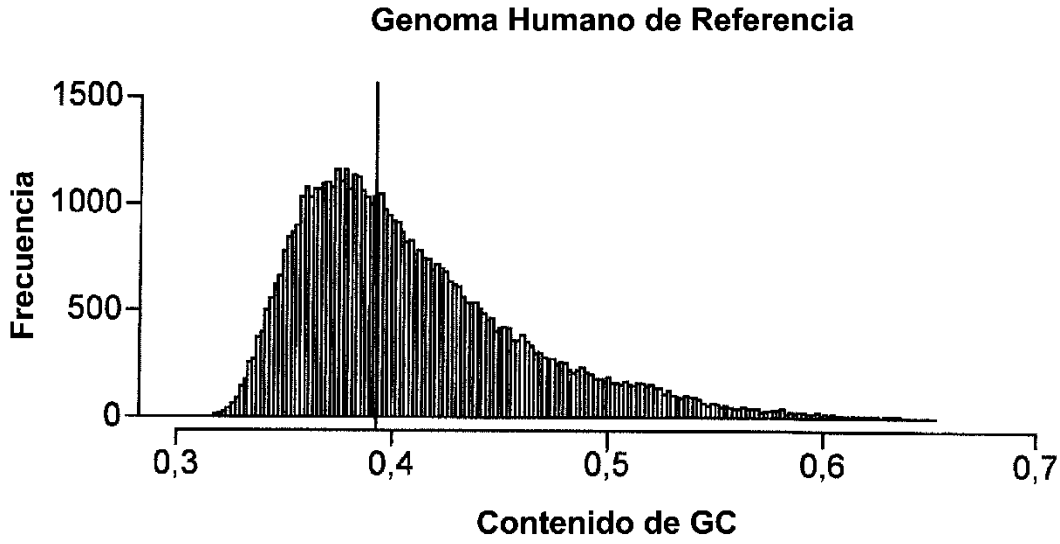

Genoma Humano de Referencia
1500
1000
500
0
Frecuencia
0,3
0,4
0,5
0,6
0,7
Contenido de GC

FIG. 88

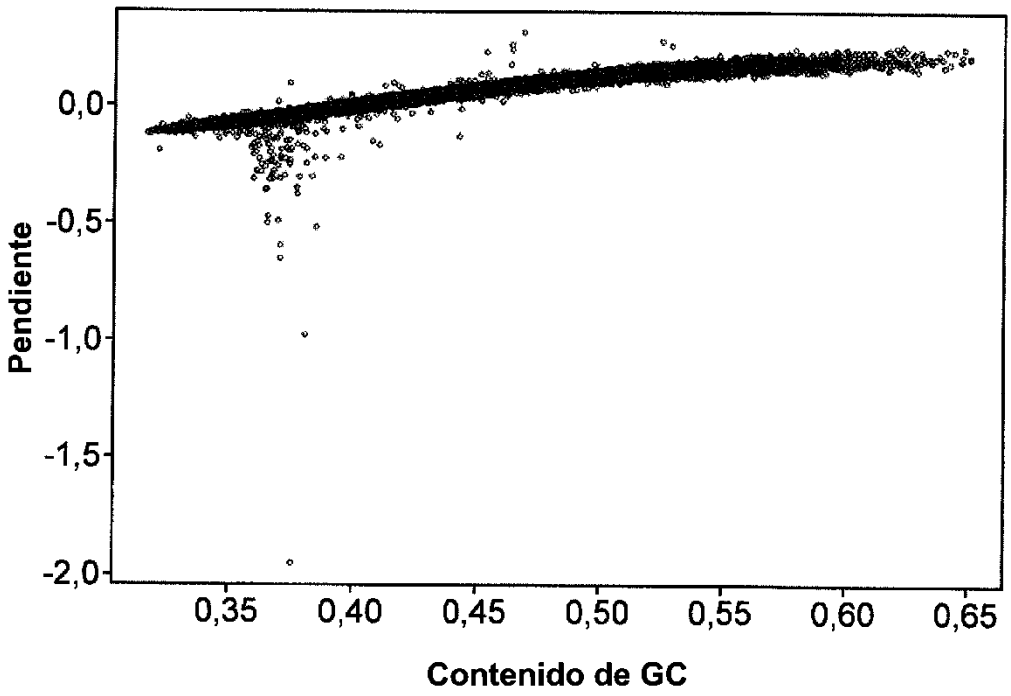

0,0
-0,5
-1,0
-1,5
-2,0
Pendiente
0,35
0,40
0,45
0,50
0,55
0,60
0,65
Contenido de GC

FIG. 89

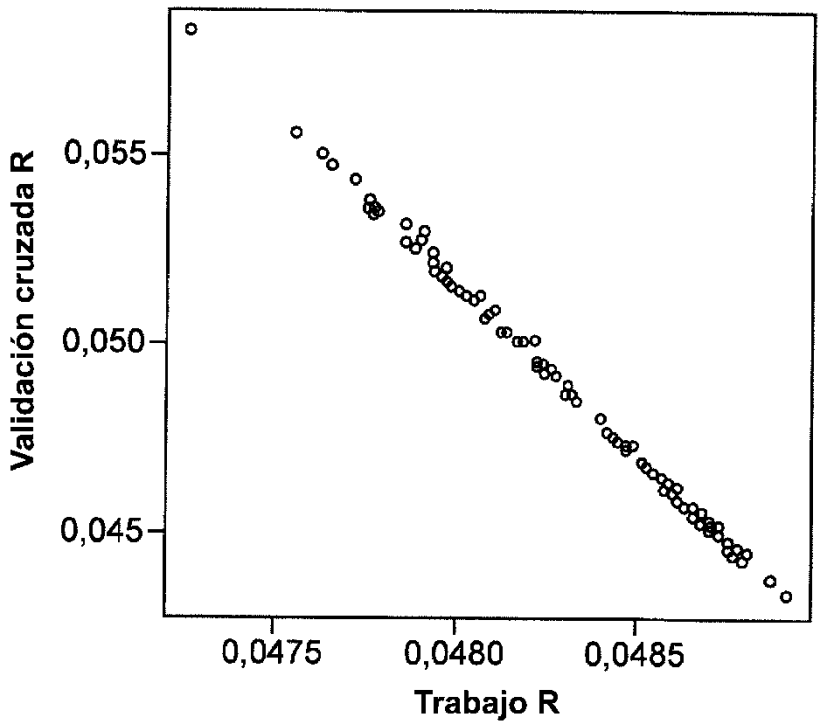

Validación cruzada R
0,055
0,050
0,045
0,0475
0,0480
0,0485
Trabajo R

FIG. 90

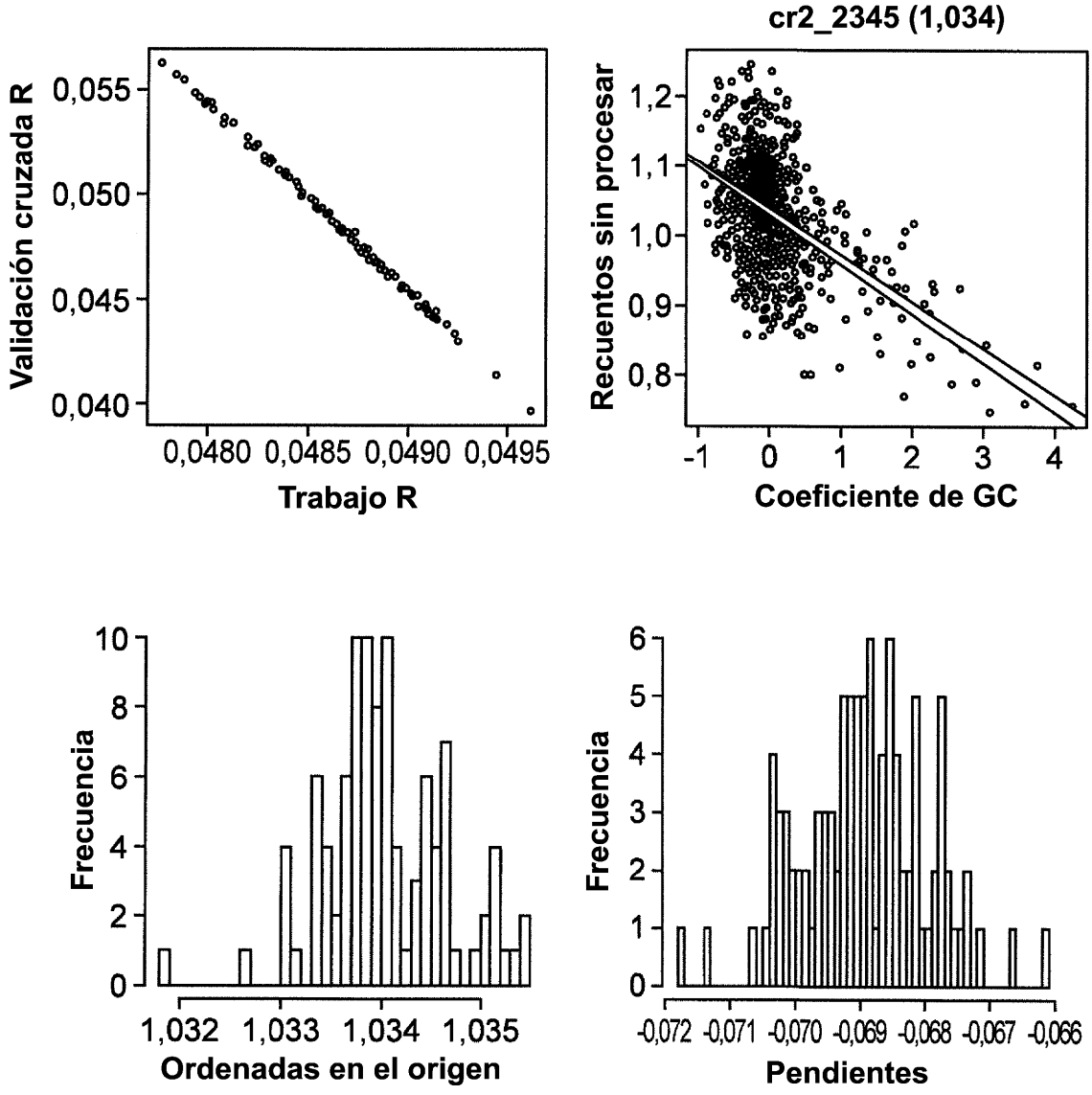

Validación cruzada R
0,055
0,050
0,045
0,040
0,0480 0,0485 0,0490 0,0495
Trabajo R
cr2_2345 (1,034)
Recuentos sin procesar
1,2
1,1
1,0
0,9
0,8
-1 0 1 2 3 4
Coeficiente de GC
Frecuencia
10
8
6
4
2
0
1,032 1,033 1,034 1,035
Ordenadas en el origen
Frecuencia
6
5
4
3
2
1
0
-0,072 -0,071 -0,070 -0,069 -0,068 -0,067 -0,066
Pendientes

FIG. 91

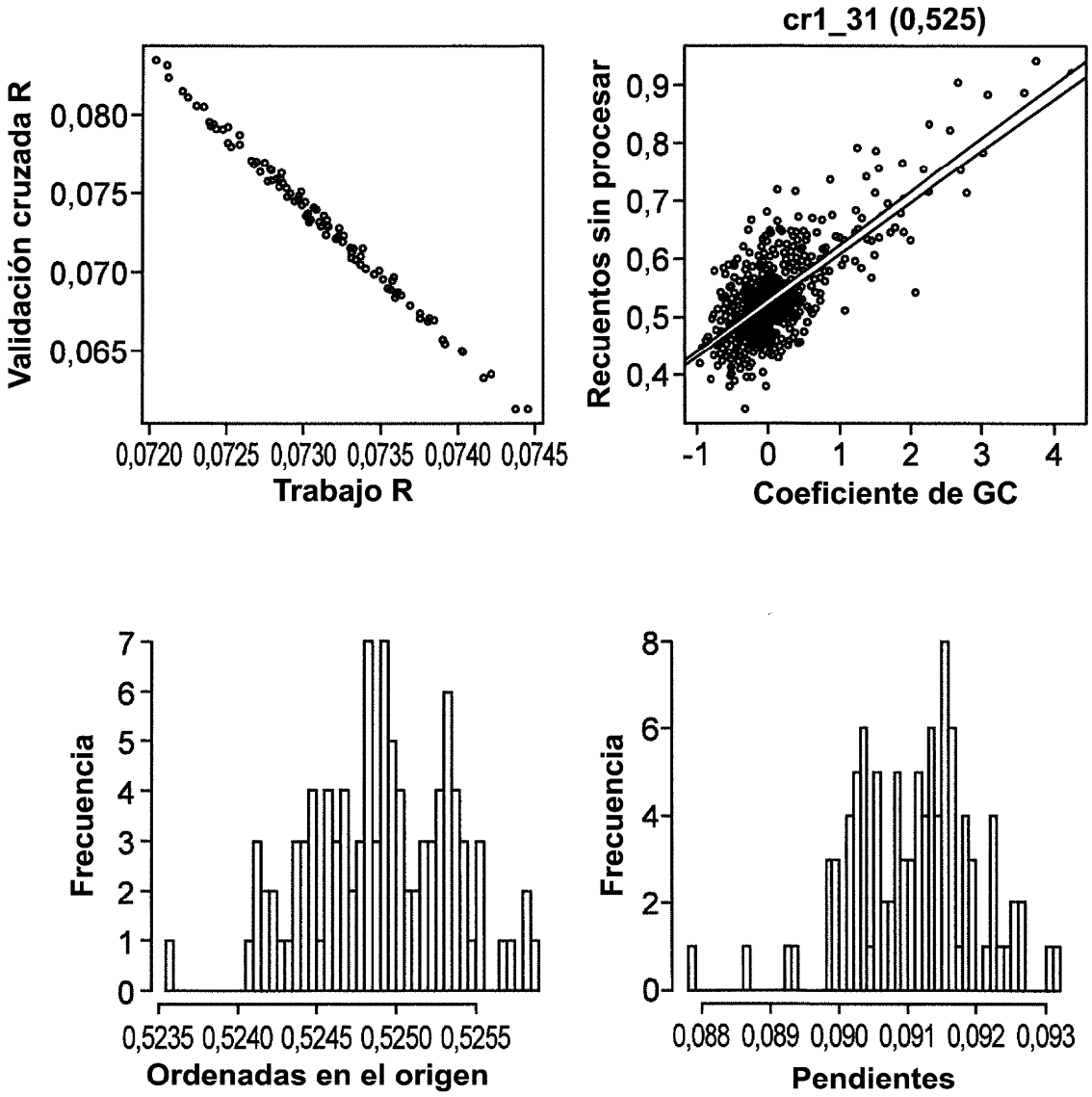

Validación cruzada R
0,080
0,075
0,070
0,065
0,0720 0,0725 0,0730 0,0735 0,0740 0,0745
Trabajo R
cr1_31 (0,525)
Recuentos sin procesar
0,9
0,8
0,7
0,6
0,5
0,4
-1 0 1 2 3 4
Coeficiente de GC
Frecuencia
7
6
5
4
3
2
1
0
0,5235 0,5240 0,5245 0,5250 0,5255
Ordenadas en el origen
Frecuencia
8
6
4
2
0
0,088 0,089 0,090 0,091 0,092 0,093
Pendientes

