

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 929 923**

51 Int. Cl.:

G16B 20/20 (2009.01)

G16B 30/10 (2009.01)

G16B 30/20 (2009.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **18.01.2013** **PCT/US2013/022290**

87 Fecha y número de publicación internacional: **25.07.2013** **WO13109981**

96 Fecha de presentación y número de la solicitud europea: **18.01.2013** **E 13707938 (0)**

97 Fecha y número de publicación de la concesión europea: **05.10.2022** **EP 2805280**

54 Título: **Procesos de diagnóstico que condicionan las condiciones experimentales**

30 Prioridad:

20.01.2012 US 201261589202 P

22.06.2012 US 201261663477 P

04.10.2012 US 201261709899 P

05.10.2012 WO PCT/US2012/059123

45 Fecha de publicación y mención en BOPI de la traducción de la patente:
02.12.2022

73 Titular/es:

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

72 Inventor/es:

DECIU, COSMIN;
EHRICH, MATHIAS;
VAN DEN BOOM, DIRK JOHANNES y
DZAKULA, ZELJKO

74 Agente/Representante:

DEL VALLE VALIENTE, Sonia

ES 2 929 923 T3

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

DESCRIPCIÓN

Procesos de diagnóstico que condicionan las condiciones experimentales

5 **Campo**

La presente invención se define mediante las reivindicaciones. En el presente documento se describen generalmente métodos, procesos y aparatos para la evaluación no invasiva de variaciones genéticas.

10 **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 delección 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. 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. Los enfoques anteriores para determinar la aneuploidía cromosómica a partir de una muestra materna que comprende ADN fetal y materno se describen en la solicitud de patente de EE. UU. n.º: 2011/201507 A1 (Rava RP *et al.*, publicada el 18 de agosto de 2011); y la solicitud de patente de RU GB 2 484 764 A (Rava RP, publicada el 25 de abril de 2012).

40 **Sumario**

La presente invención se define mediante las reivindicaciones. La invención se refiere a un método implementado por ordenador para detectar la presencia o ausencia de una aneuploidía fetal, que comprende: (a) obtener recuentos de lecturas de secuencia de nucleótidos mapeadas en secciones genómicas de referencia, en donde las lecturas de secuencia de nucleótidos se obtienen de un grupo de muestras de prueba, comprendiendo cada una ácido nucleico libre de células circulante de una mujer embarazada, en donde las muestras de prueba se secuencian bajo una o más condiciones experimentales comunes; (b) para cada muestra de prueba, filtrar secciones genómicas basándose en uno o más datos redundantes, datos no informativos, datos con ruido, secciones genómicas con secuencias sobrerrepresentadas y secciones genómicas con secuencias subrepresentadas; y/o calcular una medida de error para los recuentos de lecturas de secuencia mapeadas en algunas o todas las secciones genómicas y eliminar los recuentos de lecturas de secuencia para determinadas secciones genómicas según un umbral de la medida de error, proporcionando así secciones genómicas filtradas; (c) para cada muestra de prueba, normalizar los recuentos de las secciones genómicas filtradas, o normalizar una derivada de los recuentos de las secciones genómicas filtradas, ajustando los recuentos para la variabilidad inducida por las condiciones experimentales, en donde el ajuste comprende restar los recuentos esperados de los recuentos para las secciones genómicas, generando así un valor de resta, y dividir el valor de la resta por una estimación de la variabilidad, cuyos recuentos esperados y estimación de la variabilidad se basan en la variabilidad inducida por las condiciones experimentales de los recuentos de las secciones genómicas filtradas, obteniendo así recuentos normalizados para secciones genómicas filtradas; y (d) detectar la presencia o ausencia de una aneuploidía fetal para cada muestra de prueba basándose en los recuentos normalizados para las secciones genómicas filtradas.

En una realización, cada muestra de prueba es de plasma sanguíneo de la mujer embarazada. En una realización adicional, la aneuploidía fetal es trisomía 13, trisomía 18 o trisomía 21. En una realización adicional, la medida de error en (b) es un factor R.

En una realización adicional, el recuento esperado es una mediana de recuento y la estimación de la variabilidad es una median absolute deviation (mediana de desviación absoluta -MAD) del recuento esperado. En una realización

alternativa, el recuento esperado es una media recortada o truncada, una media Winsorizada o una estimación de remuestreo de tipo *bootstrapping*.

En una realización adicional, los recuentos se normalizan adicionalmente mediante el contenido de GC, la normalización basada en bins, GC LOESS, PERUN, GCRM o combinaciones de los mismos. En una realización adicional, la una o más condiciones experimentales comunes se eligen de una unidad de celda de flujo común, celdas de flujo comunes a un recipiente, celdas de flujo comunes a un lote o proceso de fabricación, una unidad de placa de reactivo común, placas de reactivo comunes a un recipiente y placas de reactivo comunes a un lote o proceso de fabricación. En una realización adicional, la normalización de los recuentos comprende determinar una representación porcentual. En una realización adicional, el recuento normalizado es una puntuación z. En una realización adicional, el recuento normalizado es una puntuación z robusta. En una realización adicional, la derivada de los recuentos para la sección genómica es una representación porcentual de la sección genómica. En una realización adicional, la mediana es una mediana de una representación porcentual. En una realización adicional, la representación porcentual es una representación cromosómica. En una realización adicional, el método comprende secuenciar el ácido nucleico mediante un módulo de secuenciación, proporcionando así las lecturas de secuencia de ácido nucleico y mapeando las lecturas de secuencia de ácido nucleico a las secciones genómicas de un genoma de referencia.

En el presente documento también se desvela un método para detectar la presencia o ausencia de una aneuploidía fetal, que incluye: (a) obtener lecturas de secuencia de nucleótidos a partir de ácido nucleico de muestra, incluyendo ácido nucleico libre de células circulante de una mujer embarazada; (b) mapear las lecturas de secuencia de nucleótidos en secciones genómicas de referencia; (c) contar el número de lecturas de secuencia de nucleótidos mapeadas en cada sección genómica de referencia, obteniendo así recuentos; (d) normalizar los recuentos para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que incluye muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; y (e) proporcionar un resultado determinante de la presencia o ausencia de una aneuploidía fetal a partir del recuento de muestra normalizado. En algunas realizaciones, las lecturas de secuencia se asignan a una porción de, o todas, las secciones genómicas de referencia.

En el presente documento también se proporciona un método para detectar la presencia o ausencia de una aneuploidía fetal, que incluye: (a) obtener una muestra que incluya ácido nucleico libre de células circulante de una mujer embarazada; (b) aislar el ácido nucleico de muestra de la muestra; (c) obtener lecturas de secuencia de nucleótidos a partir de un ácido nucleico de muestra; (d) mapear las lecturas de secuencia de nucleótidos en secciones genómicas de referencia, (e) contar el número de lecturas de secuencia de nucleótidos mapeadas en cada sección genómica de referencia, obteniendo así recuentos; (f) normalizar los recuentos para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que incluye muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; y (g) proporcionar un resultado determinante de la presencia o ausencia de una aneuploidía fetal a partir del recuento de muestra normalizado.

En el presente documento también se proporciona un método para detectar la presencia o ausencia de una aneuploidía fetal, que incluye: (a) mapear en secciones genómicas de referencia lecturas de secuencia de nucleótidos obtenidas de ácido nucleico de muestra, incluyendo ácido nucleico libre de células circulante de una mujer embarazada; (b) contar el número de lecturas de secuencia de nucleótidos mapeadas en cada sección genómica de referencia, obteniendo así recuentos; (c) normalizar los recuentos para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que incluye muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; y (d) proporcionar un resultado determinante de la presencia o ausencia de una aneuploidía fetal a partir del recuento de muestra normalizado.

En el presente documento también se proporciona un método para detectar la presencia o ausencia de una variación genética, que incluye: (a) obtener lecturas de secuencia de nucleótidos a partir de ácido nucleico de muestra, incluyendo ácido nucleico libre de células circulante de un sujeto de prueba; (b) mapear las lecturas de secuencia de nucleótidos en secciones genómicas de referencia; (c) contar el número de lecturas de secuencia de nucleótidos mapeadas en cada sección genómica de referencia, obteniendo así recuentos; (d) normalizar los recuentos para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que incluye muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; y (e) proporcionar un resultado determinante de la presencia o ausencia de una variación genética en el sujeto de prueba a partir del recuento de muestra normalizado.

En el presente documento también se proporciona un método para detectar la presencia o ausencia de una aneuploidía fetal, que incluye: (a) obtener una muestra que incluya ácido nucleico libre de células circulante de un sujeto de prueba; (b) aislar el ácido nucleico de muestra de la muestra; (c) obtener lecturas de secuencia de nucleótidos a partir de un ácido nucleico de muestra; (d) mapear las lecturas de secuencia de nucleótidos en secciones genómicas de referencia, (e) contar el número de lecturas de secuencia de nucleótidos mapeadas en cada sección genómica de referencia, obteniendo así recuentos; (f) normalizar los recuentos para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que incluye muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; y (g) proporcionar un resultado determinante de la presencia o ausencia de una variación genética en el sujeto de prueba a partir del recuento de muestra normalizado.

En el presente documento también se proporciona un método para detectar la presencia o ausencia de una variación genética, que incluye: (a) mapear en secciones genómicas de referencia lecturas de secuencia de nucleótidos obtenidas de ácido nucleico de muestra, incluyendo ácido nucleico libre de células circulante de un sujeto de prueba; (b) contar el número de lecturas de secuencia de nucleótidos mapeadas en cada sección genómica de referencia, obteniendo así recuentos; (c) normalizar los recuentos para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que incluye muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; y (d) proporcionar un resultado determinante de la presencia o ausencia de una variación genética en el sujeto de prueba a partir del recuento de muestra normalizado.

En el presente documento también se proporciona un método para detectar la presencia o ausencia de una variación genética, que incluye: (a) obtener lecturas de secuencia de nucleótidos a partir de ácido nucleico de muestra, incluyendo ácido nucleico libre de células circulante de un sujeto de prueba; (b) mapear las lecturas de secuencia de nucleótidos en secciones genómicas de referencia; (c) contar el número de lecturas de secuencia de nucleótidos mapeadas en cada sección genómica de referencia, obteniendo así recuentos; (d) ajustar las lecturas de secuencia mapeadas contadas en (c) según una variable o característica seleccionada, cuya variable o característica seleccionada minimiza o elimina el efecto de secuencias repetitivas y/o secuencias sobrerrepresentadas o subrepresentadas; (e) normalizar los recuentos restantes después de (d) para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que incluye muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; y (f) proporcionar un resultado determinante de la presencia o ausencia de una variación genética en el sujeto de prueba a partir del recuento de muestra normalizado.

En el presente documento también se proporciona un método para detectar la presencia o ausencia de una variación genética, que incluye: (a) obtener una muestra que incluya ácido nucleico libre de células circulante de un sujeto de prueba; (b) aislar el ácido nucleico de muestra de la muestra; (c) obtener lecturas de secuencia de nucleótidos a partir de un ácido nucleico de muestra; (d) mapear las lecturas de secuencia de nucleótidos en secciones genómicas de referencia, (e) contar el número de lecturas de secuencia de nucleótidos mapeadas en cada sección genómica de referencia, obteniendo así recuentos; (f) ajustar las lecturas de secuencia mapeadas contadas en (e) según una variable o característica seleccionada, cuya variable o característica seleccionada minimiza o elimina el efecto de secuencias repetitivas y/o secuencias sobrerrepresentadas o subrepresentadas; (g) normalizar los recuentos restantes después de (f) para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que incluye muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; y (h) proporcionar un resultado determinante de la presencia o ausencia de una variación genética en el sujeto de prueba a partir de los recuentos de muestra normalizados.

En el presente documento también se proporciona un método para detectar la presencia o ausencia de una variación genética, que incluye: (a) mapear en secciones genómicas de referencia lecturas de secuencia de nucleótidos obtenidas de ácido nucleico de muestra, incluyendo ácido nucleico libre de células circulante de un sujeto de prueba; (b) contar el número de lecturas de secuencia de nucleótidos mapeadas en cada sección genómica de referencia, obteniendo así recuentos; (c) ajustar las lecturas de secuencia mapeadas contadas en (b) según una variable o característica seleccionada, cuya variable o característica seleccionada minimiza o elimina el efecto de secuencias repetitivas y/o secuencias sobrerrepresentadas o subrepresentadas; (d) normalizar los recuentos restantes después de (c) para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que incluye muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; y (e) proporcionar un resultado determinante de la presencia o ausencia de una variación genética en el sujeto de prueba a partir de los recuentos de muestra normalizados.

En el presente documento también se proporciona un método para detectar la presencia o ausencia de una microdelección, que incluye: (a) obtener lecturas de secuencia de nucleótidos a partir de ácido nucleico de muestra, incluyendo ácido nucleico libre de células circulante de un sujeto de prueba; (b) mapear las lecturas de secuencia de nucleótidos en secciones genómicas de referencia; (c) contar el número de lecturas de secuencia de nucleótidos mapeadas en cada sección genómica de referencia, obteniendo así recuentos; (d) ajustar las lecturas de secuencia mapeadas contadas en (c) según una variable o característica seleccionada, cuya característica o variable seleccionada minimiza o elimina el efecto de secuencias repetitivas y/o secuencias sobrerrepresentadas o subrepresentadas; (e) normalizar los recuentos restantes después de (d) para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que incluye muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; y (f) proporcionar un resultado determinante de la presencia o ausencia de una variación genética en el sujeto de prueba a partir de los recuentos de muestra normalizados.

En el presente documento se proporciona un método para detectar la presencia o ausencia de un microdelección, que incluye: (a) obtener una muestra que incluya ácido nucleico libre de células circulante de un sujeto de prueba; (b) aislar el ácido nucleico de muestra de la muestra; (c) obtener lecturas de secuencia de nucleótidos a partir de un ácido nucleico de muestra; (d) mapear las lecturas de secuencia de nucleótidos en secciones genómicas de referencia, (e) contar el número de lecturas de secuencia de nucleótidos mapeadas en cada sección genómica de referencia, obteniendo así recuentos; (f) ajustar las lecturas de secuencia mapeadas contadas en (e) según una variable o característica seleccionada, cuya característica o variable seleccionada minimiza o elimina el efecto de secuencias repetitivas y/o secuencias sobrerrepresentadas o subrepresentadas; (g) normalizar los recuentos restantes después de (f) para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que incluye muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; y (h) proporcionar un resultado determinante de la presencia o ausencia de una variación genética en el sujeto de prueba a partir de los recuentos de muestra normalizados.

En el presente documento también se proporciona un método para detectar la presencia o ausencia de una microdelección, que incluye: (a) mapear en secciones genómicas de referencia lecturas de secuencia de nucleótidos obtenidas de ácido nucleico de muestra, incluyendo ácido nucleico libre de células circulante de un sujeto de prueba; (b) contar el número de lecturas de secuencia de nucleótidos mapeadas en cada sección genómica de referencia, obteniendo así recuentos; (c) ajustar las lecturas de secuencia mapeadas contadas en (b) según una variable o característica seleccionada, cuya característica o variable seleccionada minimiza o elimina el efecto de secuencias repetitivas y/o secuencias sobrerrepresentadas o subrepresentadas; (d) normalizar los recuentos restantes después de (c) para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que incluye muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; y (f) proporcionar un resultado determinante de la presencia o ausencia de una variación genética en el sujeto de prueba a partir de los recuentos de muestra normalizados.

En algunos casos, las lecturas de secuencia ajustadas, contadas mapeadas se ajustan adicionalmente para una o más condiciones experimentales antes de normalizar los recuentos restantes. En determinados casos, la variación genética es una microdelección. En algunos casos, la microdelección está en el cromosoma 22. En determinados casos, la microdelección se produce en la región 22q11.2 del cromosoma 22. En algunos casos, la microdelección se produce en el cromosoma 22 entre las posiciones de nucleótidos 19.000.000 y 22.000.000 según el genoma de referencia hg19.

En algunas realizaciones, el ácido nucleico de muestra es de plasma sanguíneo del sujeto de prueba. El ácido nucleico de muestra puede ser de suero sanguíneo del sujeto de prueba. El sujeto de prueba puede elegirse de un ser humano, un animal y una planta. Un sujeto de prueba humano puede incluir una mujer, una mujer embarazada, un hombre, un feto o un recién nacido.

En algunas realizaciones, la aneuploidía fetal es trisomía 13. En determinadas realizaciones, la aneuploidía fetal es trisomía 18. En algunas realizaciones, la aneuploidía fetal es trisomía 21.

En determinados casos, la variación genética está asociada con una afección médica. En algunos casos, la afección médica es cáncer. En determinados casos, la afección médica es una aneuploidía.

En algunos casos, las lecturas de secuencia del ácido nucleico de muestra libre de células están en forma de fragmentos de polinucleótidos. En determinados casos, los fragmentos de polinucleótidos tienen entre aproximadamente 20 y aproximadamente 50 nucleótidos de longitud. En algunos casos, los polinucleótidos tienen entre aproximadamente 30 y aproximadamente 40 nucleótidos de longitud. En algunos casos, la expresión "fragmento

de polinucleótido” es sinónimo o puede intercambiarse con la expresión “información de secuencia”, con referencia a lecturas de secuencia o una representación digital del ADN físico, o viceversa.

En determinadas realizaciones, el recuento esperado es una mediana de recuento. En algunas realizaciones, el recuento esperado es una media recortada o truncada, una media Winsorizada o una estimación de remuestreo de tipo *bootstrapping*. En determinados casos, el recuento de muestra normalizado se obtiene mediante un proceso que incluye la normalización de la derivada de los recuentos de la primera sección genómica, cuya derivada es una representación del recuento de la primera sección genómica determinada al dividir los recuentos de la primera sección genómica por los recuentos para múltiples secciones genómicas que incluyen la primera sección genómica. En algunos casos, la derivada de los recuentos de la primera sección genómica se normaliza según una derivada del recuento esperado, cuya derivada del recuento esperado es una representación del recuento de la primera sección genómica esperado determinada al dividir el recuento esperado de la primera sección genómica por el recuento esperado para múltiples secciones genómicas que incluyen la primera sección genómica. En determinados casos, la primera sección genómica es un cromosoma o parte de un cromosoma y las múltiples secciones genómicas incluyen autosomas. En algunos casos, el cromosoma es el cromosoma 21, el cromosoma 18 o el cromosoma 13.

En determinados casos, el recuento de muestra normalizado se obtiene mediante un proceso que incluye restar el recuento esperado de los recuentos para la primera sección genómica, generando así un valor de resta, y dividir el valor de resta por una estimación de la variabilidad del recuento. En algunos casos, el recuento de muestra normalizado se obtiene mediante un proceso que incluye restar la representación del recuento de la primera sección genómica esperado de la representación del recuento de la primera sección genómica, generando así un valor de resta, y dividir el valor de la resta por una estimación de la variabilidad de la representación del recuento de la primera sección genómica. En determinadas realizaciones, la estimación de la variabilidad del recuento esperado es una mediana de desviación absoluta (MAD) del recuento. En algunos casos, la estimación de la variabilidad del recuento es una alternativa a la MAD introducida por Rousseeuw y Croux o una estimación de remuestreo de tipo *bootstrapping*.

En algunas realizaciones, la una o más condiciones experimentales comunes incluyen una celda de flujo. En determinados casos, la una o más condiciones experimentales comunes incluyen un canal en una celda de flujo. En algunas realizaciones, la una o más condiciones experimentales comunes incluyen una placa de reactivo. En determinados casos, la placa de reactivo se usa para organizar el ácido nucleico para la secuenciación. En algunos casos, la placa de reactivo se usa para preparar una biblioteca de ácidos nucleicos para la secuenciación. En determinados casos, la una o más condiciones experimentales comunes incluyen un índice de etiquetas de identificación.

En determinadas realizaciones, el recuento de muestra normalizado se ajusta para el contenido de guanina y citosina de las lecturas de secuencia de nucleótidos o del ácido nucleico de muestra. En algunos casos, los métodos descritos en el presente documento incluyen someter los recuentos o el recuento de muestra normalizado a una regresión polinomial ponderada localmente. En determinados casos, la regresión polinomial ponderada localmente es una regresión LOESS o una regresión LOWESS. En algunos casos, el recuento de muestra normalizado se ajusta para las secuencias de nucleótidos que se repiten en las secciones genómicas de referencia. En determinados casos, los recuentos o el recuento de muestra normalizado se ajustan para las secuencias de nucleótidos que se repiten en las secciones genómicas de referencia. En algunos casos, el método incluye filtrar los recuentos antes de obtener el recuento de muestra normalizado.

En algunos casos, el ácido nucleico de muestra incluye ácido nucleico monocatenario. En determinados casos, el ácido nucleico de muestra incluye ácido nucleico bicatenario. En algunos casos, obtener las lecturas de secuencia de nucleótidos incluye someter el ácido nucleico de muestra a un proceso de secuenciación usando un dispositivo de secuenciación. En determinados casos, proporcionar un resultado incluye la factorización de la fracción de ácido nucleico fetal en el ácido nucleico de muestra. En algunos casos, el método incluye determinar la fracción de ácido nucleico fetal en el ácido nucleico de muestra.

En determinados casos, el recuento de muestra normalizado se obtiene sin ajustar el contenido de guanina y citosina de las lecturas de secuencia de nucleótidos o del ácido nucleico de muestra. En algunos casos, el recuento de muestra normalizado se obtiene para una condición experimental. En determinados casos, la condición experimental es una celda de flujo. En algunos casos, el recuento de muestra normalizado se obtiene para dos condiciones experimentales. En determinados casos, las condiciones experimentales son una celda de flujo y una placa de reactivo. En algunos casos, las condiciones experimentales son una celda de flujo y un índice de etiquetas de identificación. En algunos casos, el recuento de muestra normalizado se obtiene para tres condiciones experimentales. En determinados casos, las condiciones experimentales son una celda de flujo, una placa de reactivo y un índice de etiquetas de identificación.

En algunos casos, el recuento de muestra normalizado se obtiene después de (i), el ajuste según el contenido de guanina y citosina, y después de (i), (ii), el ajuste según una condición experimental. En determinados casos, el recuento de muestra normalizado se obtiene tras el ajuste según las secuencias de nucleótidos que se repiten en las secciones genómicas de referencia anteriores a (i). En algunos casos, (ii) consiste en el ajuste según la celda de flujo. En determinados casos, (ii) consiste en el ajuste según el índice de etiquetas de identificación y a continuación el ajuste según la celda de flujo. En algunos casos, (ii) consiste en el ajuste según la placa de reactivo y a continuación el

ajuste según la celda de flujo. En determinados casos, (ii) consiste en el ajuste según el índice de etiquetas de identificación y la placa de reactivo y a continuación el ajuste según la celda de flujo.

- En determinados casos, el recuento de muestra normalizado se obtiene después del ajuste según una condición experimental que consiste en el ajuste según la celda de flujo. En algunos casos, el recuento de muestra normalizado se obtiene después del ajuste según una condición experimental que consiste en el ajuste según el índice de etiquetas de identificación y a continuación el ajuste según la celda de flujo. En determinados casos, el recuento de muestra normalizado se obtiene después del ajuste según una condición experimental que consiste en el ajuste según la placa de reactivo y a continuación el ajuste según la celda de flujo. En algunos casos, el recuento de muestra normalizado se obtiene después del ajuste según una condición experimental que consiste en el ajuste según el índice de etiquetas de identificación y la placa de reactivo y a continuación el ajuste según la celda de flujo. En determinados casos, el recuento de muestra normalizado se obtiene después del ajuste según las secuencias de nucleótidos que se repiten en las secciones genómicas de referencia antes del ajuste según la condición experimental.
- En determinados casos, algunos métodos incluyen además la evaluación de la significación estadística de las diferencias entre los recuentos de muestra normalizados, o una derivada de los recuentos de muestra normalizados, para el sujeto de prueba y otras muestras, referencias o muestras y la referencia para una primera sección genómica. En algunos casos, determinados métodos incluyen además la evaluación de la significación estadística de las diferencias entre los recuentos de muestra normalizados, o una derivada de los recuentos de muestra normalizados, para el sujeto de prueba y otras muestras, referencias o muestras y la referencia para una o más secciones genómicas. En determinados casos, algunos métodos incluyen además proporcionar un resultado determinante de la presencia o ausencia de una variación genética en el sujeto de prueba basándose en la evaluación. En algunos casos, la variación genética se elige de una microdelección, duplicación y aneuploidía.
- En algunos casos, también se proporciona un producto de programa informático, incluyendo un medio utilizable por ordenador que tiene incorporado un código de programa legible por ordenador, incluyendo el código de programa legible por ordenador distintos módulos de software que incluyen un módulo de recepción de secuencias, 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 identificar la presencia o ausencia de una variación genética en un ácido nucleico de muestra, incluyendo el método: (a) obtener, mediante el módulo de recepción de secuencias, lecturas de secuencia de nucleótidos a partir de ácido nucleico de muestra; (b) mapear, mediante el módulo de procesamiento lógico, las lecturas de secuencia de nucleótidos en secciones genómicas de referencia; (c) contar, mediante el módulo de procesamiento lógico, el número de lecturas de secuencia de nucleótidos mapeadas en cada sección genómica de referencia, obteniendo así los recuentos; (d) normalizar, mediante el módulo de procesamiento lógico, los recuentos para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que comprende muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; (e) generar, mediante el módulo de procesamiento lógico, un resultado determinante de la presencia o ausencia de una variación genética en el sujeto de prueba a partir del recuento de muestra normalizado; y (f) organizar, mediante el módulo de organización de visualización de datos en respuesta a lo determinado por el módulo de procesamiento lógico, una visualización de datos que indica la presencia o ausencia de la variación genética en el ácido nucleico de muestra.
- También se proporciona en determinados casos un aparato que incluye una memoria en la que se almacena un caso de producto de programa informático descrito en el presente documento. En algunos casos, el aparato incluye un procesador que implementa una o más funciones del caso de producto de programa informático descrito en el presente documento. En determinados casos, la una o más funciones del producto de programa informático especificado en el presente documento se implementan en un entorno basado en web.
- También se proporciona en determinados casos, un aparato que incluye un sistema basado en web en el que se implementa un producto de programa informático especificado en el presente documento. En algunos casos, el sistema basado en web comprende ordenadores, enrutadores y equipos de telecomunicaciones suficientes para la funcionalidad basada en web. En determinados casos, el sistema basado en web comprende computación en la nube en red, almacenamiento en la nube en red o computación en la nube en red y almacenamiento en la nube en red.
- También se proporciona en algunos casos un sistema que incluye un aparato de secuenciación de ácidos nucleicos y un aparato de procesamiento, en donde el aparato de secuenciación obtiene lecturas de secuencia de nucleótidos de un ácido nucleico de muestra, y el aparato de procesamiento obtiene las lecturas de secuencia de nucleótidos del aparato de secuenciación y realiza un método que incluye: (a) mapear las lecturas de secuencia de nucleótidos en secciones genómicas de referencia; (b) contar el número de lecturas de secuencia de nucleótidos mapeadas en cada sección genómica de referencia, obteniendo así recuentos; (c) normalizar los recuentos para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que comprende muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; y (d) proporcionar un resultado

determinante de la presencia o ausencia de una variación genética en el ácido nucleico de muestra a partir del recuento de muestra normalizado.