FIG. 92

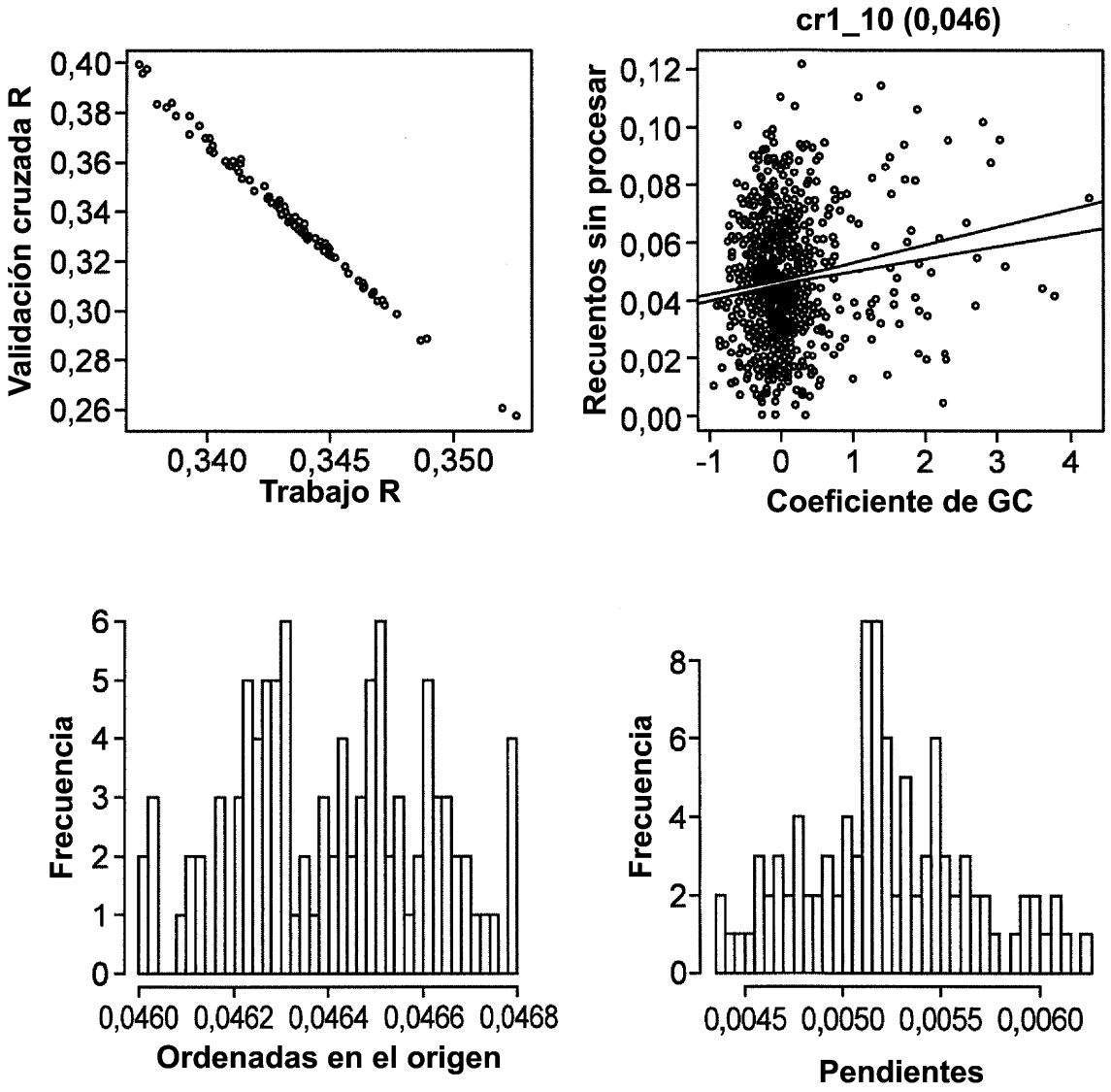

Validación cruzada R
0,40
0,38
0,36
0,34
0,32
0,30
0,28
0,26
0,340
0,345
0,350
Trabajo R
cr1_10 (0,046)
Recuentos sin procesar
0,12
0,10
0,08
0,06
0,04
0,02
0,00
-1
0
1
2
3
4
Coeficiente de GC
Frecuencia
6
5
4
3
2
1
0
0,0460
0,0462
0,0464
0,0466
0,0468
Ordenadas en el origen
Frecuencia
8
6
4
2
0
0,0045
0,0050
0,0055
0,0060
Pendientes

FIG. 93

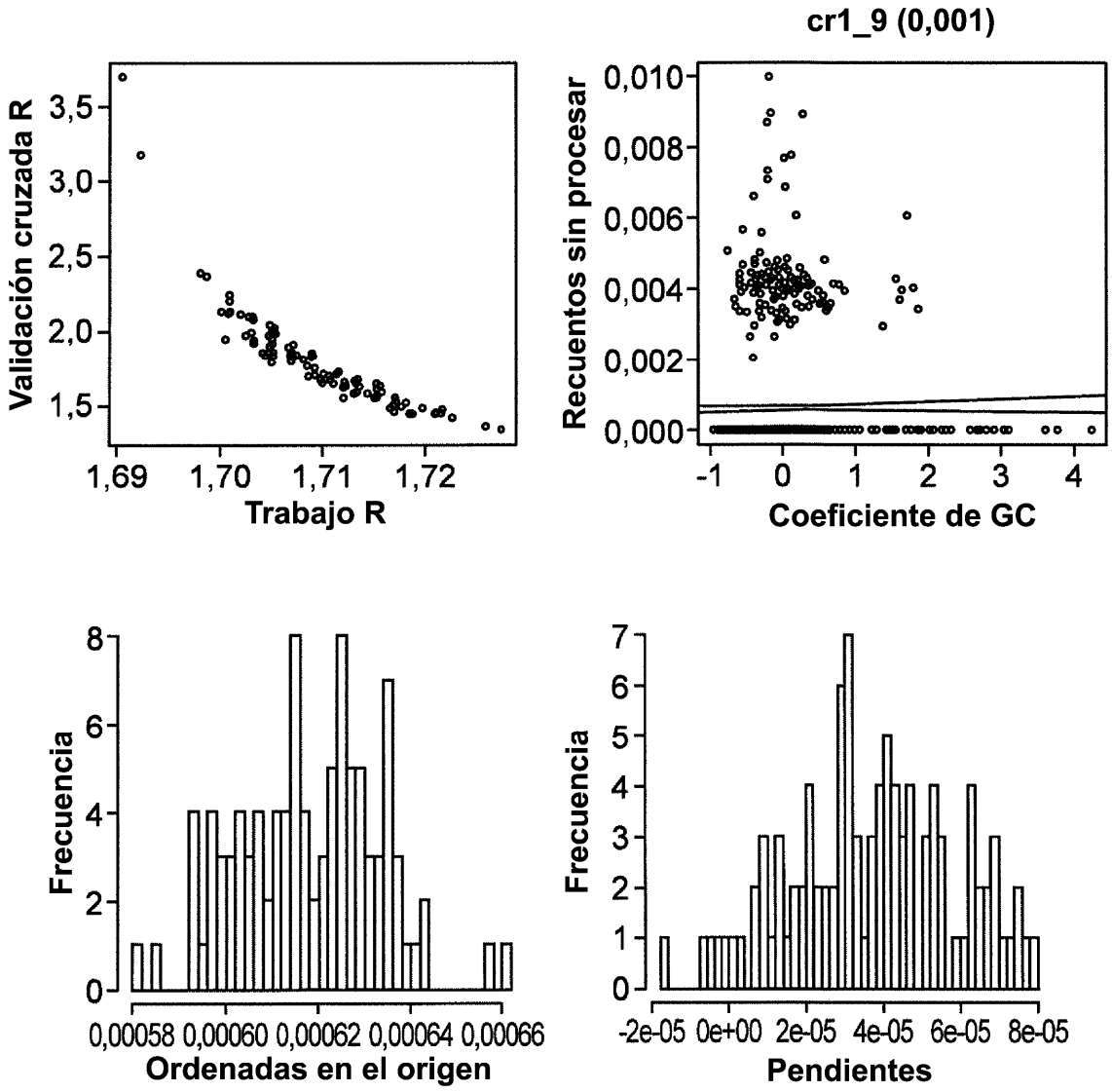

cr1_9 (0,001)
Validación cruzada R
3,5
3,0
2,5
2,0
1,5
1,69
1,70
1,71
1,72
Trabajo R
Recuentos sin procesar
0,010
0,008
0,006
0,004
0,002
0,000
-1
0
1
2
3
4
Coeficiente de GC
Frecuencia
8
6
4
2
0
0,00058
0,00060
0,00062
0,00064
0,00066
Ordenadas en el origen
Frecuencia
7
6
5
4
3
2
1
0
-2e-05
0e+00
2e-05
4e-05
6e-05
8e-05
Pendientes

FIG. 94

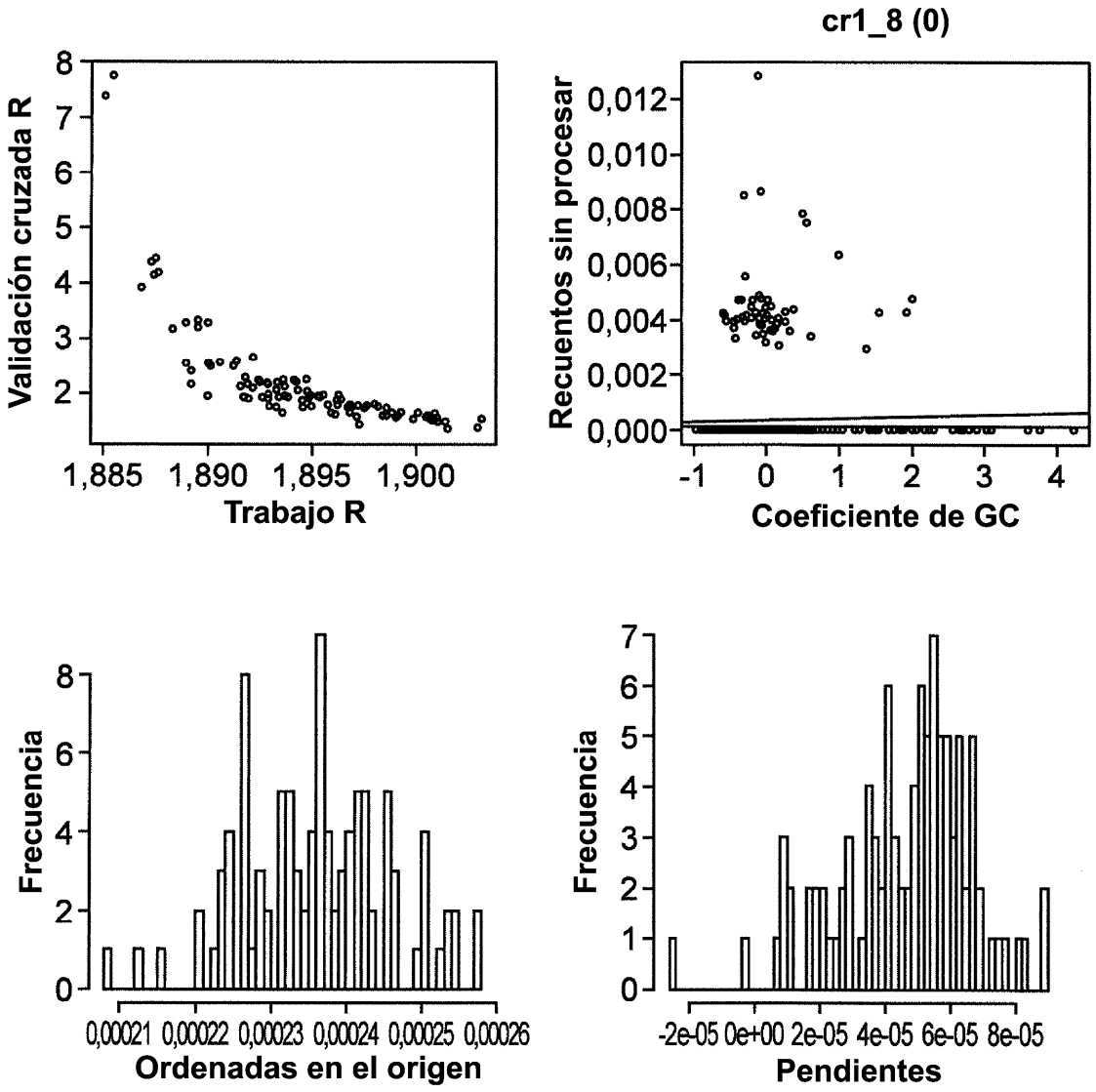

cr1_8 (0)
Validación cruzada R
Trabajo R
Recuentos sin procesar
Coeficiente de GC
Frecuencia
Ordenadas en el origen
Frecuencia
Pendientes

FIG. 95

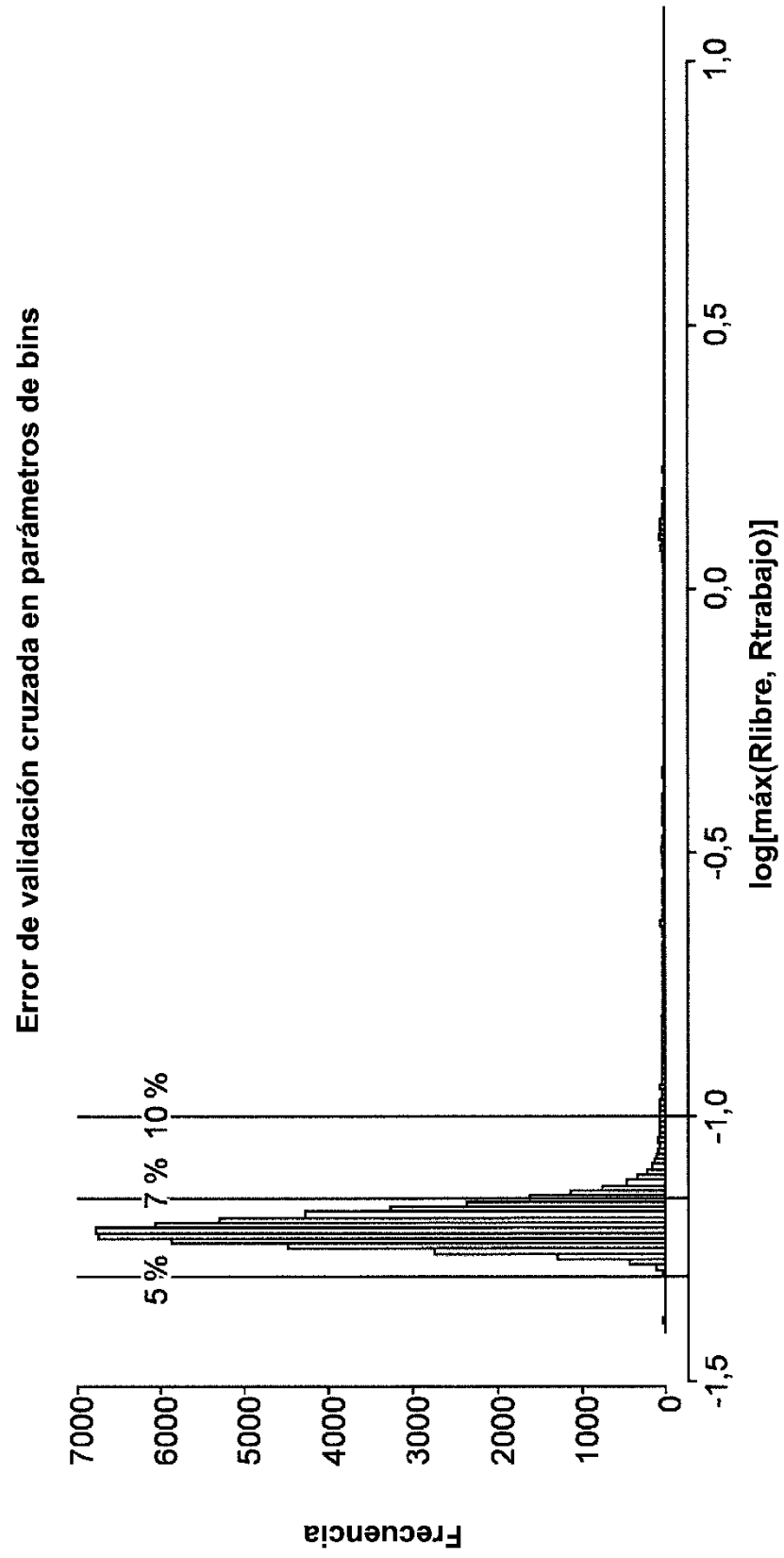

Error de validación cruzada en parámetros de bins
Frecuencia
7000
6000
5000
4000
3000
2000
1000
0
5 %
7 %
10 %
-1,5
-1,0
-0,5
0,0
0,5
1,0
log[máx(Rlibre, Rtrabajo)]

FIG. 96

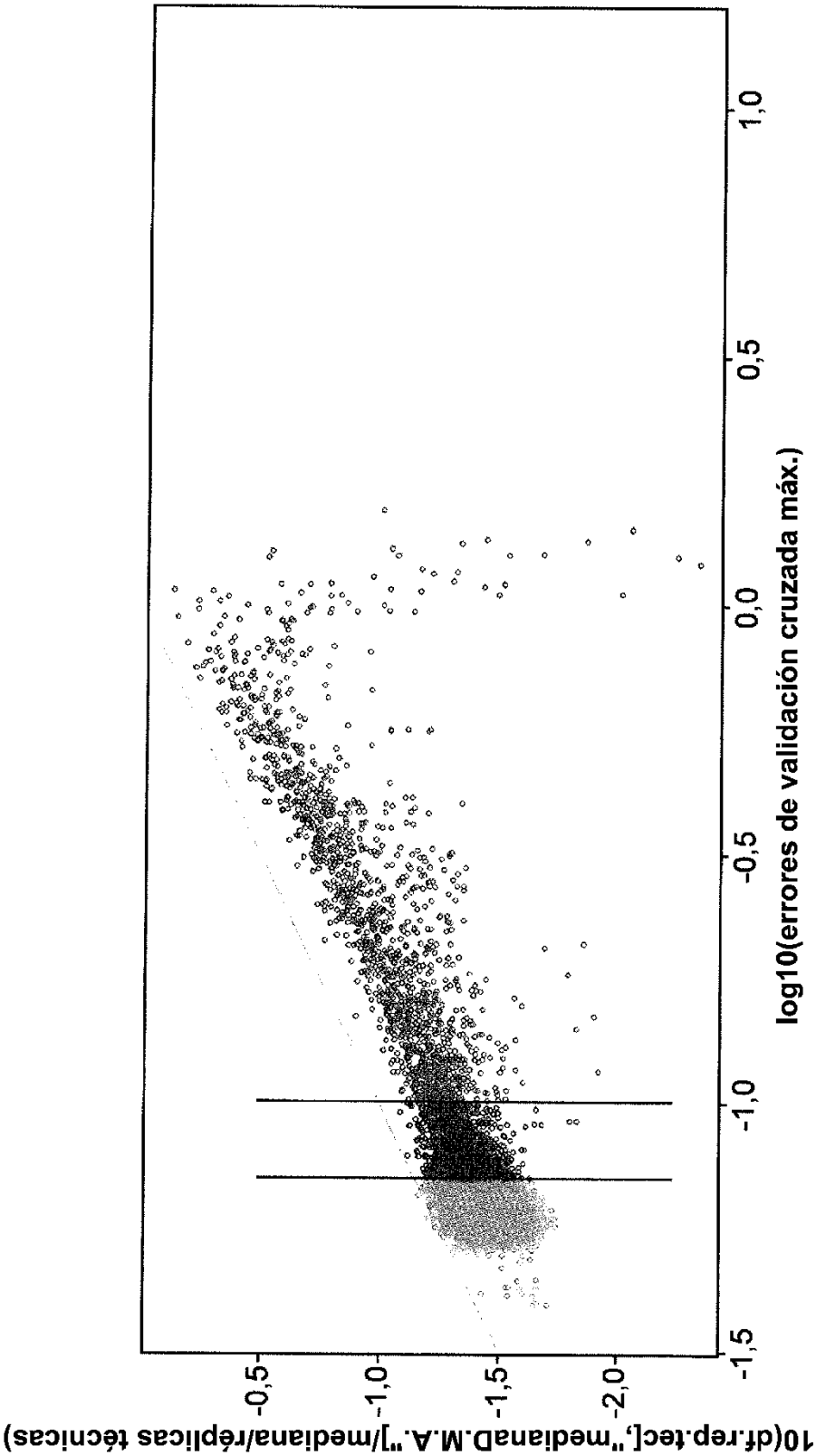

log10(df.rep.tec[,"medianaD.M.A."]/mediana/réplicas técnicas)
-0,5
-1,0
-1,5
-2,0
log10(errores de validación cruzada máx.)
-1,5
-1,0
-0,5
0,0
0,5
1,0

FIG. 97

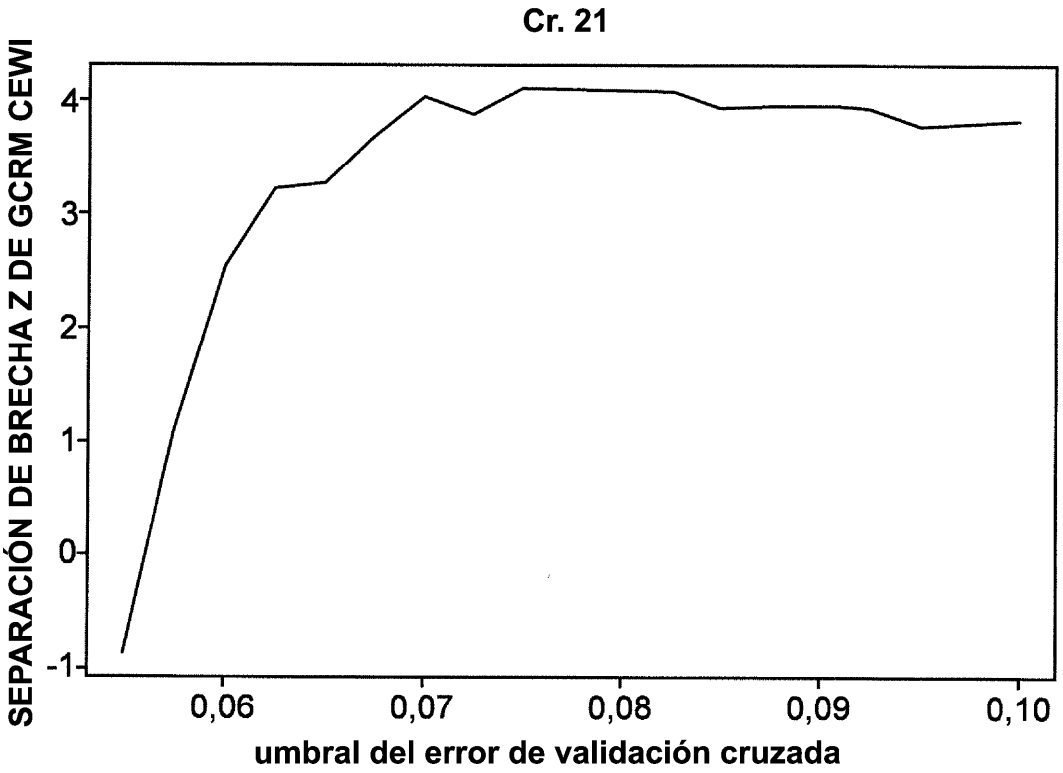

Cr. 21
SEPARACIÓN DE BRECHA Z DE GCRM CEWI
4
3
2
1
0
-1
0,06
0,07
0,08
0,09
0,10
umbral del error de validación cruzada

FIG. 98

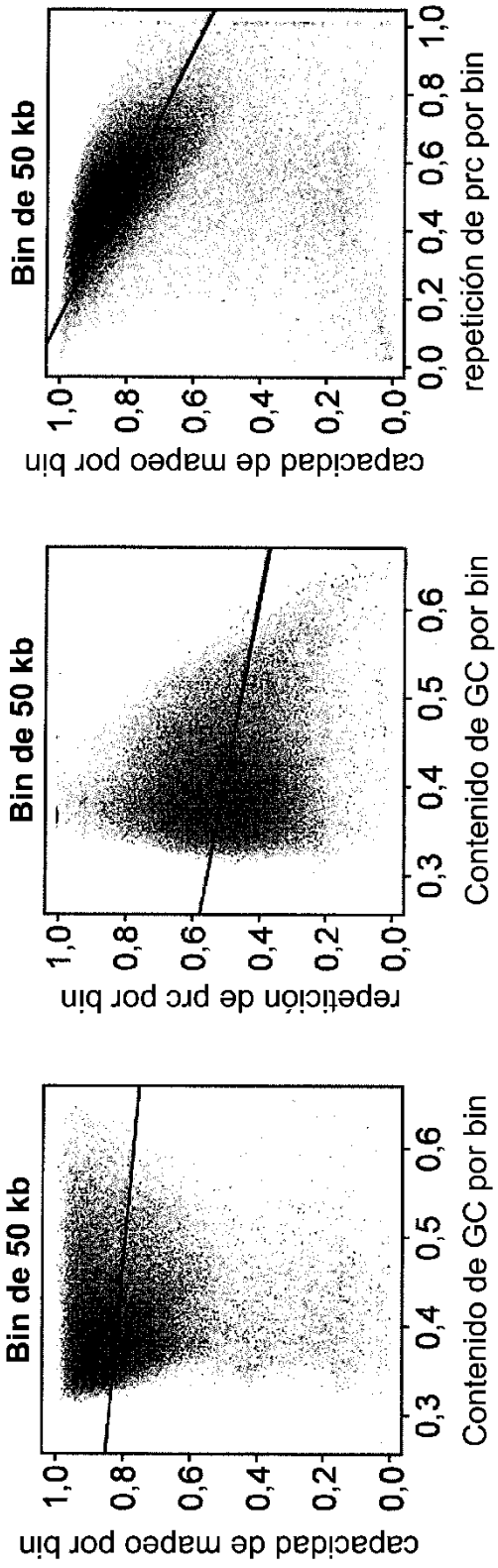


FIG. 99A

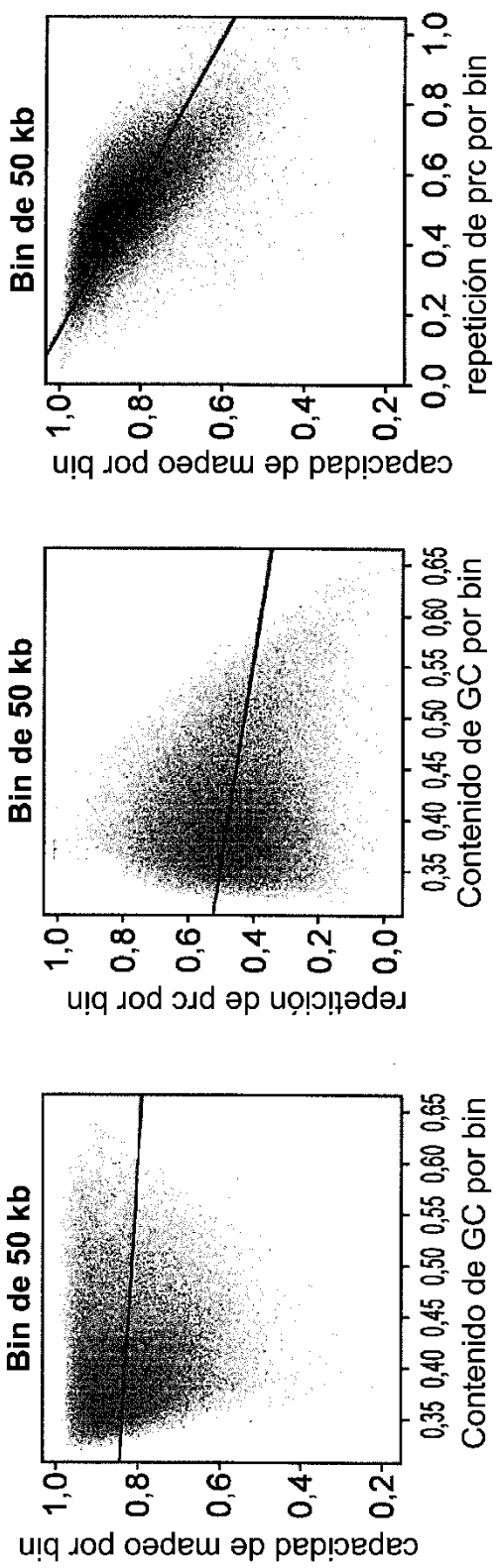


FIG. 99B

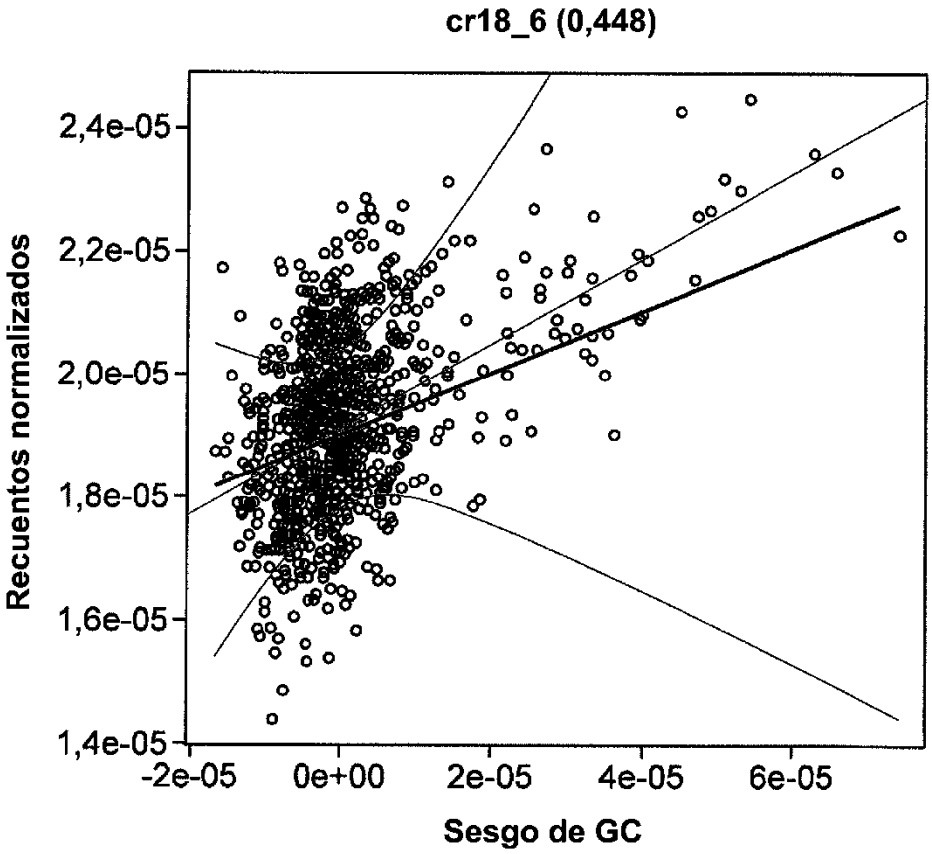

cr18_6 (0,448)
Recuentos normalizados
2,4e-05
2,2e-05
2,0e-05
1,8e-05
1,6e-05
1,4e-05
-2e-05
0e+00
2e-05
4e-05
6e-05
Sesgo de GC