También se proporciona en el presente documento un método para identificar la presencia o ausencia de una microdelección de 22q11.2 entre las posiciones de nucleótidos del cromosoma 22 19.000.000 y 22.000.000 según el genoma de referencia humano hg19, incluyendo el método: (a) obtener una muestra que comprenda ácido nucleico libre de células circulante de un sujeto de prueba; (b) aislar el ácido nucleico de muestra de la muestra; (c) obtener lecturas de secuencia de nucleótidos a partir de un ácido nucleico de muestra; (d) mapear las lecturas de secuencia de nucleótidos en secciones genómicas de referencia, (e) contar el número de lecturas de secuencia de nucleótidos mapeadas en cada sección genómica de referencia, obteniendo así recuentos; (f) ajustar las lecturas de secuencia mapeadas contadas en (e) según una variable o característica seleccionada, cuya característica o variable seleccionada minimiza o elimina el efecto de secuencias repetitivas y/o secuencias sobrerrepresentadas o subrepresentadas; (g) normalizar los recuentos restantes después de (f) para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que comprende muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; (h) evaluar la significación estadística de las diferencias entre los recuentos normalizados o una derivada de los recuentos normalizados para el sujeto de prueba y los sujetos de referencia para una o más secciones genómicas seleccionadas correspondientes al cromosoma 22 entre las posiciones de nucleótidos 19.000.000 y 22.000.000; y (i) proporcionar un resultado determinante de la presencia o ausencia de una variación genética en el sujeto de prueba a partir de la evaluación en (h).

Determinados casos y realizaciones se describen con más detalle en la siguiente descripción, ejemplos, reivindicaciones y dibujos.

Breve descripción de los dibujos

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

La Fig. 1 ilustra gráficamente la fracción de ADN fetal para cada una de las muestras seleccionadas representadas en función de la edad gestacional.

La Fig. 2 ilustra gráficamente la fracción de ADN fetal para cada una de las muestras seleccionadas representadas en función de la edad materna.

La Fig. 3 ilustra gráficamente la fracción de ADN fetal para cada una de las muestras seleccionadas representadas en función del peso materno.

La Fig. 4 ilustra gráficamente el porcentaje de cromosoma 21 para cada una de las muestras seleccionadas representadas en función de las lecturas coincidentes del cromosoma 21 por celda de flujo.

La Fig. 5 ilustra gráficamente el porcentaje de cromosoma 21 para cada una de las muestras seleccionadas representadas en función de las lecturas coincidentes del cromosoma 21 por número de placa.

La Fig. 6 ilustra gráficamente el porcentaje de cromosoma 21 para cada una de las muestras seleccionadas representadas en función del instrumento Illumina usado para la secuenciación.

La Fig. 7 ilustra gráficamente la puntuación z del cromosoma 21 para cada una de las muestras seleccionadas representadas en función de la edad gestacional.

La Fig. 8 ilustra gráficamente la puntuación z del cromosoma 21 para cada una de las muestras seleccionadas representadas en función de la edad materna.

La Fig. 9 ilustra gráficamente la puntuación z del cromosoma 21 para cada una de las muestras seleccionadas representadas en función del peso materno.

La Fig. 10 ilustra gráficamente la puntuación z del cromosoma 21 para cada una de las muestras seleccionadas representadas en función de la concentración de bibliotecas.

La Fig. 11 ilustra una optimización de preparación de bibliotecas. La Fig. 11A muestra una comparación de la concentración de bibliotecas estandarizada preparada por un método de preparación de bibliotecas manual y semiautomatizado (n = 287). La Fig. 14B muestra puntuaciones z basadas en GCRM para cada una de las 93 muestras. Las muestras euploides confirmadas (n = 83) se muestran en color gris claro. Las muestras de trisomía 21 confirmadas (n = 10) se muestran en color gris oscuro.

La Fig. 12 muestra una comparación emparejada de puntuaciones z . Las puntuaciones z se calcularon para muestras emparejadas con puntuaciones z enmascaradas, repetidas y normalizadas de GC descritas anteriormente en el eje x y puntuaciones z de las mismas bibliotecas secuenciadas en 12 plex en el eje y . Las muestras clasificadas por análisis de cariotipo como trisomías para la Fig. 12A (cromosoma 21), la Fig. 12B (cromosoma 13), o la Fig. 12C (cromosoma 18) se muestran en color gris oscuro. Las muestras no afectadas para cada condición de aneuploidía se muestran en color gris claro. Las líneas horizontales y verticales en cada gráfico representan el punto de corte de clasificación respectivo para ese cromosoma ($z = 3$ para el cromosoma 21, $z = 3,95$ para los cromosomas 13 y 18).

La Fig. 13 muestra las puntuaciones Z (eje x) frente a la fracción fetal (eje y). La puntuación z específica del cromosoma para cada cromosoma aneuploide se representa frente a la proporción de ADN fetal (fracción fetal).

Las muestras clasificadas por análisis de cariotipo como trisomías para la Fig. 13A (cromosoma 21), la Fig. 13B (cromosoma 13), o la Fig. 13C (cromosoma 18) se muestran en color gris oscuro. Las muestras no afectadas para cada condición de aneuploidía se muestran en color gris claro. Las líneas horizontales en cada gráfico representan el punto de corte de clasificación respectivo para cada cromosoma ($z = 3$ para el cromosoma 21, $z = 3,95$ para los cromosomas 13 y 18). Las líneas verticales discontinuas en cada panel representan un ajuste lineal robusto de las muestras afectadas. Las líneas horizontales discontinuas en cada panel representan un ajuste lineal robusto de todas las muestras no afectadas.

La Fig. 14 muestra una comparación emparejada de puntuaciones z . Las puntuaciones Z se calcularon para 1269 muestras emparejadas con puntuaciones z enmascaradas, repetidas y normalizadas de GC descritas anteriormente en el eje x y puntuaciones z del ensayo de alto rendimiento en el eje y . Las muestras clasificadas por análisis de cariotipo como trisomías para la Fig. 14A (cromosoma 21), la Fig. 14B (cromosoma 13), o la Fig. 14C (cromosoma 18) se muestran en color gris oscuro. Las muestras no afectadas para cada condición de aneuploidía se muestran en color gris claro. Las líneas horizontales y verticales en cada gráfico representan el punto de corte de clasificación respectivo para ese cromosoma ($z = 3$ para el cromosoma 21, $z = 3,95$ para los cromosomas 13 y 18).

Descripción detallada

Se proporcionan métodos, procesos y aparatos mejorados útiles para identificar variaciones genéticas. 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.

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 determinados casos, la presencia o ausencia de una o más variaciones genéticas se determinado según un resultado proporcionado por 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. 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 tener cualquier longitud observada y, en algunos casos, tiene de aproximadamente 1 base o par de bases (pb) a 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). En algunos casos, una variación genética es una anomalía cromosómica (por ejemplo, aneuploidía), anomalía cromosómica parcial o mosaicismo, que se describen con mayor detalle a continuación.

Una variación genética para la cual se identifica la presencia o ausencia para un sujeto se asocia con una afección médica en determinados casos. 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 algunos casos, la predicción del sexo del feto se puede determinar 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. Los individuos con XX son de sexo femenino y XY son de sexo masculino, y las variaciones no limitantes incluyen XO, XYY, XXX y XXY.

Anomalías cromosómicas

En algunos casos, 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 heterocigiosidad, 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.

La monosomía generalmente es la falta de un cromosoma del complemento normal. La monosomía parcial puede producirse en translocaciones o deleciones desequilibradas, en las cuales solo una porción del cromosoma está presente en una sola copia. La monosomía de cromosomas sexuales (45, X) provoca el síndrome de Turner, por ejemplo.

La disomía generalmente es 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).

La trisomía generalmente es 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 los cromosomas sexuales se puede observar en mujeres (p. ej., 47, XXX) o en hombres (p. ej., 47, XXY en el síndrome de Klinefelter); o 47, XYY).

La tetrasomía y la pentasomía generalmente son 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, XXXYY y XYYYY.

Las anomalías cromosómicas pueden estar provocadas por una variedad de mecanismos. Los mecanismos incluyen, pero sin limitación, (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 merotética 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.

Una monosomía parcial, o trisomía parcial, generalmente es un desequilibrio del material genético causado por la pérdida o ganancia de parte de un cromosoma. Una monosomía parcial o trisomía parcial puede resultar de una translocación desequilibrada, en donde 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 la porción que existe en el cromosoma derivado) y solo una copia de parte del otro cromosoma implicado en el cromosoma derivado.

El mosaicismo generalmente es una 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 una trisomía 21 simple, pero durante el curso 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 curso de la división celular, uno de los cromosomas 21 se duplicó. 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.

- 5 Las Tablas 1A y 1B presentan una lista no limitativa de afecciones, síndromes y/o anomalías cromosómicas que pueden identificarse potencialmente mediante los métodos y aparatos descritos en el presente documento. La Tabla 1B es de la base de datos DECIPHER al 6 de octubre de 2011 (por ejemplo, versión 5.1, basada en posiciones mapeadas en GRCh37; disponible en el uniform resource locator (localizador uniforme de recursos - URL) dechipper.sanger.ac.uk).

10 Tabla 1a

Cromosoma	Anomalia	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 linfoproliferativa ligada al cromosoma X
1	monosomía-trisomía (somática) 1p	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	Maulido 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 7 de la infancia; somáticos: adenomas corticales renales; síndrome mielodisplásico
8	deleción de 8q24.1	síndrome de Langer-Giedon
8	monosomía-trisomía	síndrome mielodisplásico; síndrome de Warkany; somática: leucemia mielógena crónica
9	monosomía 9p	Síndrome de Alfi
9	monosomía-trisomía parcial 9p	Síndrome de Rethore
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	11 q-	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	monosomía con deleción de 15q11-q13	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
	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

Cromosoma	Anomalia	Asociación con la enfermedad
17	trisomía 17p11.2-12	Síndrome de Charcot-Marie-Tooth tipo 1; HNPP
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
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 conotruncuales y de la cara, síndrome de Opitz G/BBB autosómico dominante, síndrome cardiorfacial 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	
síndrome de microdeleción de 15q13.3	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	
16p13.11 recurrente	16	15.504.454	16.284.248	0,78	
microdeleción (locus de susceptibilidad a trastornos neurocognitivos)					
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 a trombocitopenia-síndrome de radio ausente (TAR)	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	

ES 2 929 923 T3

Síndrome	Cromosoma	Inicio	Fin	Intervalo (Mb)	Grado
síndrome de delección de 22q13 (síndrome de Phelan-Mcdermid)	22	51.045.516	51.187.844	0,14	1
síndrome de microdelección de 2p15-16.1	2	57.741.796	61.738.334	4,00	
delección de 2q33.1	2	196.925.089	205.206.940	8,28	1
síndrome					
monosomía de 2q37	2	239.954.693	243.102.476	3,15	1
síndrome de microdelección 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 delección de 8p23.1	8	8.119.295	11.765.719	3,65	
síndrome de delección 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 (delección 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
Poliposis adenomatosa familiar		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) - delección de SHOX	X	751.878	867.875	0,12	
Discondrosteosis de Leri-Weill (LWD) - delección de SHOX	X	460.558	753.877	0,29	
Síndrome de Miller-Dieker (SMD)	17	1	2.545.429	2,55	1
Síndrome de microdelección 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

Síndrome	Cromosoma	Inicio	Fin	Intervalo (Mb)	Grado
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 delección 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 a menudo tienen una o más de las siguientes características; anomalía patógena; fuerte acuerdo 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 informado en las bases de datos de la DVG ni encontrado en población sana; datos funcionales que confirman el efecto de dosificación de un solo gen o de múltiples genes; genes candidatos confirmados o fuertes; implicaciones de gestión clínica definidas; riesgo de cáncer conocido con implicaciones para la vigilancia; múltiples fuentes de información (OMIM, Genereviews, Orphanet, Unique, Wikipedia); y/o disponible para uso diagnóstico (asesoramiento reproductivo).

Las afecciones de grado 2 a menudo tienen una o más de las siguientes características; probable anomalía patógena; altamente penetrante; fenotipo variable sin características consistentes además de DD; pequeño número de casos/informes en la bibliografía; todos los casos informados tienen un fenotipo clínico; sin datos funcionales ni genes patógenos confirmados; múltiples fuentes de información (OMIM, Genereviews, Orphanet, Unique, Wikipedia); y/o puede usarse con fines de diagnóstico y asesoramiento reproductivo.

Las afecciones de grado 3 a menudo tienen una o más de las siguientes características; locus de susceptibilidad; individuos sanos o padres no afectados de un índice descrito; presente en poblaciones de control; no penetrante; fenotipo leve e inespecífico; características menos consistentes; sin datos funcionales ni genes patógenos confirmados; fuentes de datos más limitadas; la posibilidad de un segundo diagnóstico sigue siendo una posibilidad para los casos que se desvían de la mayoría o si se presenta un hallazgo clínico novedoso; y/o precaución cuando se usa con fines de diagnóstico y consejos cautelosos para el asesoramiento reproductivo.

Preeclampsia

En algunos casos, 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 morbilidad 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 algunos casos, 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 algunos casos, 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 lo tanto, se contempla además que los métodos proporcionados en el presente documento pueden usarse para identificar un tipo particular de cáncer.

Otras variaciones genéticas

En algunos casos, la presencia o ausencia de una variación genética se puede determinar usando un método o aparato descrito en el presente documento. En determinados casos, una variación genética es una o más condiciones elegidas de copy number variations (variaciones del número de copias - CNV), microdeleciones, duplicaciones o cualquier condición que cause o de como resultado una variación de la dosis genética de una dosis genética esperada observada en un individuo no afectado. En algunos casos, la variación del número de copias se refiere a reordenamientos estructurales de una o más secciones genómicas, cromosomas o partes de cromosomas, cuyo reordenamiento a menudo es causado por deleciones, duplicaciones, inversiones y/o translocaciones. Las CNV pueden ser heredadas o causadas por una mutación *de novo* y, normalmente, dan como resultado un número anormal de copias de una o más secciones genómicas (por ejemplo, dosificación génica anormal con respecto a una muestra no afectada). La variación del número de copias puede producirse en regiones que van desde tan solo una kilobase hasta varias megabases, en algunos casos. Las CNV se pueden detectar usando diversos métodos citogenéticos (FISH, CGH, aCGH, análisis de cariotipo) y/o métodos de secuenciación.

Una microdelección generalmente es una dosis reducida, con respecto a las regiones no afectadas, de material genético (por ejemplo, ADN, genes, ácido nucleico representativo de una región particular) ubicado en una sección o segmento genómico seleccionado. Las microdeleciones y los síndromes causados por microdeleciones a menudo se caracterizan por una pequeña deleción (por ejemplo, generalmente menos de cinco megabases) de uno o más segmentos cromosómicos, que abarcan uno o más genes, cuya ausencia a veces confiere una patología. Las microdeleciones a veces son causadas por errores en el cruce cromosómico durante la meiosis. En muchos casos, las microdeleciones no son detectables por los métodos de cariotipado utilizados actualmente.

Una duplicación cromosómica, o microduplicación o duplicación, generalmente es una o más regiones de material genético (por ejemplo, ADN, genes, ácido nucleico representativo de una región particular) para las cuales la dosis se aumenta en relación con las regiones no afectadas. Las duplicaciones se producen a menudo como resultado de un error en la recombinación homóloga o debido a un evento de retrotransposón. Las duplicaciones pueden variar desde regiones pequeñas (miles de pares de bases) hasta cromosomas enteros en algunos casos. 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 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).

Muestras

El ácido nucleico utilizado en los métodos y aparatos descritos en el presente documento a menudo se aísla de una muestra obtenida de un sujeto. En algunos casos, un sujeto se denomina sujeto de prueba y, en determinados casos, un sujeto se denomina sujeto de muestra o sujeto de referencia. En algunos casos, sujeto de prueba se refiere a un sujeto que está siendo evaluado para determinar la presencia o ausencia de una variación genética. Un sujeto de muestra, o sujeto de referencia, a menudo se utiliza como base para la comparación con el sujeto de prueba, y a veces se selecciona un sujeto de referencia en función del conocimiento de que se sabe que el sujeto de referencia está libre o tiene la variación genética que se está evaluando para el sujeto de prueba. 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 (p. ej., ganado vacuno), equino (p. ej., caballo), caprino y ovino (p. ej., oveja, cabra), porcino (p. ej., cerdo), camélido (p. ej., camello, llama, alpaca), mono, simio (p. ej., gorila, chimpancé), úrsido (p. ej., oso), ave de corral, perro, gato, ratón, rata, pescado, delfín, ballena y tiburón. Un sujeto puede ser del sexo masculino o del sexo femenino (p. ej., una mujer).

El ácido nucleico puede aislarse de cualquier tipo de muestra o espécimen biológico adecuado. 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 algunos casos, una muestra biológica puede ser sangre y, algunas veces, plasma o suero. Como se usa en el presente documento, "sangre" generalmente se refiere a sangre completa o cualquier fracción de sangre, tal como suero y plasma tal como se definen convencionalmente, por ejemplo. 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 de la preparación adicional. Una muestra de líquido o tejido de la que se extrae ácido nucleico puede ser acelular. En algunos casos, una muestra de líquido o tejido puede contener elementos celulares o restos celulares. En algunos casos, pueden incluirse células fetales o células cancerosas en la muestra.

Una muestra puede ser 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 sin limitación, (i) ácido nucleico derivado del feto y derivado de la madre, (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 algunos casos, están presentes una especie minoritaria de ácido nucleico y una mayoría de especies de ácido nucleico.

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 sometida a prueba para un posible embarazo. La edad gestacional adecuada puede variar dependiendo de la prueba prenatal que se realiza. En determinados casos, 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 determinados casos, 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).

Aislamiento y procesamiento de ácidos nucleicos

El ácido nucleico puede proceder de una o más fuentes (por ejemplo, células, suelo, etc.) mediante métodos conocidos en la técnica. Se conocen en la técnica procedimientos y reactivos de lisis celular y generalmente pueden realizarse mediante métodos de lisis electrolítica, física o química. 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 los últimos procedimientos, una solución puede contener Tris 15 mM, pH 8,0; EDTA 10 mM y 100 ug/ml de ARNasa A; una segunda solución puede contener NaOH 0,2 N y SDS al 1 %; y una tercera solución 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), incorporado en el presente documento en su totalidad.

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 determinados casos. Un ácido nucleico en algunos casos puede ser de un solo cromosoma (por ejemplo, una muestra de ácido nucleico puede proceder de un cromosoma de una muestra obtenida de un organismo diploide). 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 donde 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 determinados casos. El ácido nucleico extracelular a menudo es un ácido nucleico aislado de una fuente que sustancialmente no tiene células. 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 plasma sanguíneo, suero sanguíneo y orina. 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 grande (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 determinados casos. 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 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 algunos casos, 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 algunos casos, 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 algunos casos, 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 algunos casos, 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 algunos casos, 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).

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 determinados casos. En algunos casos, 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 o amplificarse de la una o más muestras. Como se usa en el presente documento, "aislado" 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 lo tanto, se altera por intervención humana (por ejemplo, "por la mano del hombre") de su entorno original. Un ácido nucleico aislado se proporciona 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 aproximadamente el 90 %, 91 %, 92 %, 93 %, 94 %, 95 %, 96 %, 97 %, 98 %, 99 % o más del 99 % de componentes distintos de ácido nucleico. Como se usa en el presente documento, "purificado" se refiere 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 puede estar aproximadamente el 90 %, 91 %, 92 %, 93 %, 94 %, 95 %, 96 %, 97 %, 98 %, 99 % o más del 99 % libre de otras especies de ácido nucleico. Un ácido nucleico amplificado a menudo se prepara sometiendo el ácido nucleico de una muestra a un proceso que genera lineal o exponencialmente ácidos nucleicos de amplicón que tienen la misma o sustancialmente la misma secuencia de nucleótidos que la secuencia de nucleótidos del ácido nucleico en la muestra, o parte de la misma.

El ácido nucleico puede procesarse además al someter el ácido nucleico a un método que genera fragmentos de ácido nucleico, en determinados casos, antes de proporcionar ácido nucleico para un procedimiento descrito en el presente documento. En algunos casos, 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 cualquier 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 determinados casos, 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 algunos casos, 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 una porción del mismo. Por ejemplo, un fragmento puede tener subsecuencias x e y, y otro fragmento puede tener subsecuencias y y z, en donde 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 determinados casos. El ácido nucleico puede estar parcialmente fragmentado (por ejemplo, a partir de una reacción de escisión específica incompleta o terminada) o completamente fragmentado en determinados casos.

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.º 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 limitantes de procesos que pueden generar ácido nucleico en fragmentos no específicamente escindido incluyen, sin limitación, poner en contacto el ácido nucleico con un aparato que expone el ácido nucleico a la fuerza de cizallamiento (por ejemplo, pasando el ácido nucleico a través de una aguja de jeringa; uso de una prensa francesa); exponer el ácido nucleico a la irradiación (por ejemplo, rayos gamma, rayos X, radiación UV; el tamaño de los fragmentos pueden controlarse mediante la intensidad de la irradiación); hervir ácido nucleico en agua (por ejemplo, produce aproximadamente 500 fragmentos de pares de bases) y exponer el ácido nucleico a un proceso de hidrólisis de 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 la porción 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.

Como se usa en el presente documento, “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 determinados casos, 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).

Un ácido nucleico se puede escindir específicamente poniendo en contacto el ácido nucleico con uno o más agentes de escisión específicos. Como se usa en el presente documento, un “agente de escisión específico” 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 ejemplos de agentes de escisión específicos enzimáticos incluyen, sin limitación, endonucleasas (por ejemplo, DNasa I, II); RNasa (por ejemplo, RNasa E, F, H, P); la enzima Cleavase™; ADN polimerasa Taq; ADN polimerasa I de *E. coli* y endonucleasas específicas de estructura eucariota; 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 II, EcoR V, Hae II, Hae III, 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, emparejamiento erróneo de timina-ADN glicosilasa, hipoxantina-ADN glicosilasa, 5-hidroximetiluracilo-ADN glicosilasa (HmIDG), 5-hidroximetilcitosina-ADN glicosilasa o 1,N6-eteno-adenina-ADN glicosilasa); exonucleasas (por ejemplo, exonucleasa III); ribozimas y DNAzimas. 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, incluidas N3-metiladenina y N3-metilguanina, que son reconocidas y escindidas por la ADN-glicosilasa de alquil purina; (ii) bisulfito de sodio, que provoca la desaminación de residuos de citosina en el ADN para formar residuos de uracilo que pueden escindirse por uracil N-glicosilasa; y (iii) un agente químico que convierte la guanina en su forma oxidada, 8-hidroxiguanina, que puede escindirse por la ADN N-glicosilasa de formamidopirimidina. Los ejemplos de procesos de escisión química incluyen, sin limitación, alquilación (por ejemplo, alquilación de ácido nucleico modificado con fosforotioato); escisión de la labilidad ácida del ácido nucleico que contiene P3'-N5'-fosforamidato; y tratamiento de ácido nucleico con tetróxido de osmio y piperidina.

En algunos casos, el ácido nucleico fragmentado puede someterse a un procedimiento de fraccionamiento por tamaño y la totalidad o parte del grupo fraccionado puede aislarse o analizarse. Se conocen en la técnica procedimientos de fraccionamiento por tamaños (por ejemplo, separación en una matriz, separación mediante un tamiz molecular, separación por electroforesis en gel, separación por cromatografía en columna).

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 determinados casos, 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.

5 Obtención de lecturas de secuencia

Los métodos de secuenciación, mapeo y analíticos relacionados se conocen en la técnica (por ejemplo, la publicación de solicitud de patente estadounidense US2009/0029377). Determinados aspectos de tales procedimientos se describen más adelante en el presente documento.

10 Generalmente, las lecturas 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 ("lecturas de doble extremo"). En determinados casos, "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 algunos casos, "obtener" puede implicar recibir información de secuencia obtenida directamente de un ácido nucleico por otro.

20 En algunos casos, se secuencia una muestra de ácido nucleico de un individuo. En determinados casos, las muestras de ácido nucleico de dos o más muestras biológicas, en donde cada muestra biológica es de un individuo o dos o más individuos, se combinan y se secuencia la combinación. En los últimos casos, una muestra de ácido nucleico de cada muestra biológica se identifica a menudo por medio de una o más etiquetas de identificación únicas.

25 En algunos casos, 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 donde una región dada del genoma puede cubrirse por dos o más lecturas o lecturas solapantes (p. ej., "veces" de cobertura mayores de 1). En algunos casos, 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 (p. ej., 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).

35 En determinados casos, una fracción de una combinación de ácidos nucleicos que se secuencia en una ejecución se subselecciona adicionalmente antes de la secuenciación. En determinados casos, 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 algunos casos, el ácido nucleico puede fraccionarse por tamaños (por ejemplo, mediante electroforesis en gel, cromatografía de exclusión molecular o mediante un enfoque basado en microfluidos) 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 algunos casos, el ácido nucleico fetal puede enriquecerse suprimiendo el ácido nucleico de fondo materno, tal como mediante la adición de formaldehído. En algunos casos, una porción o un subconjunto de una combinación preseleccionada de ácidos nucleicos se secuencia aleatoriamente. En algunos casos, el ácido nucleico se amplifica antes de la secuenciación. En algunos casos, una porción o un subconjunto del ácido nucleico se amplifica antes de la secuenciación.

50 Puede usarse cualquier método de secuenciación adecuado para realizar los métodos descritos en el presente documento. En algunos casos, 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 Chem* 55:641-658 (2009)). Tales métodos de secuenciación pueden proporcionar además información cuantitativa digital, en donde cada lectura de secuencia es una "etiqueta de secuencia" contable que representa un molde de ADN clonal individual o una molécula de ADN individual. 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.

60 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 una 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 algunos casos, la tecnología de primera generación, tal como, por ejemplo, secuenciación de Sanger que incluye la secuenciación automatizada de Sanger, puede usarse en los métodos proporcionados en el presente documento. Además, en el presente documento también se contemplan tecnologías de secuenciación adicionales que incluyen el uso de tecnologías de obtención de imágenes de ácidos nucleicos en desarrollo (por ejemplo, transmission electron microscopy (microscopía electrónica de transmisión -TEM) y atomic force microscopy (microscopía 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 los métodos descritos en el presente documento es secuenciación por síntesis y secuenciación basada en el terminador reversible (por ejemplo, el analizador genómico de Illumina y Genome Analyzer II). 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 se fragmenta ADN molde (por ejemplo, ADN circulante, libre de células (ADNclic)) en longitudes de varios cientos de pares de bases en la preparación para la generación de bibliotecas. En algunos casos, la preparación de bibliotecas puede realizarse sin fragmentación adicional o selección de tamaño del ADN de molde (por ejemplo, ADNclic). En determinados casos, la generación de bibliotecas se realiza mediante una modificación del protocolo de los fabricantes, como se describe en el ejemplo 2. El aislamiento de muestras y la generación de bibliotecas se realizan usando métodos y aparatos automatizados, en determinados casos. Brevemente, el ADNclic 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 ADNclic 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 ADNclic modificado (por ejemplo, reparado en los extremos y extendido con un solo nucleótido) con un soporte sólido, la superficie interior de una celda de flujo, por ejemplo. En algunos casos, el cebador adaptador incluye nucleótidos indexadores o nucleótidos de "código de barras" (por ejemplo, una secuencia única de nucleótidos que puede usarse como un cebador indexador para permitir la identificación inequívoca de una muestra), 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 emparejados, cebadores de secuenciación multiplexados, y similares), o combinaciones de los mismos (por ejemplo, adaptador/secuenciación, adaptador/indexación, adaptador/indexación/secuenciación). Los cebadores indexadores o nucleótidos contenidos en un cebador adaptador tienen a menudo seis o más nucleótidos de longitud y frecuentemente se posicionan en el cebador de tal manera que los nucleótidos indexadores son los primeros nucleótidos secuenciados durante la reacción de secuenciación. En determinados casos, los nucleótidos indexadores o de código de barras 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 de código de barras y la secuenciación de la muestra se ligan entre sí y las lecturas se desmultiplexan. Después del ligamiento y la desmultiplexación, las lecturas de secuencia pueden ajustarse o procesarse adicionalmente como se describe en el presente documento.