FIG. 100

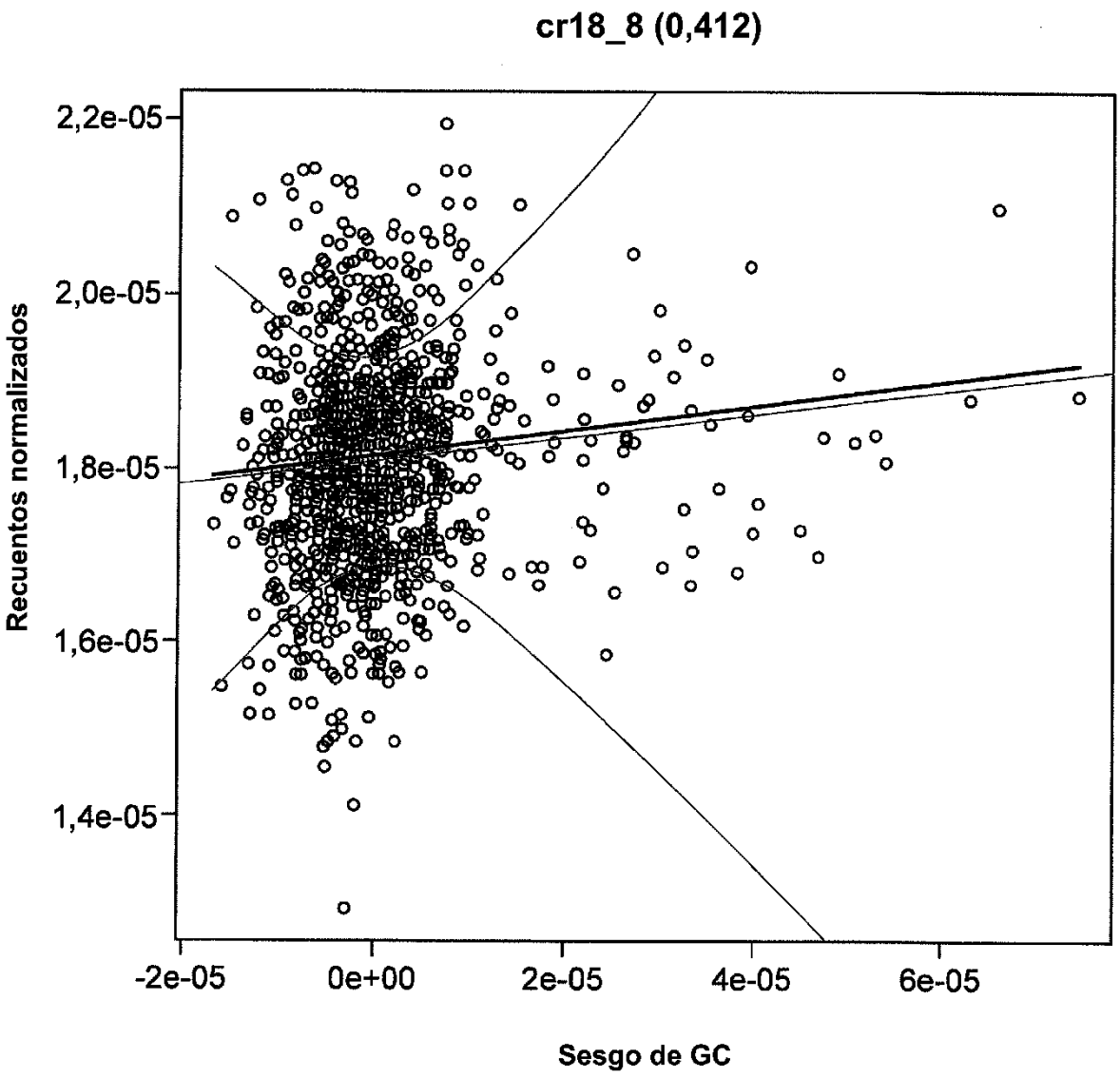

cr18_8 (0,412)
Recuentos normalizados
2,2e-05
2,0e-05
1,8e-05
1,6e-05
1,4e-05
-2e-05
0e+00
2e-05
4e-05
6e-05
Sesgo de GC

FIG. 101

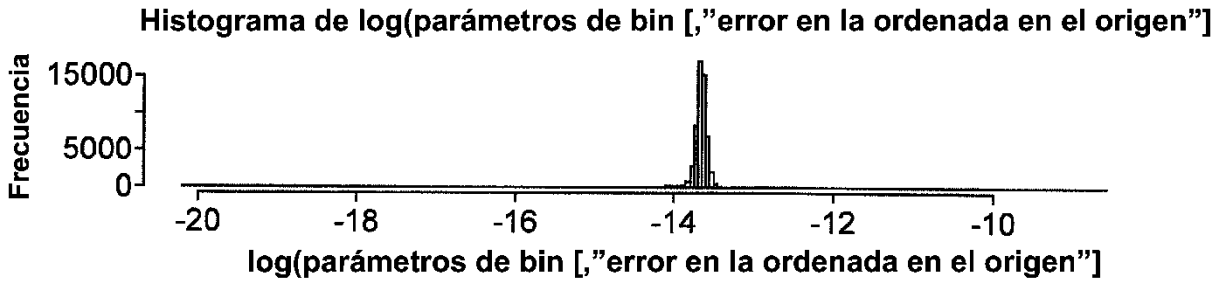

Histograma de log(parámetros de bin [,"error en la ordenada en el origen"]
Frecuencia
15000
5000
0
-20
-18
-16
-14
-12
-10
log(parámetros de bin [,"error en la ordenada en el origen"]

FIG. 102

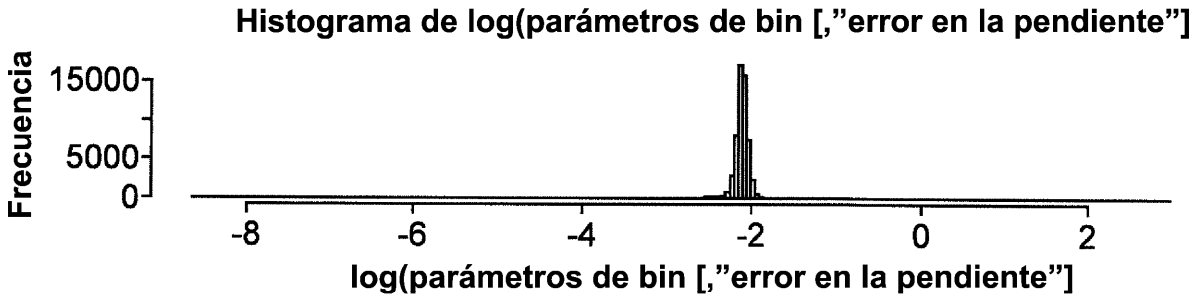

Histograma de log(parámetros de bin [,"error en la pendiente"]
Frecuencia
15000
5000
0
-8
-6
-4
-2
0
2
log(parámetros de bin [,"error en la pendiente"]

FIG. 103

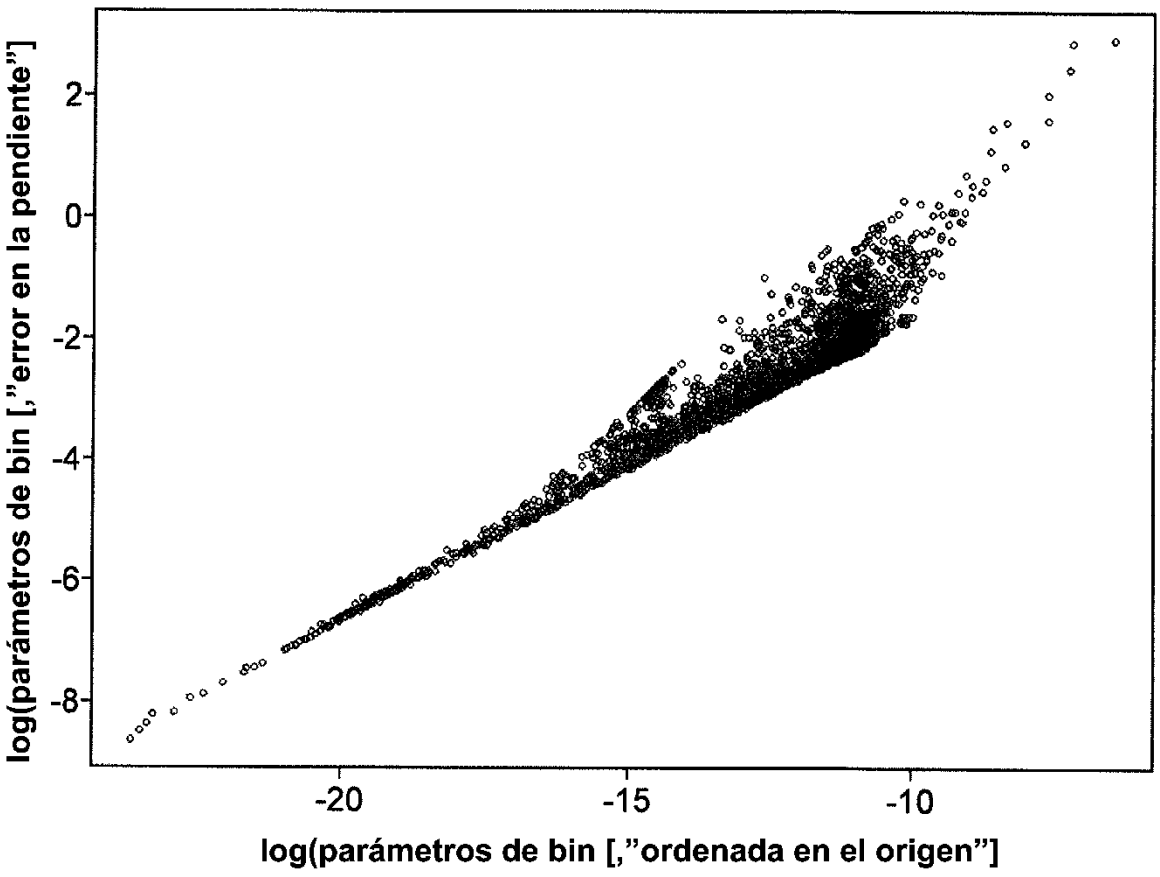

log(parámetros de bin [,"error en la pendiente"]
2
0
-2
-4
-6
-8
-20
-15
-10
log(parámetros de bin [,"ordenada en el origen"]

FIG. 104

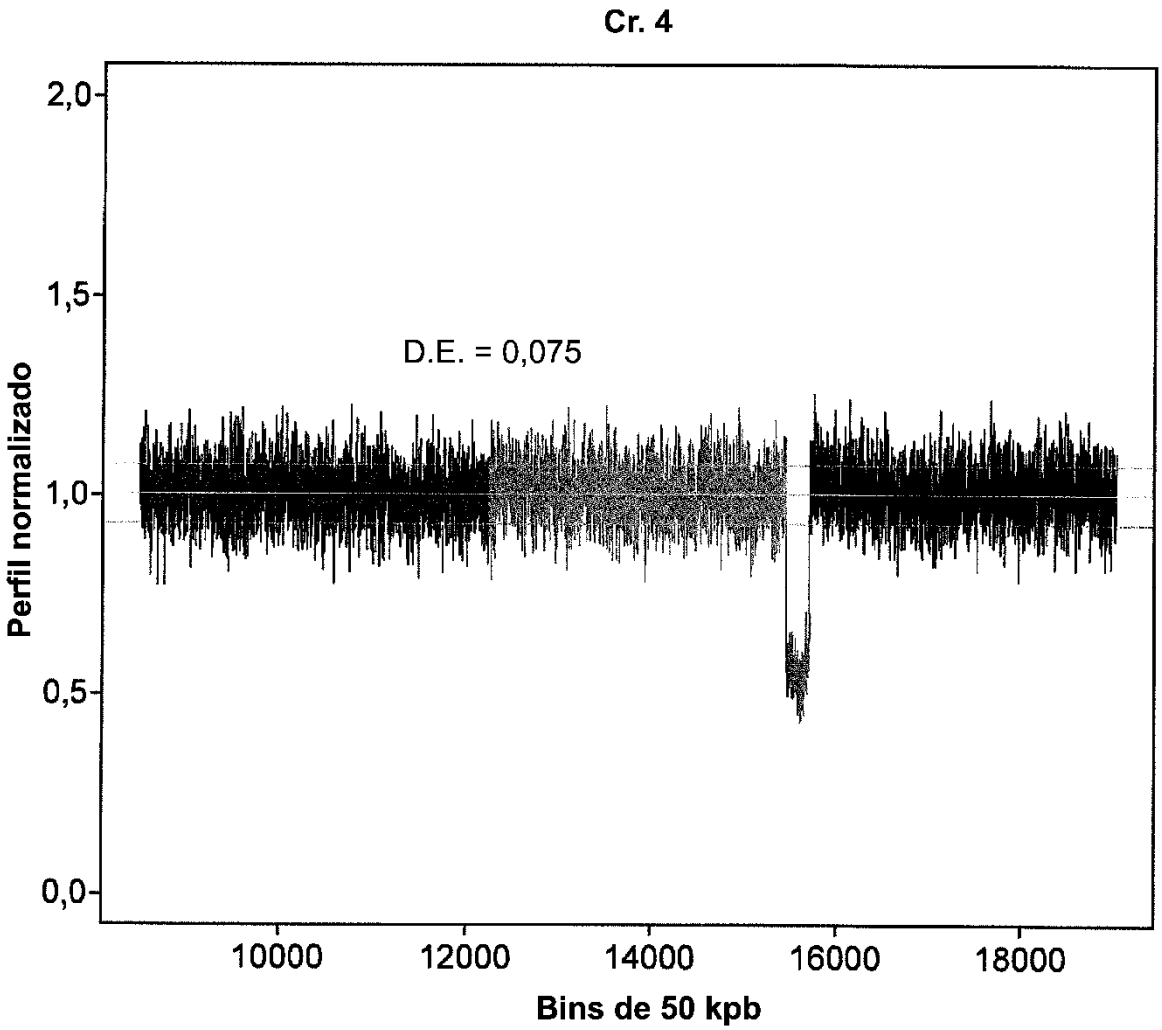

Cr. 4
2,0
1,5
1,0
0,5
0,0
Perfil normalizado
D.E. = 0,075
10000
12000
14000
16000
18000
Bins de 50 kpb

FIG. 105

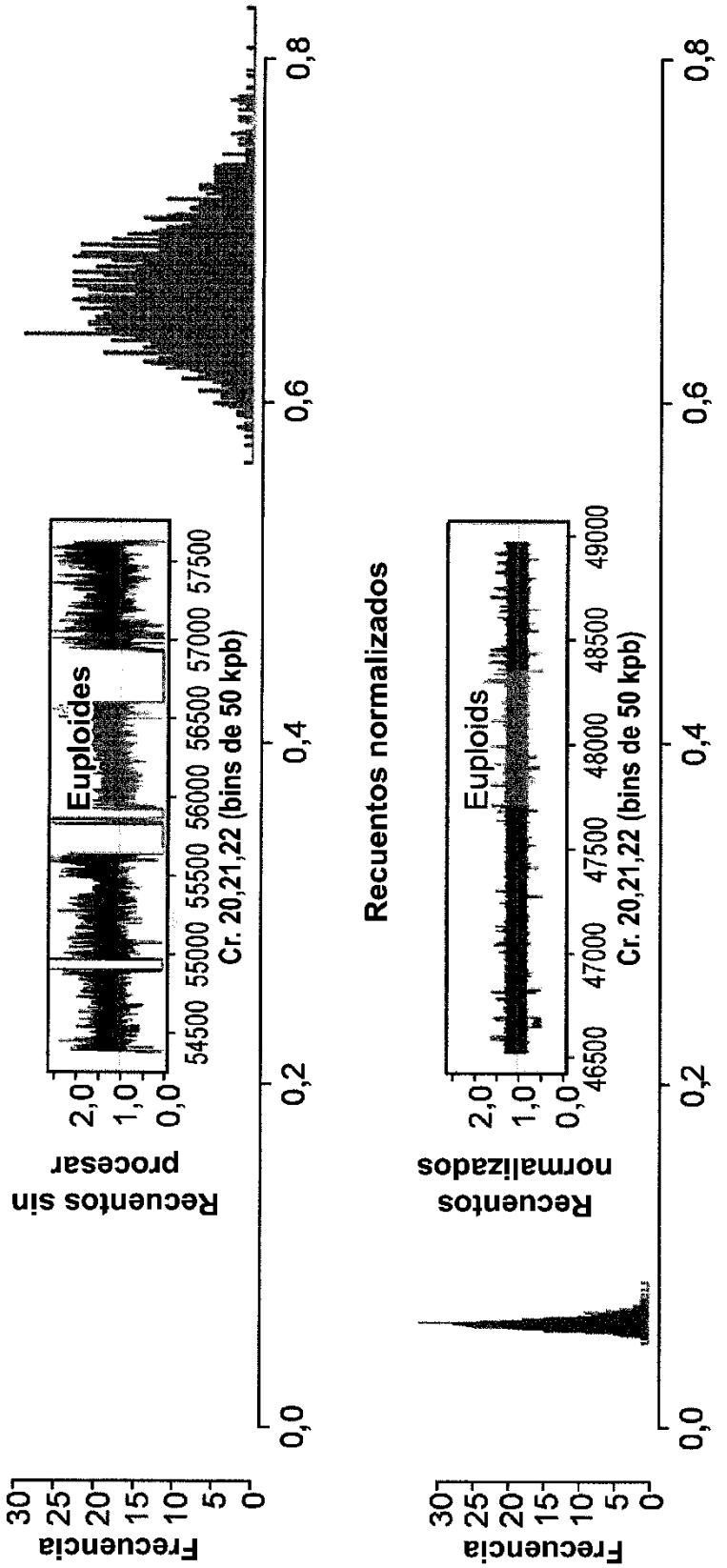

Recuentos sin procesar
Euploides
Cr. 20,21,22 (bins de 50 kpb)
Recuentos sin procesar
Frecuencia
Recuentos normalizados
Euploids
Cr. 20,21,22 (bins de 50 kpb)
Recuentos normalizados
Frecuencia
D.E.