En determinados procedimientos de secuenciación por síntesis, el uso de cebadores indexadores 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 cebadores indexadores únicos utilizados durante la preparación de la biblioteca. Están disponibles cebadores indexadores en varias fuentes comerciales (por ejemplo, Illumina, Life Technologies, NEB). Las reacciones descritas en el ejemplo 2 se realizaron usando uno de los pocos kits disponibles comercialmente disponibles en el momento del estudio, que incluía 12 cebadores indexadores únicos. Los ejemplos no limitativos de kits de secuenciación múltiple 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). Los métodos descritos en el presente documento no se limitan a 12 cebadores indexadores y se pueden realizar usando cualquier número de cebadores indexadores únicos (por ejemplo, 4, 8, 12, 24, 48, 96 o más). Cuanto mayor sea el número de cebadores indexadores únicos, mayor será el número de muestras que se pueden multiplexar en un solo carril de celda de flujo.

La multiplexación usando 12 cebadores indexadores permite analizar simultáneamente 96 muestras (por ejemplo, el mismo número de pocillos en una placa de micropocillos de 96 pocillos) en una celda de flujo de 8 carriles. De manera similar, la multiplexación con 48 cebadores indexadores permite analizar simultáneamente 384 muestras (por ejemplo, el mismo 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 50×10^6 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 50×10^6 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 los métodos descritos 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 donde 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 sulfúrida y luciferasa en un dispositivo que contiene pocillos de tamaño de micropocillos. 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 sulfúrida 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 los métodos proporcionados 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 los métodos descritos 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/cm². 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 los métodos proporcionados 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 los métodos descritos 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 los métodos descritos en el presente documento es la matriz de chemical-sensitive field effect transistor (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 patente estadounidense n.º 2009/0026082).

Otra tecnología de secuenciación de ácidos nucleicos que puede usarse en los métodos descritos en el presente documento es la microscopía 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 Individual Molecule Placement Rapid Nano Transfer (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 publicación de patente PCT 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 algunos casos. 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 algunos casos, 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 donde 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 algunos casos. El patrón de hibridación al alineamiento puede usarse para determinar las secuencias de polinucleótidos presentes en la muestra. En algunos casos, 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 algunos casos, puede usarse secuenciación por nanoporos en los métodos descritos 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 dan 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. Por lo tanto, la cantidad de corriente que puede pasar a través del nanoporo en un momento dado varía dependiendo de si el nanoporo está bloqueado por una A, una C, una G, una T o, en algunos casos, metil-C. El cambio en la corriente a través del nanoporo cuando 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 Chem 53: 1996-2001 (2007); publicación PCT n.º WO2010/004265).

Existen varias maneras en que pueden usarse nanoporos para secuenciar moléculas de ácido nucleico. En algunos casos, 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 caso, 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 interactuar 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 algunos casos, 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 caso, la enzima puede unirse a la estructura del nanoporo de tal manera que pueda empujar o extraer el ácido nucleico 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 algunos casos, 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 algunos casos, 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 algunos casos, las lecturas de secuencia tienen una longitud media, mediana de longitud o longitud promedio de aproximadamente 1000 pb o más.

En algunos casos, 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.

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”. En algunos casos, 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.

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 BLAST, BLITZ y FASTA, o variaciones de los mismos. En algunos casos, 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. A continuación, 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. Las lecturas de secuencia generadas en los ejemplos 1, 2 y 3 se mapearon en el genoma de referencia humano UCSC hg19 usando CASAVA versión 1.6, como se describe en los ejemplos 2 y 3. En algunos casos, el mapeo de lectura de secuencia se puede realizar antes del ajuste de secuencias repetitivas y/o contenido de GO y, en determinados casos, el mapeo de lectura de secuencia se puede realizar después del ajuste de secuencias repetitivas y/o contenido de GC.

Una “etiqueta de secuencia” es 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 una sola porción del genoma de referencia (por ejemplo, un cromosoma). En algunos casos, las etiquetas de secuencia repetitiva se eliminan del análisis adicional (por ejemplo, cuantificación). En algunos casos, 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 algunos casos, las lecturas mapeadas de manera no única se eliminan del análisis posterior (por ejemplo, cuantificación). Puede permitirse que un determinado grado pequeño de emparejamiento 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

determinados casos. En algunos casos, no se permite ningún grado de emparejamiento erróneo para que una lectura se mapee en una secuencia de referencia.

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 determinados casos, en donde 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 algunos casos. 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 determinados casos, la referencia externa proviene del ADN de uno o más individuos que tienen 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).

Secciones genómicas

En algunos casos, 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 algunos casos, 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" también se puede usar indistintamente con "ventana de secuencia", "sección", "bin", "locus", "región", "división" o "segmento". En algunos casos, una sección genómica es un cromosoma completo, una porción de un cromosoma, múltiples porciones de cromosomas, múltiples cromosomas, porciones de múltiples cromosomas y/o combinaciones de los mismos. 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. En algunos casos, una sección genómica se basa en una longitud particular de secuencia genómica. En algunos casos, los métodos incluyen el análisis de múltiples lecturas de secuencias mapeadas en una pluralidad de secciones genómicas. Las secciones genómicas pueden tener aproximadamente la misma longitud o las secciones genómicas pueden tener longitudes diferentes. En algunos casos, 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 algunos casos, la sección genómica tiene de aproximadamente 10 kb a aproximadamente 20 kb. Las secciones genómicas analizadas en el presente documento no se limitan a ejecuciones contiguas de secuencia. Por lo tanto, las secciones genómicas pueden estar formadas por secuencias contiguas o no contiguas. Las secciones genómicas analizadas en el presente documento no se limitan a un solo cromosoma y, en algunos casos, pueden trascender los cromosomas individuales. 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 algunos casos, las secciones genómicas pueden ser secciones cromosómicas particulares en un cromosoma de interés, tales como, por ejemplo, un cromosoma en donde se evalúa una variación genética (por ejemplo, una aneuploidía de los cromosomas 13, 18 y/o 21). 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 algunos casos, 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 algunos casos, la división del genoma en regiones que trascienden 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). En algunos casos, la división del genoma en regiones que trascienden 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, presencia de secuencias repetitivas, 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, nivel 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 de etiquetas o lecturas de secuencia para una sección genómica definida, en donde la densidad de etiqueta de secuencia se usa para comparar diferentes muestras y para el análisis posterior. En algunos casos, el valor de las etiquetas de secuencia es un valor normalizado de las etiquetas de secuencia. El valor de la densidad de etiqueta de secuencia a veces se normaliza dentro de una muestra y, a veces, se normaliza a una mediana de valor para un grupo de muestras (por ejemplo, muestras procesadas en un carril de flujo, muestras preparadas en una placa de generación de bibliotecas, muestras recogidas en una placa de clasificación, similares y combinaciones de los mismos).

En algunos casos, la normalización se puede realizar contando el número de etiquetas que se encuentran dentro de cada sección genómica; obteniendo un valor de la mediana, moda, promedio o punto medio del recuento total de etiquetas de secuencia para cada cromosoma; obteniendo un valor de la mediana, moda, promedio o punto medio de todos los valores autosómicos; y usando este valor como una constante de normalización para tener en cuenta las diferencias en el número total de etiquetas de secuencia obtenidas para diferentes muestras. En determinados casos, la normalización se puede realizar contando el número de etiquetas que se encuentran dentro de cada sección genómica para todas las muestras en una celda de flujo; obteniendo un valor de la mediana, moda, promedio o punto medio del recuento total de etiquetas de secuencia para cada cromosoma para todas las muestras en una celda de flujo; obteniendo una mediana del valor, un valor de moda, promedio o punto medio de todos los valores autosómicos para todas las muestras en una celda de flujo; y usando este valor como una constante de normalización para tener en cuenta las diferencias en el número total de etiquetas de secuencia obtenidas para diferentes muestras procesadas en paralelo en una celda de flujo. En algunos casos, la normalización se puede realizar contando el número de etiquetas que se encuentran dentro de cada sección genómica para todas las muestras preparadas en una placa (por ejemplo, placa de reactivo, placa de micropocillos); obteniendo un valor de la mediana, moda, promedio o punto medio del recuento total de etiquetas de secuencia para cada cromosoma para todas las muestras preparadas en una placa, obteniendo una mediana del valor, un valor de moda, promedio o punto medio de todos los valores autosómicos para todas las muestras preparadas en una placa; y usando este valor como una constante de normalización para tener en cuenta las diferencias en el número total de etiquetas de secuencia obtenidas para diferentes muestras procesadas en paralelo en una placa.

Una densidad de etiqueta de secuencia algunas veces es de aproximadamente 1 para un cromosoma disómico. Las densidades de etiquetas de secuencia pueden variar según los artefactos de secuenciación, más particularmente el sesgo de G/C, los efectos de procesamiento por lotes (por ejemplo, preparación de muestras) y similares, que pueden corregirse mediante el uso de un estándar externo o una referencia interna (por ejemplo, procedente de sustancialmente todas las etiquetas de secuencia (secuencias genómicas), que pueden ser, por ejemplo, un solo cromosoma, un valor calculado de todos los autosomas, un valor calculado de todas las muestras analizadas en una celda de flujo (un solo cromosoma o todos los autosomas), o un valor calculado de todas las muestras procesadas en una placa y analizadas en una o más celdas de flujo, en algunos casos). 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 algunos casos, 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 algunos casos. Si el feto tiene una aneuploidía en el 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 algunos casos.

Resultados y determinación de la presencia o ausencia de una variación genética

Algunas variaciones genéticas están asociadas con afecciones médicas. Las variaciones genéticas a menudo incluyen una ganancia, una pérdida y/o alteración (por ejemplo, duplicación, delección, fusión, inserción, mutación, reorganización, sustitución o metilación aberrante) de información genética (por ejemplo, cromosomas, porciones 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 un sujeto referencia libre de la variación genética. La presencia o ausencia de una variación genética se puede determinar analizando y/o manipulando lecturas de secuencia que se han mapeado en secciones genómicas (por ejemplo, bins genómicos) como se conoce en la técnica y se describe en el presente documento. En algunos casos, la presencia o ausencia de una afección, síndrome y/o anomalía conocida,

cuyos ejemplos no limitativos se proporcionan en las Tablas 1A y 1B, se puede detectar y/o determinar utilizando métodos descritos en el presente documento.

Recuento

Las lecturas de secuencia que se han mapeado o dividido en función de una característica o variable seleccionada se pueden cuantificar para determinar el número de lecturas que se mapearon en cada sección genómica (por ejemplo, bin, división, segmento genómico y similares), en algunos casos. En determinados casos, el número total de lecturas de secuencia mapeadas se determina contando todas las lecturas de secuencia mapeadas y, en algunos casos, el número total de lecturas de secuencia mapeadas se determina sumando los recuentos mapeadas en cada bin o división. En algunos casos, el recuento se realiza en el proceso de mapeo de lecturas. En determinados casos, un subconjunto de lecturas de secuencia mapeadas se determina contando un subconjunto predeterminado de lecturas de secuencia mapeadas y, en algunos casos, un subconjunto predeterminado de lecturas de secuencia mapeadas se determina sumando recuentos mapeados en cada bin o división predeterminados. En algunos casos, los subconjuntos predeterminados de lecturas de secuencia mapeadas pueden incluir de 1 a n lecturas de secuencia, en donde n representa un número igual a la suma de todas las lecturas de secuencia generadas a partir de una muestra del sujeto de prueba, una o más muestras del sujeto de referencia, todas las muestras procesadas en una celda de flujo, o todas las muestras preparadas en una placa para su análisis usando una o más celdas de flujo. Las lecturas de secuencia que se han mapeado y contado para una muestra del sujeto de prueba, una o más muestras del sujeto de referencia, todas las muestras procesadas en una celda de flujo o todas las muestras preparadas en una placa a veces se denominan recuento de muestra. Los recuentos de muestra a veces se distinguen además por referencia al sujeto del que se aisló la muestra (por ejemplo, recuento de muestras del sujeto de prueba, recuento de muestras del sujeto de referencia y similares).

En algunos casos, una muestra de prueba también se usa como muestra de referencia. A veces se usa una muestra de prueba como muestra de referencia y se determina una mediana del recuento esperado y/o una derivada de la mediana del recuento esperado para una o más secciones genómicas seleccionadas (por ejemplo, una primera sección genómica, una segunda sección genómica, una tercera sección genómica, 5 o más secciones genómicas, 50 o más secciones genómicas, 500 o más secciones genómicas y similares) que se sabe que están libres de variación genética (por ejemplo, no tienen microdeleciones, duplicaciones, aneuploidías y similares en la una o más secciones genómicas seleccionadas). La mediana del recuento esperado o una derivada de la mediana del recuento esperado para la una o más secciones genómicas libres de variación genética se puede usar para evaluar la importancia estadística de los recuentos obtenidos de otras secciones genómicas seleccionadas (por ejemplo, secciones genómicas diferentes a las utilizadas como secciones de muestra de referencia) de la muestra de prueba. En algunos casos, también se determina la mediana de desviación absoluta y, en determinados casos, la mediana de desviación absoluta también se usa para evaluar la significación estadística de los recuentos obtenidos de otras secciones genómicas seleccionadas de la muestra de prueba.