FIG. 106

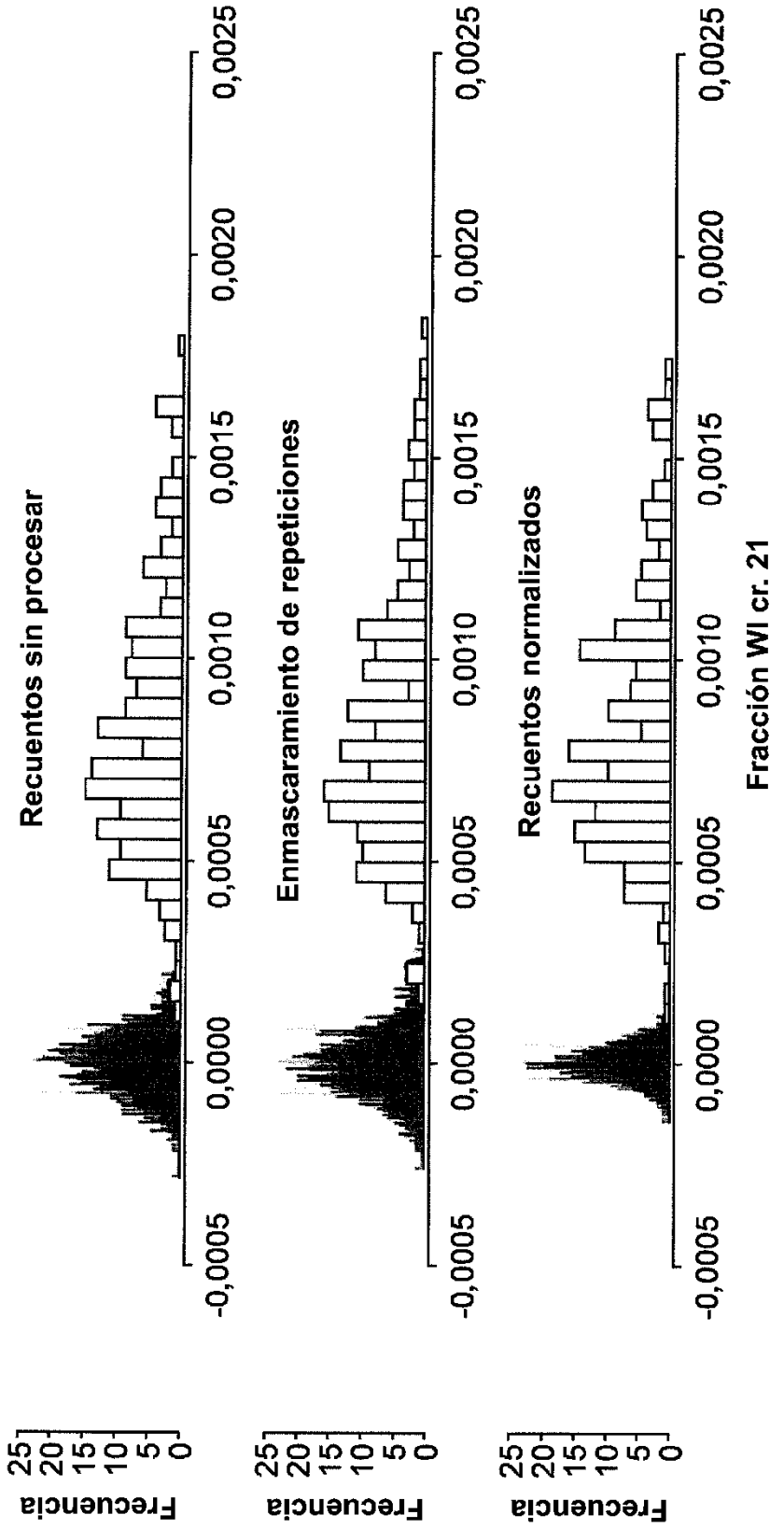

Recuentos sin procesar
Enmascaramiento de repeticiones
Recuentos normalizados
Frecuencia
25
20
15
10
5
0
-0,0005
0,0000
0,0005
0,0010
0,0015
0,0020
0,0025
Fracción WI cr. 21

FIG. 107

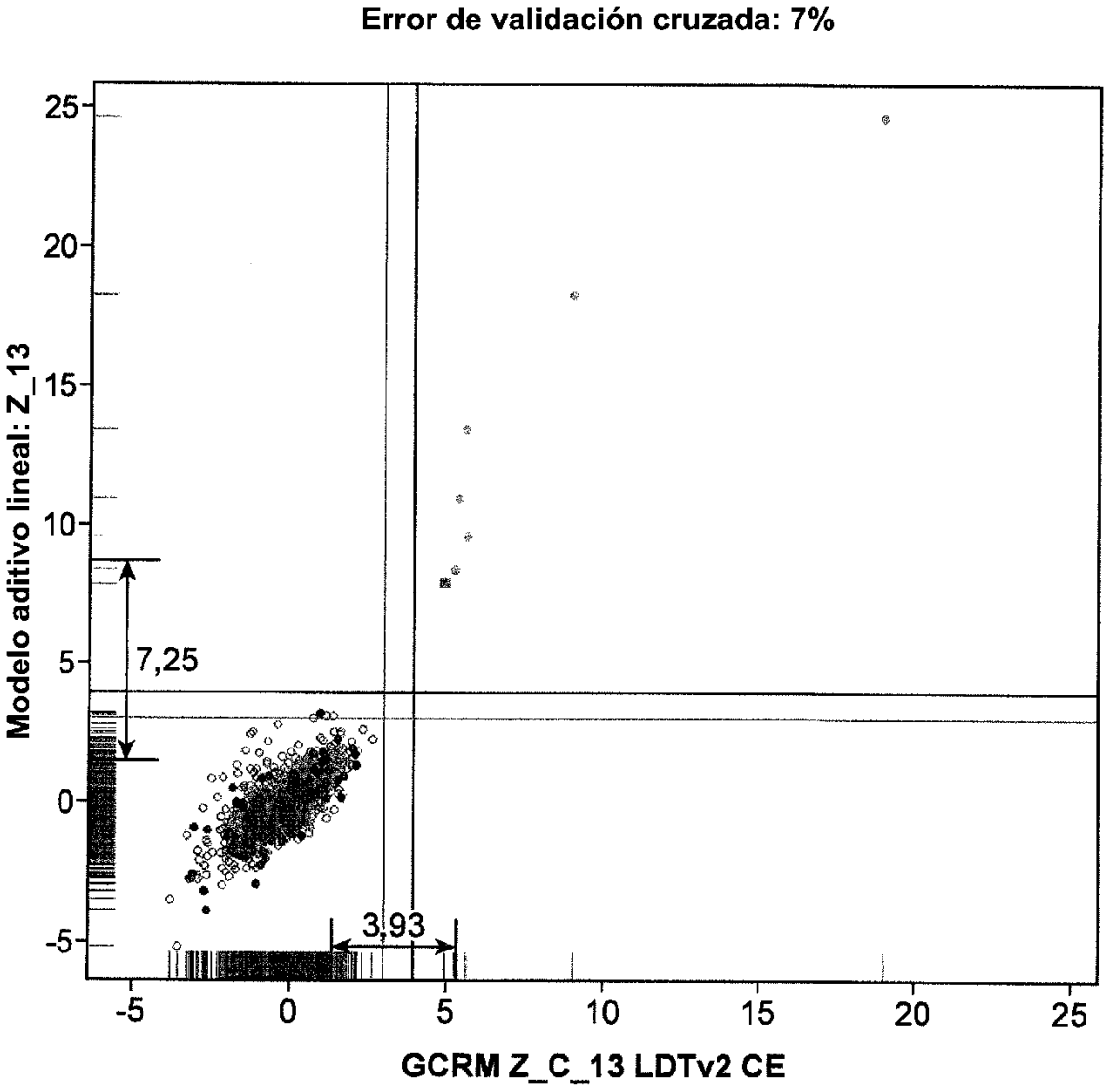

Error de validación cruzada: 7%
Modelo aditivo lineal: Z_13
GCRM Z_C_13 LDTv2 CE
7,25
3,93
25
20
15
10
5
0
-5
-5
0
5
10
15
20
25

FIG. 108

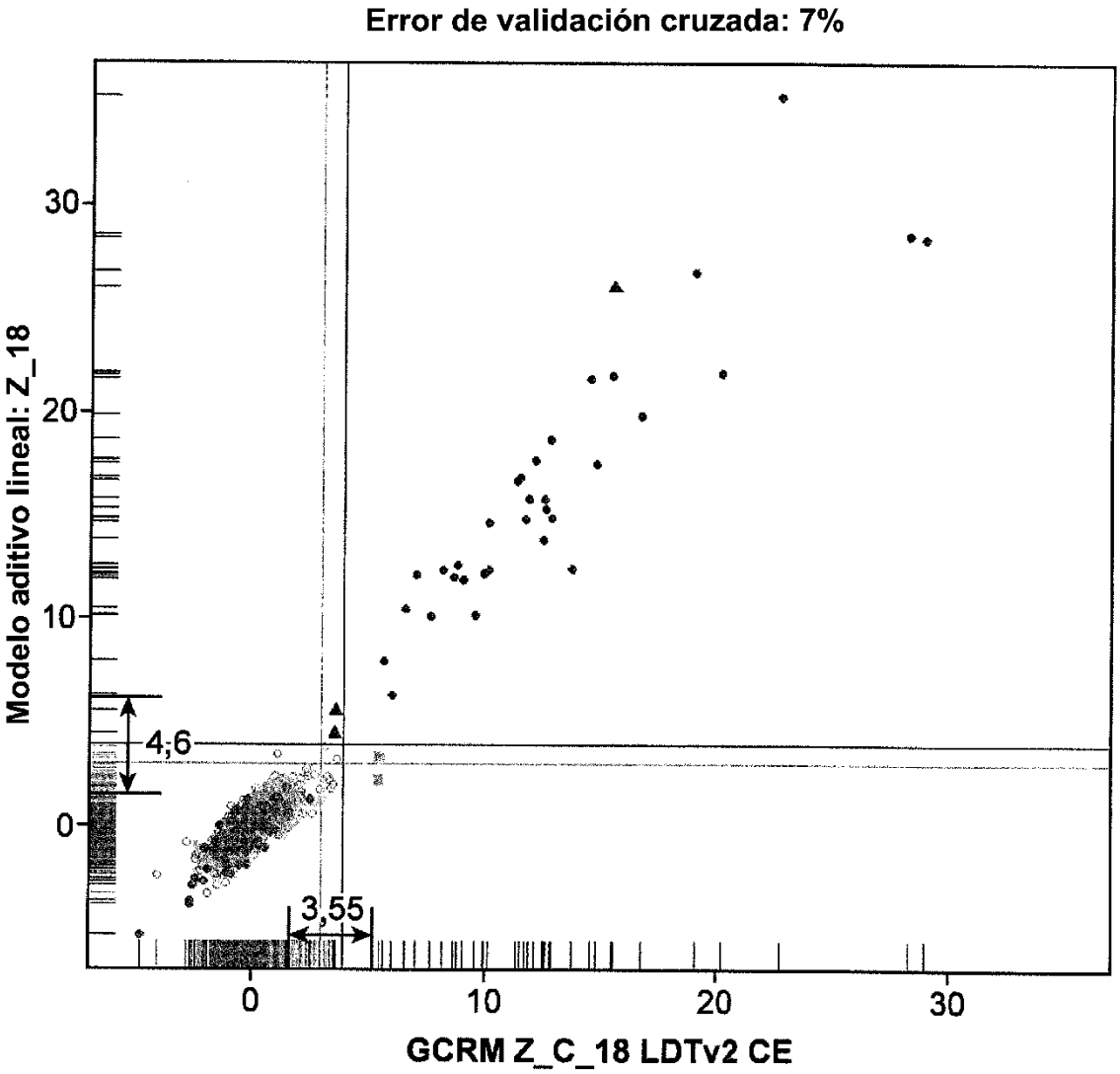

Error de validación cruzada: 7%
Modelo aditivo lineal: Z_18
GCRM Z_C_18 LDTv2 CE
4,6
3,55
0
10
20
30

FIG. 109

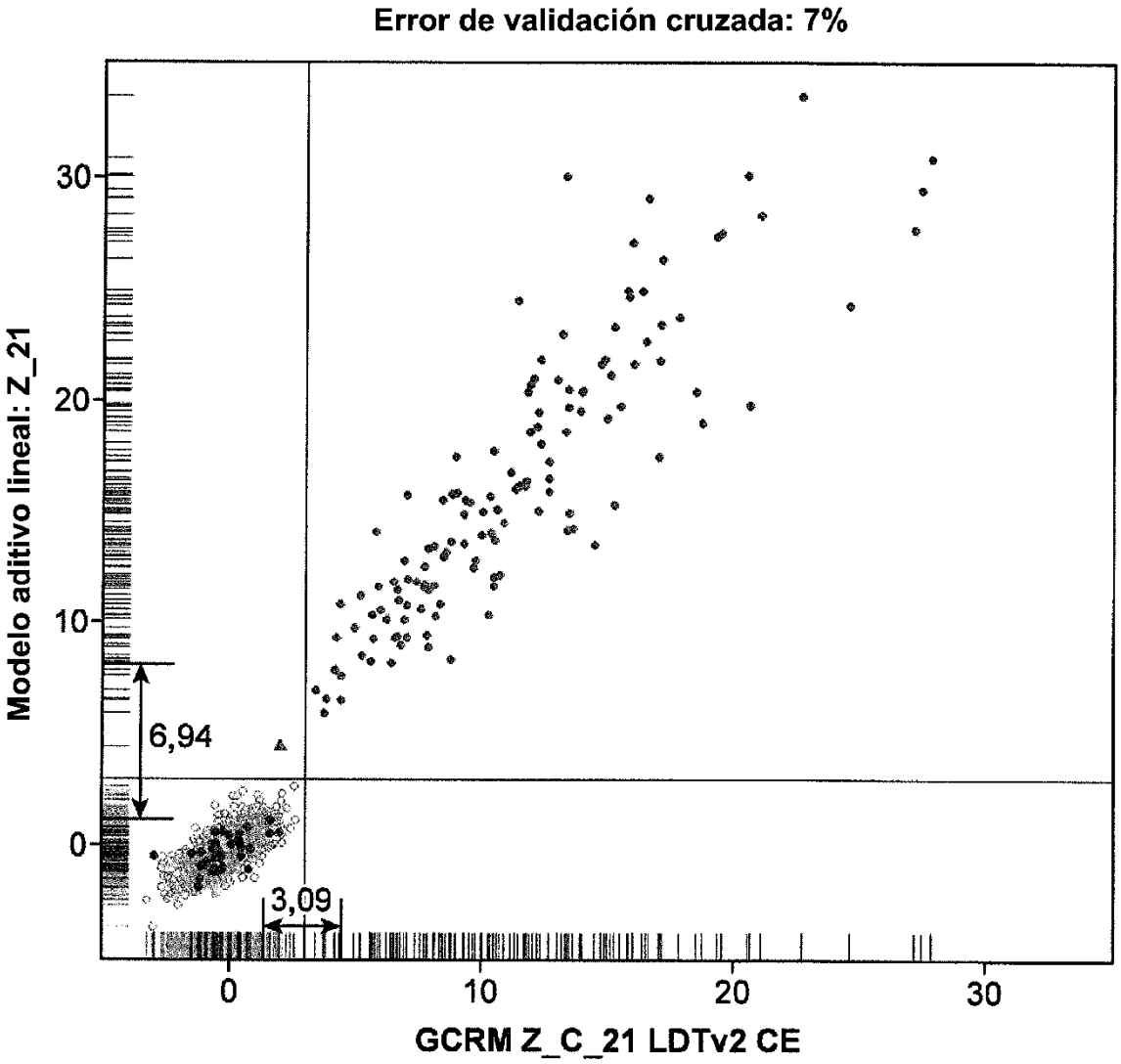

Error de validación cruzada: 7%
Modelo aditivo lineal: Z_21
GCRM Z_C_21 LDTv2 CE
6,94
3,09
0
10
20
30