Según método de la invención, un proceso de normalización que normaliza los recuentos incluye el uso de un recuento esperado. En algunos casos, los recuentos de muestra se obtienen a partir de subconjuntos predeterminados de lecturas de secuencia mapeadas. En determinados casos, se pueden seleccionar subconjuntos predeterminados de lecturas de secuencia mapeadas utilizando cualquier característica o variable adecuada. En algunos casos, se utiliza un conjunto predeterminado de lecturas de secuencia mapeadas como base para la comparación, y puede denominarse “recuento de muestra esperado” o “recuento esperado” (colectivamente, un “recuento esperado”). Un recuento esperado a menudo es un valor obtenido en parte al sumar los recuentos para una o más secciones genómicas seleccionadas (por ejemplo, una primera sección genómica, una segunda sección genómica, una tercera sección genómica, cinco o más secciones genómicas, 50 o más secciones genómicas, 500 o más secciones genómicas, y similares). A veces, las secciones genómicas seleccionadas se eligen como referencia o base de comparación, debido a la presencia o ausencia de una o más variables o características. A veces, un recuento esperado se determina a partir de recuentos de una sección genómica (por ejemplo, una o más secciones genómicas, un cromosoma, genoma o parte de los mismos) que está libre de una variación genética (por ejemplo, una duplicación, delección, inserciones, una aneuploidía fetal, trisomía). En determinados casos, un recuento esperado se deriva de recuentos de una sección genómica (por ejemplo, una o más secciones genómicas, un cromosoma, genoma o parte de los mismos) que comprende una variación genética (por ejemplo, una duplicación, delección, inserciones, una aneuploidía fetal, trisomía). A veces, un recuento esperado se determina a partir de recuentos de una o más secciones genómicas en donde algunas de las secciones genómicas comprenden una variación genética y algunas de las secciones genómicas están sustancialmente libres de una variación genética. Un recuento esperado a menudo se determina usando datos (por ejemplo, recuentos de lecturas de secuencia mapeadas) de un grupo de muestras obtenidas bajo al menos una condición experimental común. A veces, un recuento esperado se determina aplicando a los recuentos una o más manipulaciones matemáticas o estadísticas descritas en el presente documento o conocidas de otro modo en la técnica. Ejemplos no limitantes de los valores de recuento esperado o valores de recuento de muestra esperado resultantes de dichas manipulaciones matemáticas o estadísticas incluyen mediana, media, moda, promedio y/o punto medio, mediana de desviación absoluta, una alternativa a la mediana de desviación absoluta tal como la introdujeron Rousseeuw y Croux, una estimación de remuestreo de tipo *bootstrapping*, similares y combinaciones de los mismos. En algunos casos, un recuento esperado es una mediana, moda, promedio y/o punto medio de recuentos (por ejemplo, recuentos de una sección genómica, cromosoma, genoma o parte del mismo). Un

recuento esperado a veces es una mediana, moda, promedio y/o punto medio o media de recuentos o recuentos de muestra. Los ejemplos no limitantes de recuentos y recuentos esperados incluyen recuentos filtrados, recuentos esperados filtrados, recuentos normalizados, recuentos esperados normalizados, recuentos ajustados y recuentos esperados ajustados. Los procesos de filtrado, normalización y ajuste se describen con mayor detalle en el presente documento.

En algunos casos, una derivada de un recuento esperado es un recuento esperado derivado de recuentos que se han normalizado y/o manipulado (por ejemplo, manipulado matemáticamente). Los recuentos que se han normalizado y/o manipulado (por ejemplo, manipulados matemáticamente) a veces se denominan derivada de recuentos. Una derivada de recuentos a veces es una representación de los recuentos de una primera sección genómica, cuya representación a menudo son los recuentos de una primera sección genómica en relación con (por ejemplo, divididos por) los recuentos de las secciones genómicas que incluyen la primera sección genómica. A veces, una derivada de recuentos se expresa como una representación porcentual o una representación de proporción. A veces, la representación es de una sección genómica a múltiples secciones genómicas, en donde las múltiples secciones genómicas son de todo o parte de un cromosoma. A veces, la representación es de múltiples secciones genómicas a un mayor número de secciones genómicas, en donde las múltiples secciones genómicas son de todo o parte de un cromosoma y la mayor cantidad de secciones genómicas es de múltiples cromosomas, sustancialmente todos los autosomas o sustancialmente el genoma completo. En algunos casos, un proceso de normalización que normaliza una derivada de recuentos incluye el uso de una derivada de un recuento esperado. En el presente documento, un recuento esperado obtenido a partir de una derivada de recuentos se denomina "derivada del recuento esperado". A veces, una derivada de un recuento esperado es un recuento esperado derivado de una representación de recuentos (por ejemplo, una representación porcentual, una representación cromosómica). En algunos casos, una derivada de un recuento esperado es una mediana, moda, promedio y/o punto medio de una representación de recuentos (por ejemplo, una representación porcentual, una representación cromosómica). En determinados casos, una mediana es una mediana, media, moda, punto medio, promedio o similares.

A veces, se determina una estimación de la variabilidad para recuentos, recuentos esperados o una derivada de un recuento esperado. Los ejemplos no limitativos de una estimación de la variabilidad incluyen una mediana de desviación absoluta (MAD) de los recuentos, recuentos esperados o derivada de los recuentos esperados; una alternativa a la MAD presentada por Rousseeuw y Croux; una estimación de remuestreo de tipo *bootstrapping*; una desviación estándar de los recuentos, recuentos esperados o derivada de los recuentos esperados; similares y combinaciones de los mismos. A veces se utiliza una estimación de la variabilidad en un proceso de normalización para obtener un recuento de muestra normalizado.

En determinados casos, un proceso de normalización para obtener un recuento de muestra normalizado incluye restar un recuento esperado de los recuentos de una primera sección genómica, generando así un valor de resta y dividiendo el valor de resta por una estimación de la variabilidad de los recuentos o recuentos esperados. Ejemplos no limitativos de la variabilidad de los recuentos o recuentos esperados son una mediana de desviación absoluta (MAD) de los recuentos o recuentos esperados, una alternativa a la MAD introducida por Rousseeuw y Croux o una estimación de remuestreo de tipo *bootstrapping*. En algunos casos, un proceso de normalización para obtener un recuento de muestra normalizado incluye restar la representación de recuentos de la primera sección genómica esperada de la representación de recuentos de la primera sección genómica, generando así un valor de resta, y dividir el valor de resta por una estimación de la variabilidad de la representación de recuentos de la primera sección genómica o la representación de recuentos de la primera sección genómica esperada. Ejemplos no limitativos de la variabilidad de la representación de recuento o la representación de recuento esperada son una mediana de desviación absoluta (MAD) de la representación de recuentos o la representación de recuentos esperado, una alternativa a MAD como introdujeron Rousseeuw y Croux o una estimación de remuestreo de tipo *bootstrapping*. En algunos casos, un recuento esperado es una mediana, moda, promedio, media y/o punto medio de los recuentos de la primera sección genómica y, a veces, una representación de recuentos esperados es una mediana, media, moda, promedio y/o punto medio de la representación de recuentos de primera sección genómica.

En algunos casos, un recuento esperado, una derivada de un recuento esperado (por ejemplo, una representación de recuentos esperados), o una estimación de la variabilidad de recuentos, una derivada de recuentos, un recuento esperado o derivada de un recuento esperado, se determina independientemente según los datos de muestra adquiridos bajo una o más condiciones experimentales comunes. A veces se obtiene una estimación de la variabilidad de los datos de muestra generados a partir de una o más condiciones experimentales comunes; a veces se obtiene una estimación de la variabilidad para datos de muestra no generados a partir de una o más condiciones experimentales comunes; a veces se obtiene un recuento esperado para datos de muestra generados a partir de una o más condiciones experimentales comunes; a veces se obtiene un recuento esperado para datos de muestra no generados a partir de una o más condiciones experimentales comunes; y a veces se obtiene una estimación de la variabilidad y un recuento esperado para datos de muestra generados a partir de una o más condiciones experimentales comunes. A veces se obtiene una estimación de la variabilidad de una derivada de un recuento esperado (por ejemplo, una representación de recuento esperado) para datos de muestra generados a partir de una o más condiciones experimentales comunes; a veces se obtiene una estimación de la variabilidad de una derivada de un recuento esperado (por ejemplo, una representación de recuento esperado) para datos de muestra no generados a partir de una o más condiciones experimentales comunes; a veces se obtiene una derivada de un recuento esperado

(por ejemplo, una representación de recuento esperado) para datos de muestra generados a partir de una o más condiciones experimentales comunes; a veces se obtiene una derivada de un recuento esperado (por ejemplo, una representación de recuento esperado) para datos de muestra no generados a partir de una o más condiciones experimentales comunes; y a veces se obtienen una estimación de la variabilidad de una derivada de un recuento esperado (por ejemplo, una representación de recuento esperado) y una derivada de un recuento esperado (por ejemplo, una representación de recuento esperada) para datos de muestra generados a partir de una o más condiciones experimentales comunes.

En algunos casos, un recuento esperado o una derivada de un recuento esperado (por ejemplo, una representación de recuento esperado) se determina usando datos de muestra adquiridos bajo una o más condiciones experimentales comunes, y una estimación de la variabilidad de recuentos, una derivada de recuentos, un recuento esperado o una derivada de un recuento esperado se determina usando datos de muestra no adquiridos bajo una condición experimental común. En determinados casos, se determina una estimación de la variabilidad de recuentos, una derivada de recuentos, un recuento esperado o la derivada de un recuento esperado usando datos de muestra adquiridos para un primer número de muestras, y no adquiridos bajo una condición experimental común, y se determina un recuento esperado o una derivada de un recuento esperado (por ejemplo, una representación de recuento esperado), usando datos de muestra adquiridos bajo una o más condiciones experimentales comunes y adquiridos para un segundo número de muestras menor que el primer número de muestras. El segundo número de muestras a veces se adquiere en un marco de tiempo más corto que el marco de tiempo en el que se adquirió el primer número de muestras.

Los datos de muestra adquiridos bajo una o más condiciones experimentales comunes a veces se adquieren bajo 1 a aproximadamente 5 condiciones experimentales comunes (por ejemplo, 1, 2, 3, 4 o 5 condiciones experimentales comunes). Los ejemplos no limitantes de condiciones experimentales comunes incluyen un canal en una celda de flujo, una unidad de celda de flujo, celdas de flujo comunes a un recipiente, celdas de flujo comunes a un lote o proceso de fabricación; una unidad de placas de reactivo, placas de reactivo comunes a un recipiente, placas de reactivo comunes a un lote o proceso de fabricación; un operador; un instrumento (por ejemplo, un instrumento de secuenciación); humedad, temperatura; índice de etiquetas de identificación; similares y combinaciones de los mismos. Las placas de reactivo a veces se utilizan para la preparación de bibliotecas de ácidos nucleicos y/o la secuenciación de ácidos nucleicos.

La cuantificación o el recuento de lecturas de secuencia se pueden realizar de cualquier manera adecuada, incluyendo, pero sin limitación, métodos de recuento manual y métodos de recuento automatizados. En algunos casos, se puede incorporar un método de recuento automatizado en un software que determina o cuenta el número de lecturas de secuencia o mapeo de etiquetas de secuencia en cada cromosoma y/o una o más secciones genómicas seleccionadas. El software generalmente son instrucciones de programas legibles por ordenador que, cuando son ejecutadas por un ordenador, realizan operaciones informáticas, como se describe en el presente documento.

El número de lecturas de secuencia mapeadas en cada bin y el número total de lecturas de secuencia para muestras derivadas del sujeto de prueba y/o sujetos de referencia pueden analizarse y procesarse adicionalmente para proporcionar un resultado determinante de la presencia o ausencia de una variación genética. Las lecturas de secuencia mapeadas que se han contado a veces se denominan “datos” o “conjuntos de datos”. En algunos casos, 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], basadas en condiciones experimentales [por ejemplo, basadas en índices, basadas en celdas de flujo, basadas en placas], similares y combinaciones de los mismos). En determinados casos, los datos o conjuntos de datos pueden organizarse y/o estratificarse en una matriz que tiene dos o más dimensiones en función de una o más características o variables (por ejemplo, fracción fetal y edad materna; fracción fetal y ubicación geográfica; porcentaje de representación del cromosoma 21 y número de celdas de flujo; puntuación z del cromosoma 21 y peso materno; puntuación z del cromosoma 21 y edad gestacional, y similares). Los datos organizados y/o estratificados en matrices pueden organizarse y/o estratificarse 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. Los ejemplos no limitantes de datos estratificados usando características o variables se presentan en las Figs. 4 a 45. En determinados casos, los conjuntos de datos caracterizados por una o más características o variables a veces se procesan después del recuento.

Elevaciones

En algunos casos, 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). 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 algunos casos, una elevación se basa en recuentos que están sin procesar. Algunas veces, una elevación se asocia con un valor de incertidumbre. Una elevación para una sección genómica a veces se denomina “elevación de sección genómica” y 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, moda, punto medio o promedio de recuentos, una elevación para un conjunto de secciones genómicas que comprenden recuentos normalizados, similares o cualquier combinación de los mismos. En algunos casos, una primera elevación y una segunda elevación en un perfil se derivan de recuentos de secciones genómicas mapeadas en el mismo cromosoma. En algunos casos, una primera elevación y una segunda elevación en un perfil se derivan de recuentos de secciones genómicas mapeadas en cromosomas diferentes.

En algunos casos, una elevación se determina a partir de recuentos normalizados o no normalizados mapeados en una o más secciones genómicas. En algunos casos, una elevación se determina a partir de recuentos normalizados o no normalizados mapeados en dos o más secciones genómicas, en donde 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 algunos casos, los recuentos de 2 a aproximadamente 100.000 secciones genómicas determinan una elevación. En algunos casos, 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 algunos casos, los recuentos de aproximadamente 10 a aproximadamente 50 secciones genómicas determinan una elevación. En algunos casos, los recuentos de aproximadamente 20 a aproximadamente 40 o más secciones genómicas determinan una elevación. En algunos casos, 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 algunos casos, 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 algunos casos, 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 algunos casos, 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 algunos casos, 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 algunos casos, 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 algunos casos, 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 algunos casos, 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, mediana de elevación, de moda o punto medio se deriva de una segunda elevación en donde 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 donde 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 algunos casos, un valor de una segunda elevación está más próximo al valor medio, promedio, de moda, punto medio o valor de mediana de un perfil de recuento para un cromosoma, o segmento del mismo, que la primera elevación. En algunos casos, una segunda elevación es una elevación media de un cromosoma, una porción de un cromosoma o un segmento de los mismos. En algunos casos, 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 algunos casos, 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, delección, 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 incluye 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.

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 algunos casos, los datos de lectura de secuencia en un conjunto de datos pueden ajustarse y/o procesarse adicionalmente (por ejemplo, manipularse matemática y/o estadísticamente) y/o visualizarse para facilitar la provisión de un resultado. Los datos de lectura de secuencia ajustados a menudo son resultado de la manipulación de una porción de o todas las lecturas de secuencia, datos en un conjunto de datos y/o ácido nucleico de muestra. Se puede usar cualquier manipulación adecuada para ajustar una parte de o todas las lecturas de secuencia, datos en un conjunto de datos y/o ácido nucleico de muestra. En algunos casos, un ajuste a las lecturas de secuencia, datos en un conjunto de datos y/o ácido nucleico de muestra es un proceso elegido de filtrado (por ejemplo, eliminar una porción de los datos en función de una característica o variable seleccionadas; eliminar secuencias repetitivas, eliminar bins no informativos o bins que tienen una mediana de recuentos cero, por ejemplo), ajustar (por ejemplo, volver a escalar y/o volver a ponderar una porción de o todos los datos en función de un estimador; volver a ponderar los recuentos de muestras en función del contenido de G/C, volver a escalar y/o volver a ponderar una porción de o todos los datos en función de la fracción fetal, por ejemplo), normalizar usando uno o más estimadores o manipulaciones estadísticas (por ejemplo, normalizar todos los datos en una celda de flujo dada con respecto a la mediana de desviación absoluta de todos los datos en la celda de flujo), y similares. En algunos casos, el estimador es un estimador robusto. En determinados casos, se ajusta y/o se procesa una porción de los datos de lectura de secuencia y, en algunos casos, se ajustan y/o procesan todos los datos de lectura de secuencia.

Las lecturas de secuencia ajustadas o procesadas, los datos en un conjunto de datos y/o el ácido nucleico de muestra a veces se denominan derivada (por ejemplo, una derivada de los recuentos, datos de derivada, derivada de las lecturas de secuencia y similares). Una derivada de recuentos, datos o lecturas de secuencia a menudo se genera mediante el uso de una o más manipulaciones matemáticas y/o estadísticas en los recuentos, datos o lecturas de secuencia. Cualquier manipulación matemática y/o estadística adecuada descrita en el presente documento o conocida en la técnica puede usarse para generar recuentos de derivada, datos o lecturas de secuencia. Los ejemplos no limitativos de manipulaciones matemáticas y/o estadísticas que se pueden utilizar para filtrar, ajustar, normalizar o manipular recuentos, datos o secuencias de lectura para generar una derivada incluyen, promedio, media, mediana, moda, punto medio, mediana de desviación absoluta, alternativa a la mediana de desviación absoluta como introdujeron Rousseeuw y Croux, estimación de remuestreo de tipo *bootstrapping*, otros métodos descritos en el presente documento y conocidos en la técnica, similares o combinaciones de los mismos.

En determinados casos, 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 genómicas o bins 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 [por ejemplo, secuencias de G/C], secuencias repetitivas). 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, (v) reducir la complejidad de conjuntos de datos más grandes, (v) reducir o eliminar la variabilidad de datos inducida por condiciones experimentales, (vi) volver a escalar y/o volver a ponderar una porción de o todos los datos en un conjunto de datos, y/o (vii) 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 algunos casos.

A menudo los datos con ruido son (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 determinados casos, 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.

A menudo, los datos no informativos, bins no informativos y secciones genómicas no informativas son secciones genómicas o datos derivados de las mismas que tienen un valor numérico que es significativamente diferente de un valor umbral de punto de corte predeterminado o se encuentran fuera de un rango de valores de punto de corte predeterminado. A menudo, se calcula un valor umbral de punto de corte o rango de valores al manipular matemática y/o estadísticamente los datos leídos de secuencias (por ejemplo, a partir de una referencia, sujeto, celda de flujo y/o placa), en algunos casos, y en determinados casos, los datos leídos de secuencias manipulados para generar un valor de punto de corte umbral o rango de valores son datos leídos de secuencias (por ejemplo, a partir de una referencia, sujeto, celda de flujo y/o placa). En algunos casos, se obtiene un valor de punto de corte de umbral al calcular la desviación estándar y/o la mediana de desviación absoluta (por ejemplo, MAD o alternativa a MAD como introdujeron Rousseeuw y Croux, o estimación de remuestreo de tipo *bootstrapping*) de un perfil de recuento sin procesar o normalizado y multiplicando la desviación estándar para el perfil por una constante que representa el número de desviaciones estándar elegidas como umbral de punto de corte (por ejemplo, multiplicar por 3 para 3 desviaciones estándar), por lo que se genera un valor para una incertidumbre. En determinados casos, una porción o la totalidad de las secciones genómicas que exceden el valor de punto de corte de umbral de incertidumbre calculado, o que están fuera del rango de los valores de punto de corte de umbral, se eliminan como parte del proceso de normalización, antes o después de este. En algunos casos, una porción o la totalidad de las secciones genómicas que exceden el valor de punto de corte de umbral de incertidumbre calculado, o que están fuera del rango de los valores de punto de corte de umbral o los puntos de datos sin procesar, se ponderan como parte del proceso de normalización o clasificación o antes del mismo. En el presente documento se describen ejemplos de ponderación. En algunos casos, datos redundantes, y lecturas mapeadas redundantes 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.

Condiciones experimentales

Las muestras a veces se ven afectadas por condiciones experimentales comunes. Las muestras procesadas sustancialmente al mismo tiempo o usando sustancialmente las mismas condiciones y/o reactivos a veces muestran una variabilidad de datos inducida por condiciones experimentales similares (por ejemplo, condiciones experimentales comunes) cuando se comparan con otras muestras procesadas en un momento diferente y/o al mismo tiempo usando

diferentes condiciones y/o reactivos. A menudo, existen consideraciones prácticas que limitan el número de muestras que se pueden preparar, procesar y/o analizar en un momento dado durante un procedimiento experimental. En determinados casos, el marco de tiempo para procesar una muestra desde la materia prima hasta generar un resultado a veces es de días, semanas o incluso meses. Debido al tiempo que transcurre entre el aislamiento y el análisis final, los experimentos de alto rendimiento que analizan un gran número de muestras a veces generan efectos por lotes o variabilidad de datos inducida por condiciones experimentales. La variabilidad de datos inducida por condiciones experimentales a menudo incluye cualquier variabilidad de datos que sea el resultado del aislamiento, almacenamiento, preparación y/o análisis de muestras. Los ejemplos no limitantes de la variabilidad inducida por condiciones experimentales incluyen la variabilidad basada en celdas de flujo y/o la variabilidad basada en placas que incluye: sobrerrepresentación o subrepresentación de secuencias; datos con ruido; puntos de datos falsos o atípicos, efectos de reactivos, efectos del personal, efectos de las condiciones de laboratorio y similares. La variabilidad inducida por condiciones experimentales a veces se produce en subpoblaciones de muestras en un conjunto de datos (por ejemplo, efecto de lote). A menudo, un lote consiste en muestras procesadas usando sustancialmente los mismos reactivos, muestras procesadas en la misma placa de preparación de muestras (por ejemplo, placa de micropocillos usada para la preparación de muestras; aislamiento de ácido nucleico, por ejemplo), muestras clasificadas para análisis en la misma placa de clasificación (por ejemplo, placa de micropocillos usada para organizar las muestras antes de cargarlas en una celda de flujo), muestras procesadas prácticamente al mismo tiempo, muestras procesadas por el mismo personal, y/o muestras procesadas bajo sustancialmente las mismas condiciones experimentales (por ejemplo, temperatura, niveles de CO₂, niveles de ozono, similares o combinaciones de los mismos). Los efectos del lote de condiciones experimentales a veces afectan a las muestras analizadas en la misma celda de flujo, preparadas en la misma placa de reactivo o placa de micropocillos y/o preparadas para el análisis (por ejemplo, preparación de una biblioteca de ácidos nucleicos para la secuenciación) en la misma placa de reactivo o placa de micropocillos. Las fuentes adicionales de variabilidad pueden incluir la calidad del ácido nucleico aislado, la cantidad de ácido nucleico aislado, el tiempo de almacenamiento después del aislamiento del ácido nucleico, el tiempo de almacenamiento, la temperatura de almacenamiento, similares y combinaciones de los mismos. La variabilidad de los puntos de datos en un lote (por ejemplo, subpoblación de muestras en un conjunto de datos que se procesan al mismo tiempo y/o que usan los mismos reactivos y/o condiciones experimentales) a veces es mayor que la variabilidad de los puntos de datos observados entre lotes. Esta variabilidad de datos a veces incluye datos falsos o atípicos cuya magnitud puede afectar a la interpretación de algunos o todos los demás datos en un conjunto de datos. Una porción o la totalidad de un conjunto de datos se puede ajustar para las condiciones experimentales usando las etapas de procesamiento de datos descritas en el presente documento y conocidas en la técnica; normalización con respecto a la mediana de desviación absoluta calculada para todas las muestras analizadas en una celda de flujo o procesadas en una placa de micropocillos, por ejemplo.

La variabilidad inducida por las condiciones experimentales se puede observar para determinar los datos obtenidos durante un período de semanas a meses o años (por ejemplo, 1 semana, 1-4 semanas, 1 mes, 1-3 meses, 1-6 meses). A veces se llevan a cabo varios experimentos durante un período de semanas a meses en donde una o más condiciones experimentales son condiciones experimentales comunes. Los ejemplos no limitativos de condiciones experimentales comunes incluyen el uso del mismo instrumento, máquina o parte de los mismos (por ejemplo, un secuenciador, un dispositivo de manipulación de líquidos, un espectrofotómetro, fotocélula, etc.), el uso del mismo dispositivo (por ejemplo, celda de flujo, canal de celda de flujo, placa, chip, similares o parte de los mismos), uso del mismo protocolo (procedimiento operativo, procedimiento operativo estándar, receta, métodos y/o condiciones (por ejemplo, tiempo de incubación, temperatura, presión, humedad, volumen, concentración), el mismo operador (por ejemplo, un técnico, científico) y los mismos reactivos (por ejemplo, nucleótidos, oligonucleótidos, etiqueta de secuencia, índice de etiquetas de identificación, muestra (por ejemplo, muestra de ADNcl), proteínas (por ejemplo, enzimas, tampones, sales, agua), similares).

El uso del mismo dispositivo, aparato o reactivo puede incluir un dispositivo, aparato, reactivo o parte de los mismos del mismo fabricante, el mismo ciclo de fabricación, el mismo lote (por ejemplo, un material procedente de la misma planta, fabricante, ciclo de fabricación o ubicación, una colección etiquetada con la misma fecha), el mismo ciclo de limpieza, el mismo protocolo de preparación, el mismo recipiente (bolsa, caja, paquete, contenedor de almacenamiento, palé, remolque), el mismo envío (por ejemplo, la misma fecha de entrega, el mismo pedido, tener la misma factura), la misma planta de fabricación, la misma línea de montaje, similares o combinaciones de los mismos. El uso del mismo operador, en algunos casos, significa que uno o más operadores de una máquina, aparato o dispositivo son los mismos.

Ajustar los datos en un conjunto de datos a menudo puede reducir o eliminar el efecto de los valores atípicos en un conjunto de datos, volver a escalar o volver a ponderar los datos para facilitar el suministro de un resultado y/o reducir la complejidad y/o la dimensionalidad de un conjunto de datos. En determinados casos, los datos se pueden clasificar (por ejemplo, estratificar, organizar) según una o más condiciones experimentales comunes (por ejemplo, reactivos usados, celda de flujo usada, placa usada, personal que procesó las muestras, secuencias indexadoras usadas, similares o combinaciones de los mismos). En algunos casos, los datos pueden normalizarse o ajustarse según una o más condiciones experimentales comunes.

Los datos se pueden volver a escalar o volver a ponderar usando estimadores robustos. Un estimador robusto a menudo es una manipulación matemática o estadística que minimiza o elimina el efecto de datos falsos o atípicos,

cuya magnitud puede afectar al suministro de un resultado (por ejemplo, hacer una determinación de la presencia o ausencia de una variación genética). Se puede usar cualquier estimador robusto adecuado para ajustar un conjunto de datos. En algunos casos, un estimador robusto es un estimador robusto de escala (por ejemplo, variabilidad; similar y/o incluye la mediana de desviación absoluta [MAD] o una alternativa a la MAD tal introdujeron Rousseeuw y Croux, o una estimación de remuestreo de tipo *bootstrapping*), y en determinados casos, un estimador robusto es un estimador robusto de ubicación (por ejemplo, un valor esperado; similar a un promedio o mediana). Los ejemplos no limitativos de estimadores robustos de escala y ubicación se describen en el ejemplo 2 y también son conocidos en la técnica (por ejemplo, mediana, ANOVA y similares). En algunos casos, una porción o todos los datos de un conjunto de datos se pueden ajustar usando un recuento esperado o la derivada de un recuento esperado obtenido mediante un estimador robusto. En algunos casos, un recuento esperado es un recuento derivado de una referencia o una muestra de referencia (por ejemplo, una muestra euploide conocida).