FIG. 110

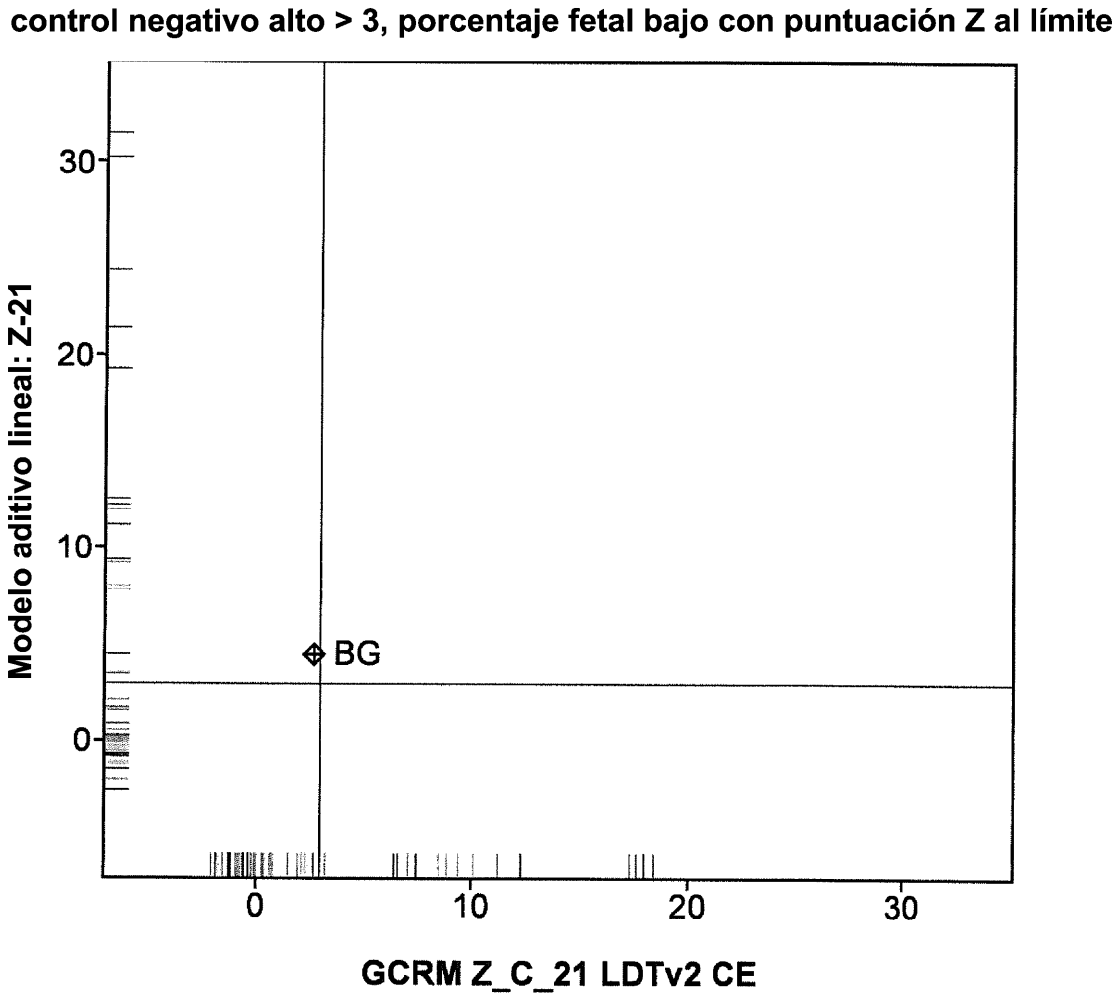

control negativo alto > 3, porcentaje fetal bajo con puntuación Z al límite
Modelo aditivo lineal: Z-21
30
20
10
0
BG
0
10
20
30
GCRM Z_C_21 LDTv2 CE

FIG. 111

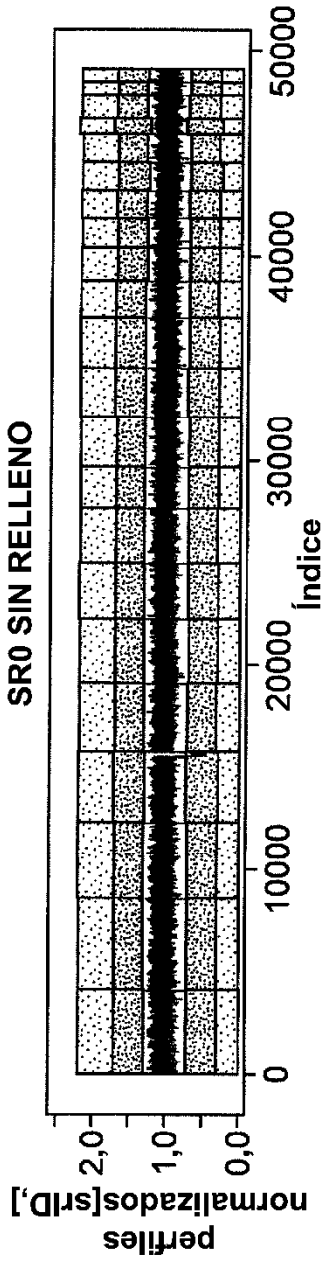


FIG. 112A

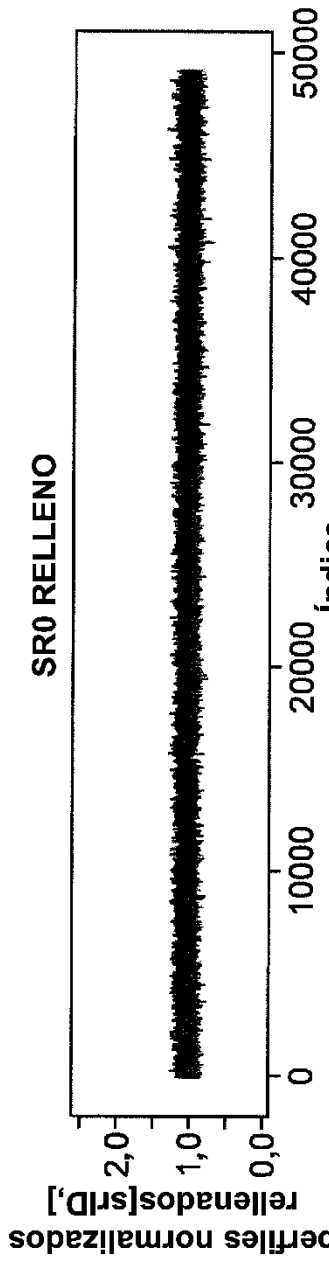


FIG. 112B

FIG. 112C

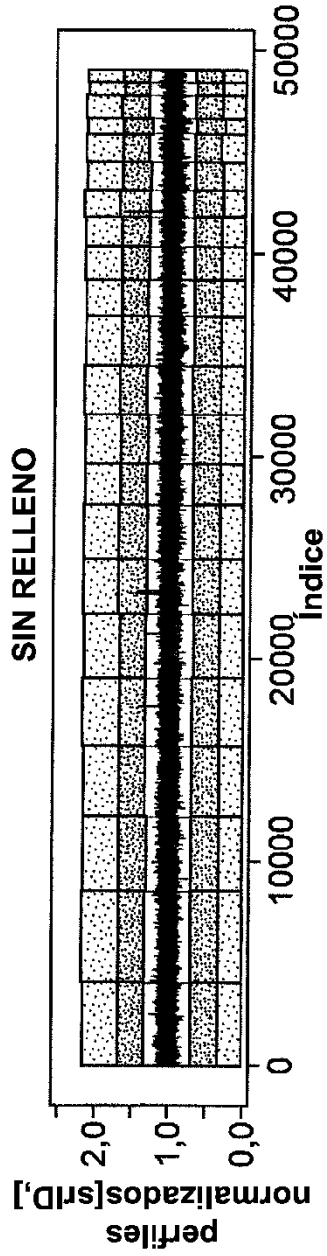


FIG. 113A

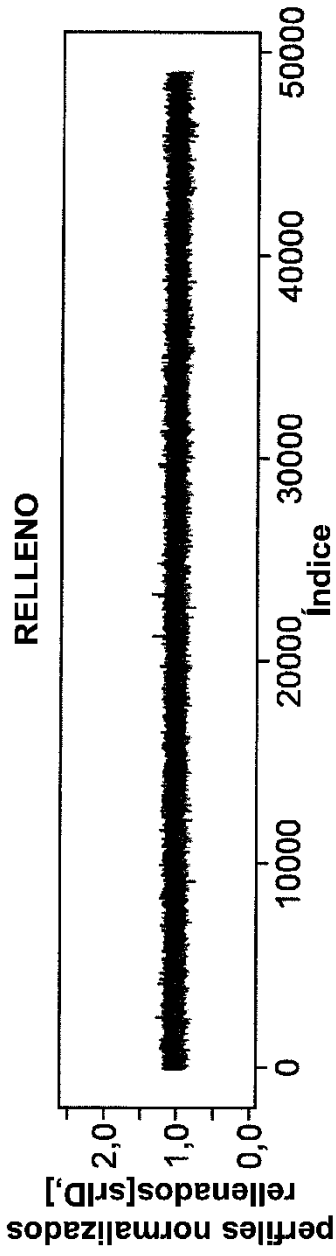


FIG. 113B

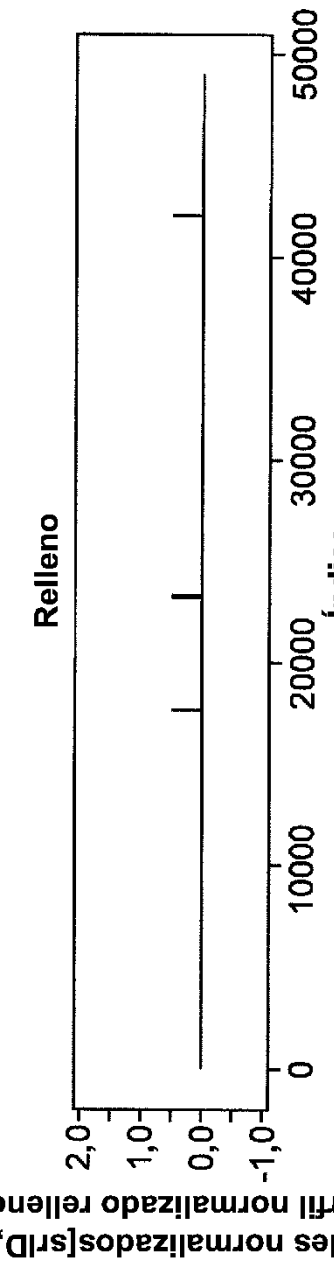


FIG. 113C

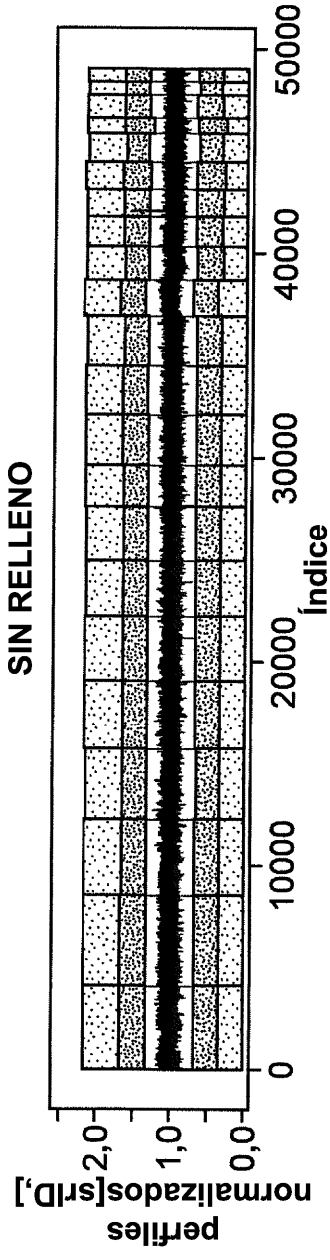


FIG. 114A

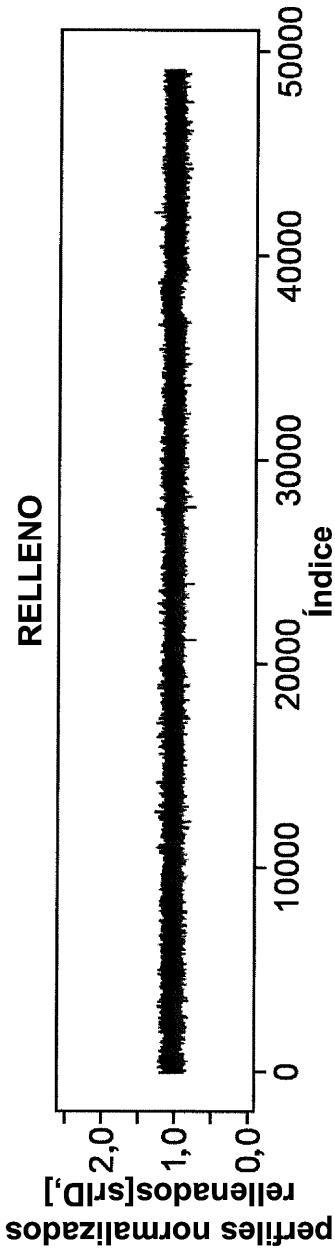


FIG. 114B

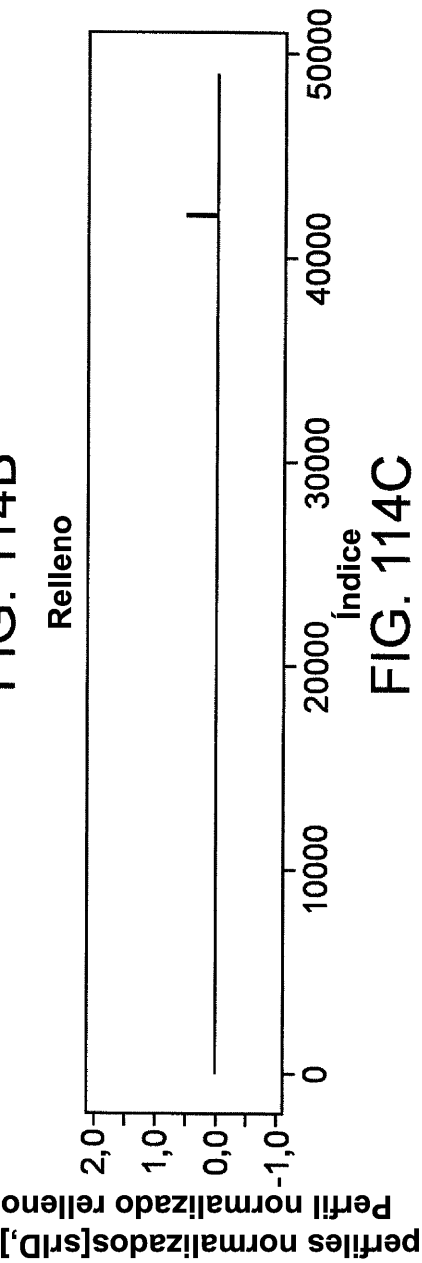


FIG. 114C

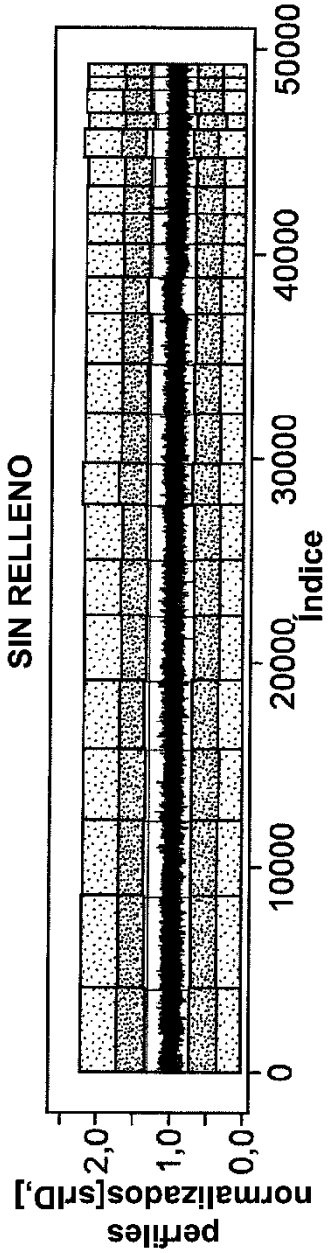


FIG. 115A

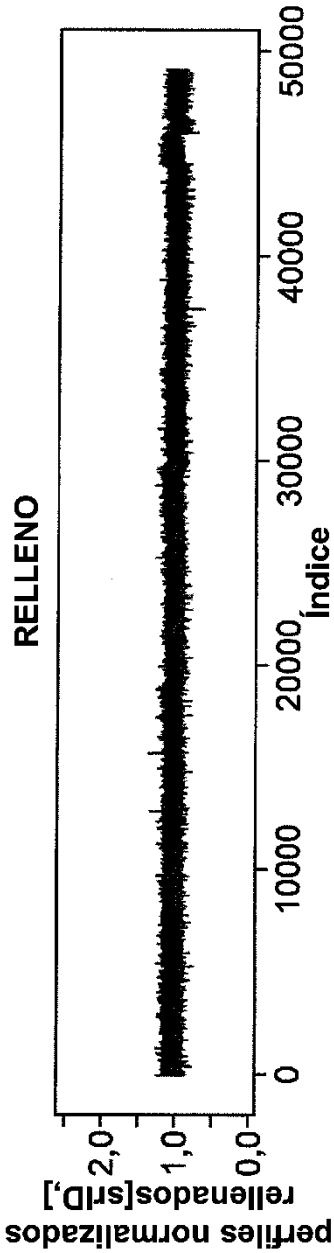


FIG. 115B

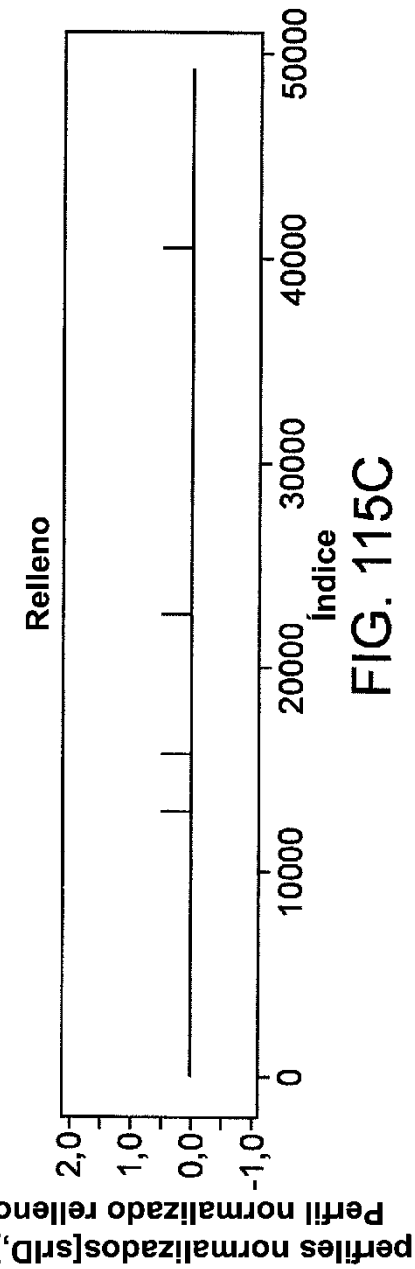


FIG. 115C

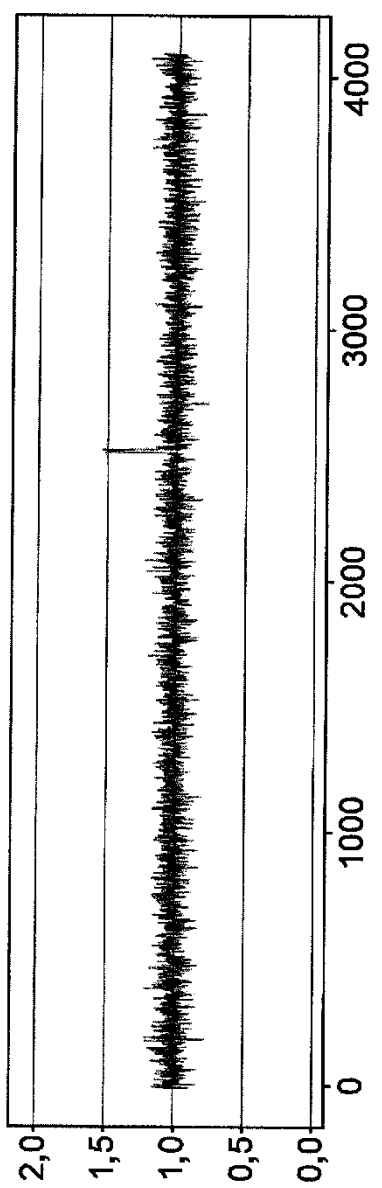


FIG. 116

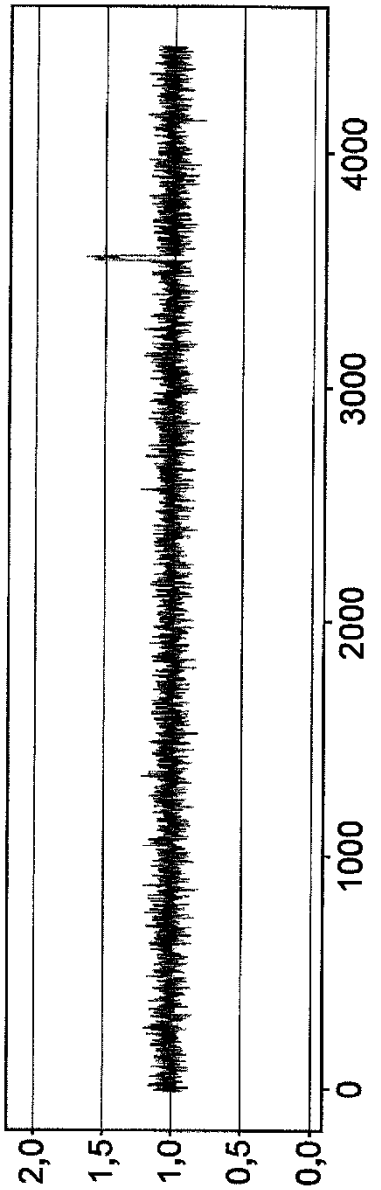


FIG. 117

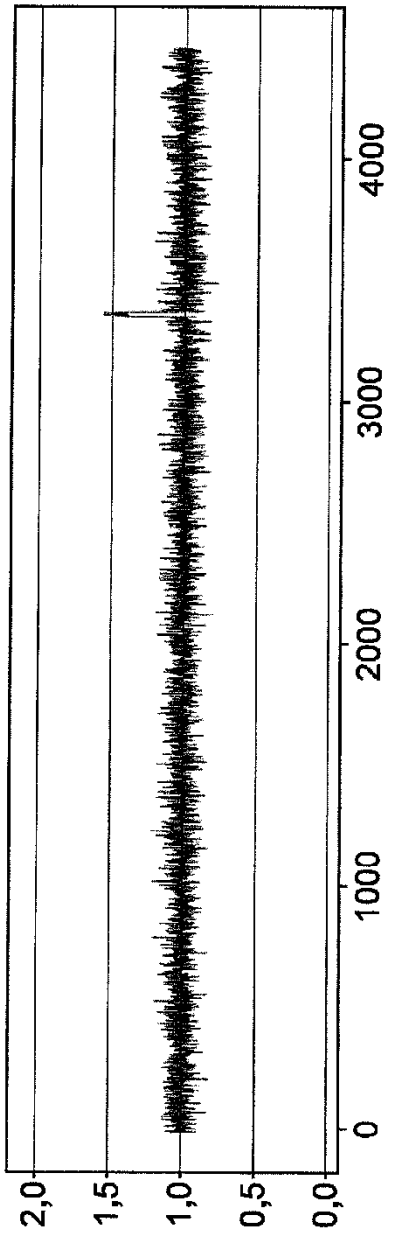


FIG. 118

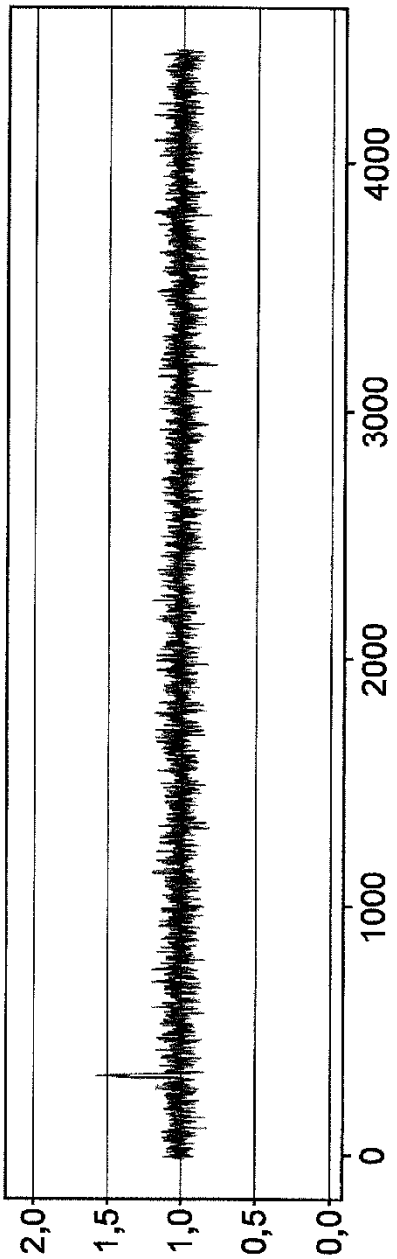


FIG. 119

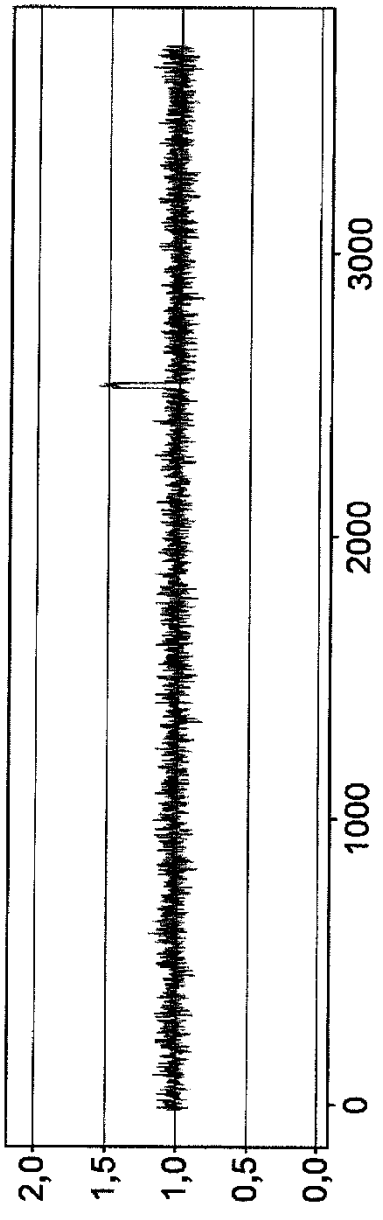


FIG. 120

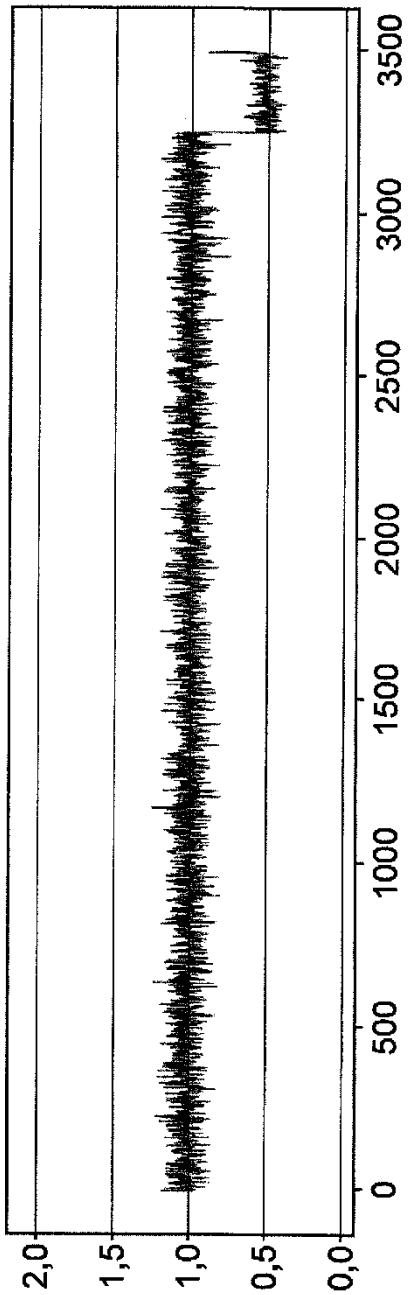


FIG. 121

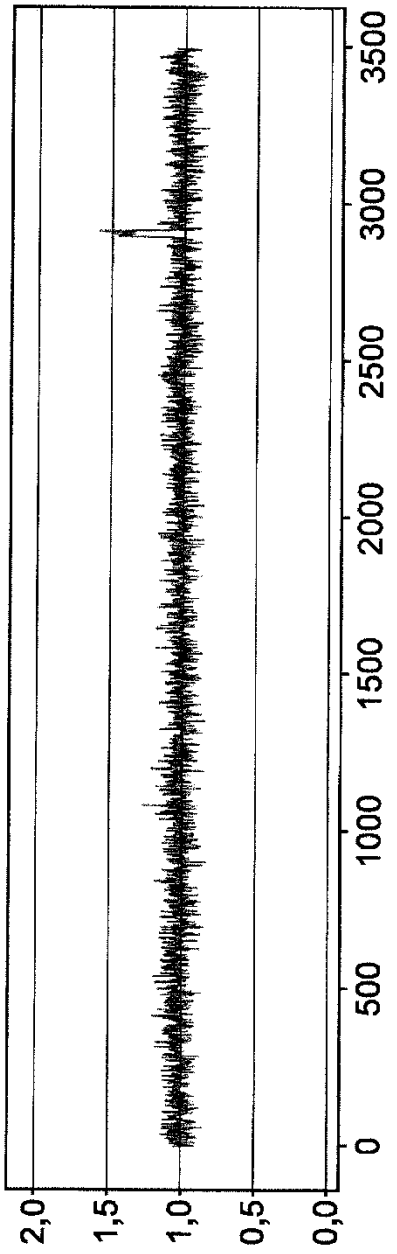


FIG. 122

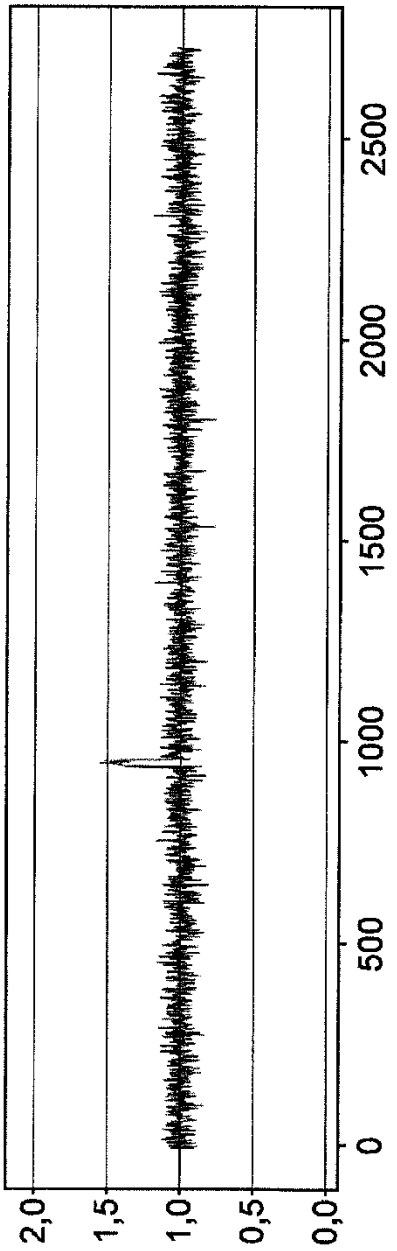


FIG. 123

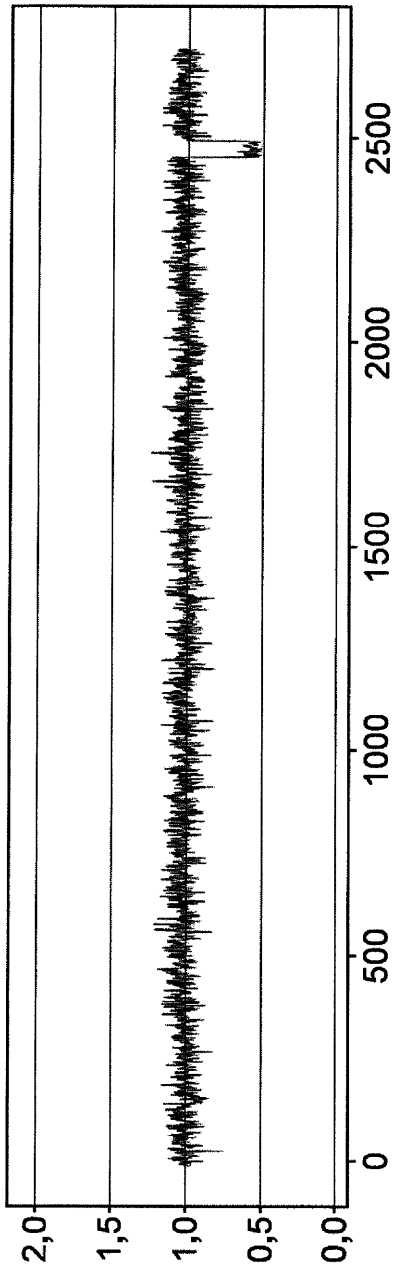


FIG. 124

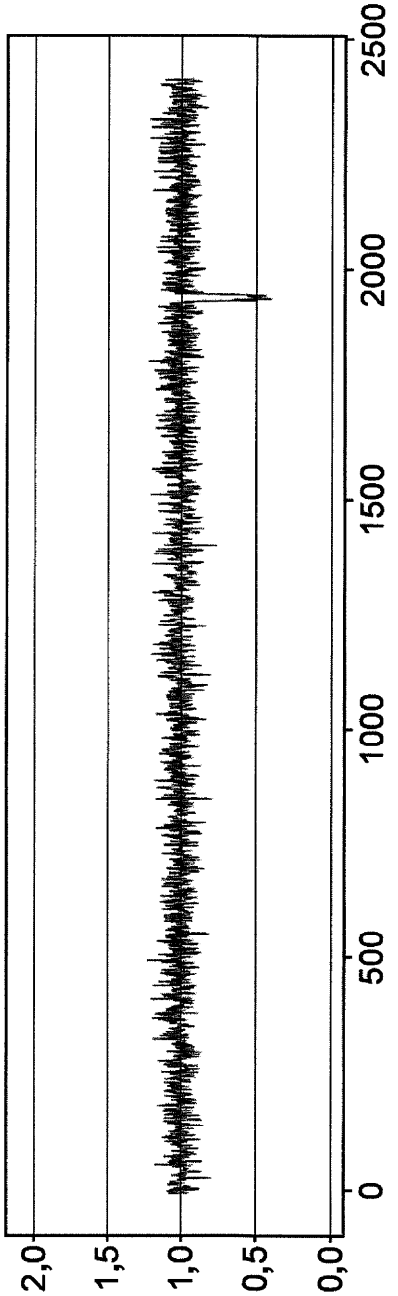


FIG. 125

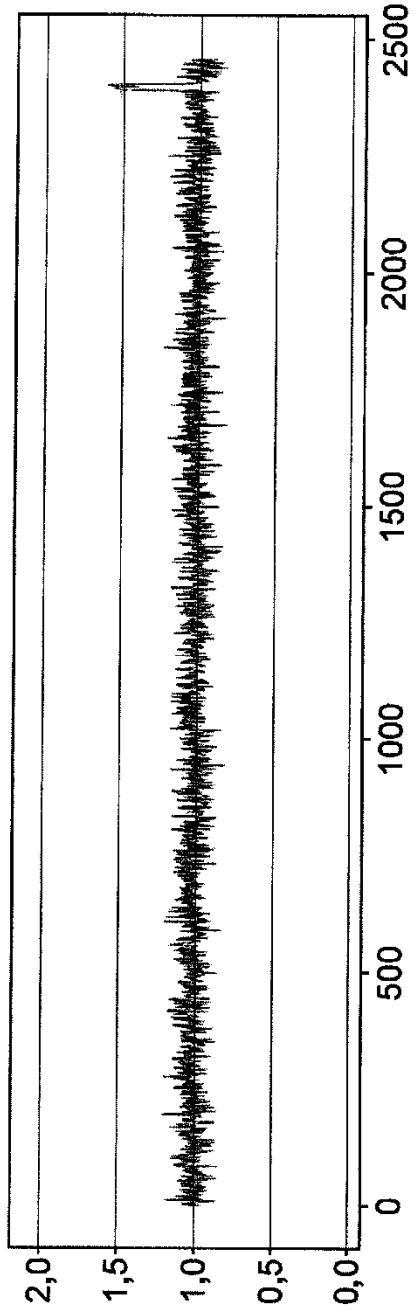


FIG. 126

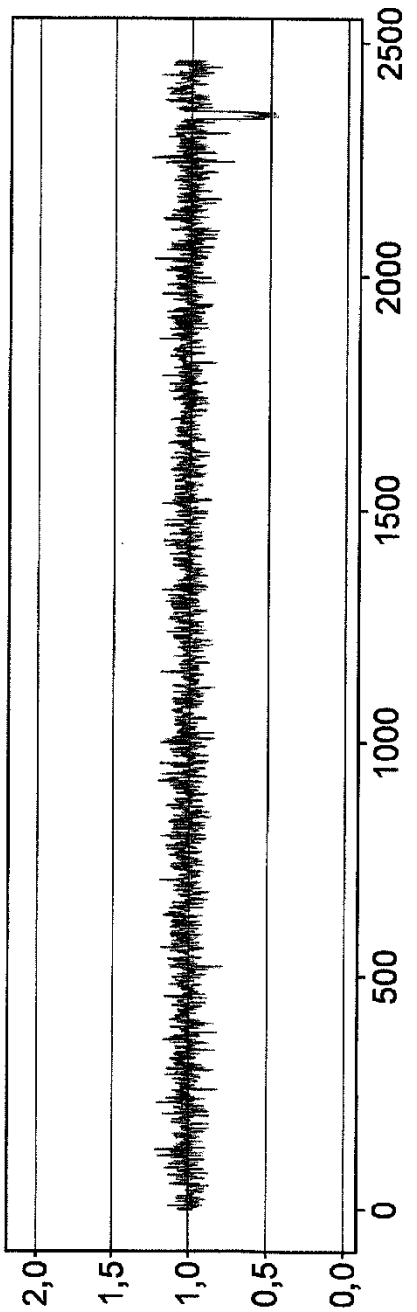


FIG. 127

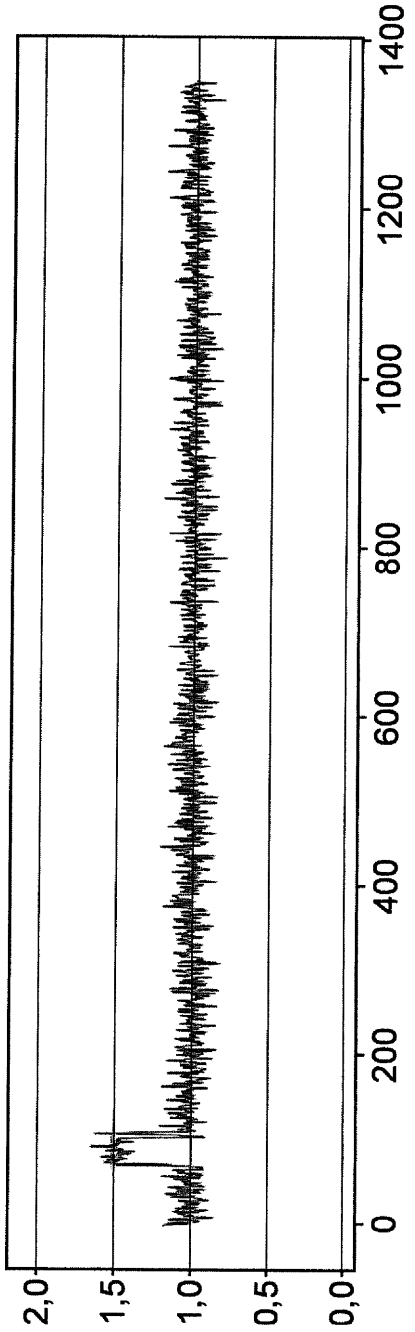


FIG. 128

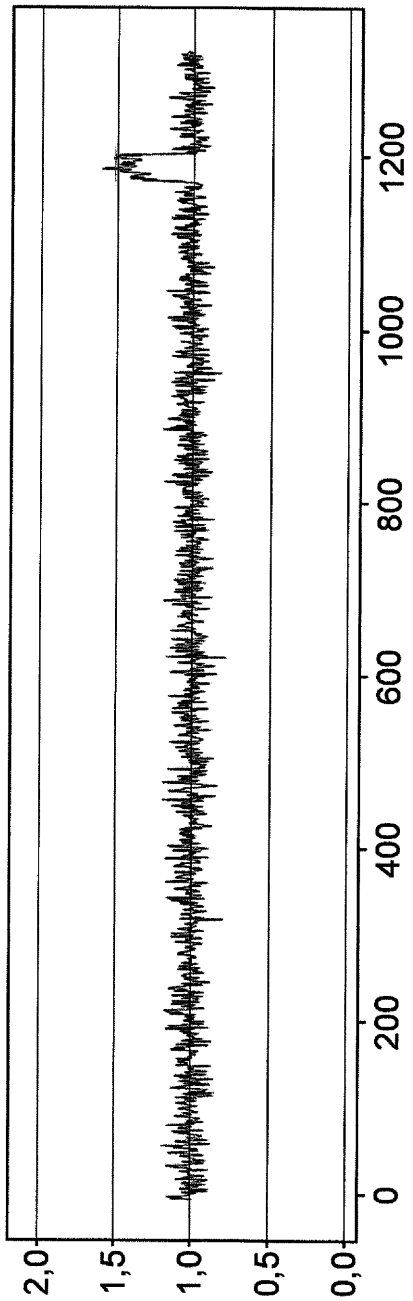


FIG. 129

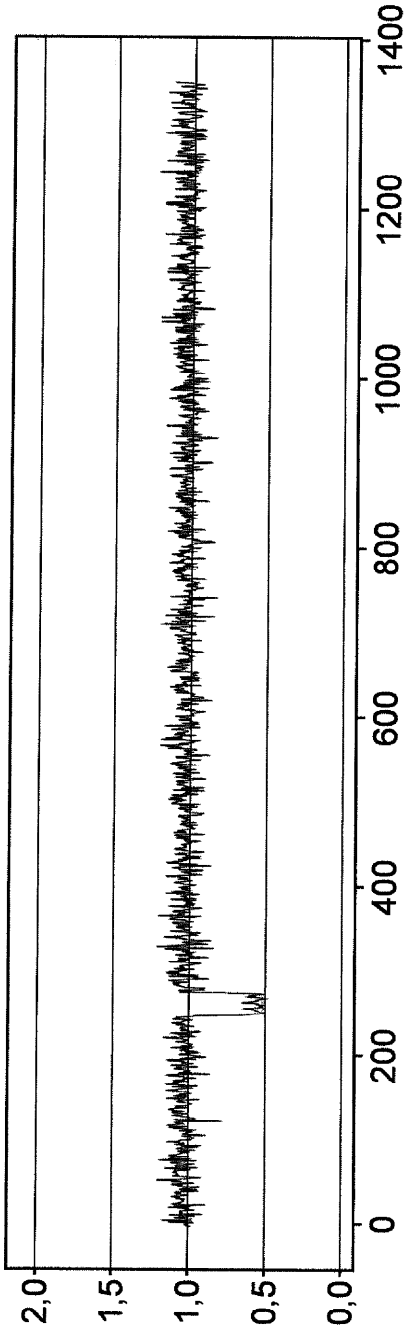


FIG. 130

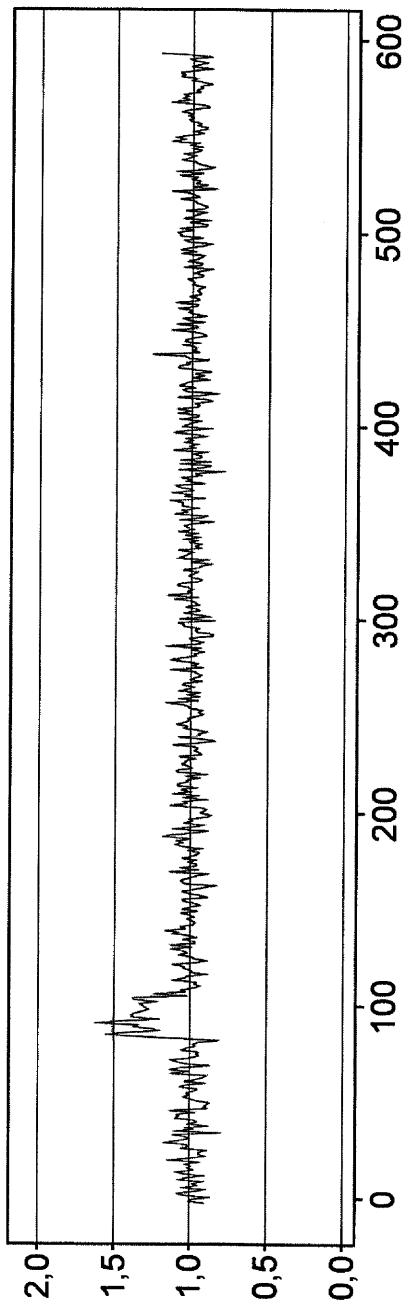


FIG. 131