Puede utilizarse cualquier procedimiento adecuado para ajustar y/o procesar conjuntos de datos descritos en el presente documento. Los ejemplos no limitativos de procedimientos que se pueden usar para ajustar conjuntos de datos incluyen ajustes basados en las condiciones experimentales (por ejemplo, normalización basada en placas, normalización basada en celdas de flujo [por ejemplo, comparaciones de medianas basadas en celdas de flujo], ajuste de enmascaramiento de repetición (por ejemplo, eliminación de secuencias repetitivas); ajuste de contenido de G/C; ajuste de regresión de polinomial ponderada localmente (por ejemplo, LOESS), normalización mediante estimadores robustos (por ejemplo, estimación de ubicación [por ejemplo, valor esperado; similar al promedio], estimación de escala [por ejemplo, variabilidad]; y análisis de variabilidad [por ejemplo, ANOVA]). Además, en determinados casos, los conjuntos de datos se pueden procesar adicionalmente utilizando uno o más de los siguientes métodos de procesamiento de datos: filtrado, normalización, ponderación, monitorización de alturas de pico, monitorización de áreas de pico, monitorización de bordes de pico, determinación de relaciones de área, procesamiento matemático de datos, procesamiento estadístico de datos, aplicación de algoritmos estadísticos, análisis con variables fijas, análisis con variables optimizadas, representación de datos para identificar patrones o tendencias para procesamiento adicional, procesamiento de ventana deslizante (por ejemplo, normalización de ventana deslizante), procesamiento de ventana estática (por ejemplo, normalización de ventana estática), similares y combinaciones de los anteriores, y en determinados casos, se puede aplicar un método de procesamiento a un conjunto de datos antes de una etapa de ajuste. En algunos casos, los conjuntos de datos se ajustan y/o procesan basándose en diversas características (por ejemplo, contenido de GC, lecturas mapeadas redundantes, regiones de centrómero, regiones de telómero, secuencias repetitivas, 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 determinados casos, 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 algunos casos, los conjuntos de datos pueden incluir de miles a millones de lecturas de secuencia para cada sujeto de prueba y/o referencia.

El ajuste y/o procesamiento de datos se puede realizar en cualquier número de etapas, en determinados casos, y en aquellos casos con más de una etapa, y las etapas se pueden realizar en cualquier orden. Por ejemplo, los datos pueden ajustarse y/o procesarse usando solo un único procedimiento de ajuste/procesamiento en algunos casos, y en determinados casos, los datos pueden ajustarse/procesarse usando 1 o más, 5 o más, 10 o más o 20 o más etapas de ajuste/procesamiento (por ejemplo, 1 o más etapas de ajuste/procesamiento, 2 o más etapas de ajuste/procesamiento, 3 o más etapas de ajuste/procesamiento, 4 o más etapas de ajuste/procesamiento, 5 o más etapas de ajuste/procesamiento, 6 o más etapas de ajuste/procesamiento, 7 o más etapas de ajuste/procesamiento, 8 o más etapas de ajuste/procesamiento, 9 o más etapas de ajuste/procesamiento, 10 o más etapas de ajuste/procesamiento, 11 o más etapas de ajuste/procesamiento, 12 o más etapas de ajuste/procesamiento, 13 o más etapas de ajuste/procesamiento, 14 o más etapas de ajuste/procesamiento, 15 o más etapas de ajuste/procesamiento, 16 o más etapas de ajuste/procesamiento, 17 o más etapas de ajuste/procesamiento, 18 o más etapas de ajuste/procesamiento, 19 o más etapas de ajuste/procesamiento, o 20 o más etapas de ajuste/procesamiento). En algunos casos, las etapas de ajuste/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 determinados casos, las etapas de ajuste/procesamiento pueden ser dos o más etapas de ajuste/procesamiento diferentes (por ejemplo, enmascaramiento de repetición, normalización basada en celdas de flujo; ajuste de contenido de G/C basado en bins, normalización basada en celdas de flujo; enmascaramiento de repetición, ajuste de contenido de G/C basado en bins, normalización basada en placas; 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), realizadas de forma simultánea o secuencial. En algunos casos, puede usarse cualquier número y/o combinación adecuada de las mismas o diferentes etapas de ajuste/procesamiento para procesar datos de lectura de secuencia para facilitar la provisión de un resultado. En determinados casos, el ajuste/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 algunos casos, una o más etapas de ajuste/procesamiento pueden comprender el ajuste para una o más condiciones experimentales descritas en el presente documento. Los ejemplos no limitantes de condiciones

experimentales que a veces conducen a la variabilidad de los datos incluyen: sobrerepresentación o subrepresentación de secuencias (por ejemplo, variabilidad basada en amplificación sesgada); datos con ruido; puntos de datos falsos o atípicos; variabilidad basada en celdas de flujo (por ejemplo, variabilidad observada en muestras analizadas en una celda de flujo, pero no observada en otras celdas de flujo utilizadas para analizar muestras del mismo lote (por ejemplo, preparadas en la misma placa de reactivo o placa de micropocillos)); y/o variabilidad basada en placas (por ejemplo, variabilidad observada en algunas o todas las muestras preparadas en la misma placa de reactivo o placa de micropocillos y/o clasificadas para el análisis en la misma placa de micropocillos independientemente de la celda de flujo usada para el análisis).

En algunos casos, el porcentaje de representación se calcula para una sección genómica (por ejemplo, una sección genómica, cromosoma, genoma o parte del mismo). En algunos casos, una representación porcentual se determina como un número de recuentos mapeados en una sección genómica normalizada (por ejemplo, dividida por) con respecto al número de recuentos mapeados en múltiples secciones genómicas. A veces, la determinación de una representación porcentual excluye secciones genómicas y/o recuentos derivados de cromosomas sexuales (por ejemplo, cromosomas X y/o Y). A veces, la determinación de una representación porcentual incluye solo secciones genómicas y/o recuentos derivados de autosomas. A veces, la determinación de una representación porcentual incluye secciones genómicas y/o recuentos derivados de autosomas y cromosomas sexuales. Por ejemplo, $perc_i$ representa la representación porcentual de una sección genómica i seleccionada,

$$perc_i = \frac{recuentos_i}{\sum_{j=1}^{22} recuentos_j}$$

en donde los $recuentos_i$ son recuentos de lecturas mapeadas en la sección genómica seleccionada i y los $recuentos_j$ son el número de recuentos de lecturas mapeadas en múltiples secciones genómicas j (por ejemplo, múltiples secciones genómicas en el cromosoma j , secciones genómicas de todos los autosomas j , secciones genómicas del genoma j). Por ejemplo, chr_i representa la representación cromosómica del cromosoma i ,

$$chr_i = \frac{recuentos_i}{\sum_{j=1}^{22} recuentos_j}$$

en donde los $recuentos_j$ son el número de lecturas alineadas en el cromosoma j . En algunos casos, una representación porcentual es una “representación de recuento de secciones genómicas”. A veces, una representación porcentual es una “representación de sección genómica” o una “representación cromosómica”.

Una o más etapas de ajuste/procesamiento pueden comprender el ajuste para la variabilidad inducida por las condiciones experimentales. La variabilidad se puede ajustar usando un estimador robusto de escala y/o ubicación. En algunos casos, las puntuaciones z se pueden ajustar para la variabilidad inducida por las condiciones experimentales al determinar (1) la representación porcentual de una sección genómica seleccionada (por ejemplo, una representación de recuentos de la primera sección genómica; cromosoma, cromosoma 21, por ejemplo), (2) la mediana, media, moda, promedio y/o punto medio de todos los valores de representación porcentual para una sección genómica seleccionada, (3) la mediana de desviación absoluta (MAD) de todos los valores de representación porcentual, y ajuste de la puntuación z usando un estimador robusto basado en celdas de flujo que minimiza o elimina el efecto de los valores atípicos. En determinados casos, se calcula un ajuste de puntuación z robusta basada en celdas de flujo para un cromosoma diana, una región genómica diana o una sección genómica diana (por ejemplo, el cromosoma 21) usando la fórmula a continuación.

$$Z \text{ robusta} = \frac{perc_i - \text{Mediana}(\{perc_{i \text{ ec}}\})}{\text{MAD}(\{perc_{i \text{ ec}}\})}$$

La fórmula, tal como está escrita, está configurada para calcular una puntuación z robusta para una sección genómica, en donde **perc** es una representación porcentual (por ejemplo, representación de recuento de la primera sección genómica, representación cromosómica) de una sección genómica seleccionada i (por ejemplo, cualquier sección genómica adecuada, cromosoma, genoma o parte del mismo). En algunos casos, la mediana se calcula a partir de uno o más valores porcentuales de representación para la sección genómica seleccionada i obtenida para las condiciones experimentales **ec**. Se calcula una MAD a partir de uno o más valores porcentuales de representación para la sección genómica seleccionada i obtenida para las condiciones experimentales **ec**. La fórmula generalizada se puede utilizar para obtener puntuaciones z robustas para cualquier sección genómica sustituyendo los valores equivalentes para la sección genómica diana elegida en determinados casos. En algunos casos, se calcula una mediana, media, moda, promedio, punto medio y/o MAD para un conjunto seleccionado de muestras o un subconjunto de muestras. A veces se calcula una mediana y/o MAD para el mismo conjunto de muestras. En algunos casos, se calcula una mediana y/o MAD para un conjunto diferente de muestras. En algunos casos, las condiciones experimentales **ec** son las mismas. En algunos casos, las condiciones experimentales **ec** pueden comprender o consistir en una o más condiciones

experimentales comunes. En algunos casos, las condiciones experimentales **ec** son diferentes. En algunos casos, las condiciones experimentales **ec'** son las mismas. En algunos casos, las condiciones experimentales **ec'** pueden comprender o consistir en una o más condiciones experimentales comunes. En algunos casos, las condiciones experimentales **ec'** son diferentes. A veces, las condiciones experimentales **ec** y **ec'** son diferentes. En algunos casos, las condiciones experimentales **ec** y **ec'** pueden comprender o consistir en una o más condiciones experimentales comunes. Por ejemplo, se puede calcular una puntuación z robusta para una sección genómica seleccionada a partir de (a) una media derivada de un conjunto seleccionado de datos recopilados de un conjunto seleccionado de muestras y en donde los datos se obtuvieron bajo una o más condiciones experimentales comunes (por ejemplo, de la misma celda de flujo), y (b) una MAD derivada de otro conjunto seleccionado de datos recopilados de otro conjunto seleccionado de muestras y en donde los datos se obtuvieron bajo una o más condiciones experimentales comunes (por ejemplo, usando diferentes celdas de flujo y el mismo lote de reactivos seleccionados). En algunos casos, una media y una MAD se derivan de datos que comparten al menos una condición experimental común. A veces, una media y una MAD se derivan de datos que no comparten una condición experimental común.

En algunos casos, se obtiene un recuento de muestra normalizado (por ejemplo, una puntuación z) mediante un proceso que comprende restar un recuento esperado (por ejemplo, una mediana de recuentos, una mediana de representaciones porcentuales) de los recuentos de una primera sección genómica (por ejemplo, recuentos, una representación porcentual), generando así un valor de resta y dividiendo el valor de resta por una estimación de la variabilidad del recuento (por ejemplo, una MAD, una MAD de recuentos, una MAD de representaciones porcentuales). En algunos casos, un recuento esperado (por ejemplo, una mediana de recuentos, una mediana de representaciones porcentuales) y una estimación de la variabilidad del recuento (por ejemplo, una MAD, una MAD de recuentos, una MAD de representaciones porcentuales) se derivan de datos que comparten al menos una condición experimental común. A veces, un recuento esperado (por ejemplo, una mediana de recuentos, una mediana de representaciones porcentuales) y una estimación de la variabilidad del recuento (por ejemplo, una MAD, una MAD de recuentos, una MAD de representaciones porcentuales) se derivan de datos que no comparten una condición experimental común. En algunos casos, una mediana puede ser una mediana, media, moda, promedio y/o punto medio.

En determinados casos, una o más etapas de ajuste/procesamiento pueden comprender ajustes para la variabilidad basada en celdas de flujo. La variabilidad basada en celdas de flujo se puede ajustar usando un estimador robusto de escala y/o ubicación. En algunos casos, las puntuaciones z se pueden ajustar para la variabilidad basada en células de flujo determinando (1) la representación porcentual de un cromosoma seleccionado (por ejemplo, una representación de recuento de la primera sección genómica; el cromosoma 21, por ejemplo), (2) la mediana de todos los valores de representación cromosómica observados en una celda de flujo, (3) la mediana de desviación absoluta de todos los valores de representación cromosómica observados en una celda de flujo, y ajustando la puntuación z usando un estimador robusto basado en celdas de flujo que minimiza o elimina el efecto de los valores atípicos. En determinados casos, se calcula un ajuste de puntuación z robusta basada en celdas de flujo para un cromosoma diana, una región genómica diana o una sección genómica diana (por ejemplo, el cromosoma 21) usando la fórmula a continuación.

$$Z_{robusta_{FC}} = \frac{\text{perc.chr21} - \text{Mediana}(\{\text{perc.chr21}\})_{FC}}{\text{MAD}(\{\text{perc.chr21}\})_{FC}}$$

La fórmula, tal como está escrita, está configurada para calcular una puntuación z robusta para el cromosoma 21, en donde perc.chr21 es el porcentaje de representación del cromosoma 21 (por ejemplo, la representación del recuento de la primera sección genómica), MAD representa la mediana de desviación absoluta y FC representa la celda de flujo. La fórmula generalizada se puede utilizar para obtener puntuaciones z robustas para cualquier cromosoma sustituyendo los valores equivalentes por el cromosoma diana elegido, la región genómica diana o la sección genómica diana, en donde se designa la referencia del cromosoma 21 (por ejemplo, cr21), en determinados casos.

En algunos casos, una o más etapas de ajuste/procesamiento pueden comprender ajustes para la variabilidad basada en placas. La variabilidad basada en placas se puede ajustar usando un estimador robusto de escala y/o ubicación. En determinados casos, las puntuaciones z se pueden ajustar para la variabilidad basada en placas determinando (1) la representación porcentual de un cromosoma seleccionado (por ejemplo, una representación de recuento de la primera sección genómica; el cromosoma 21, por ejemplo), (2) la mediana de todos los valores de representación cromosómica observados en una o más placas, (3) la mediana de desviación absoluta de todos los valores de representación cromosómica observados en una o más placas, y ajustando la puntuación z usando un estimador robusto basado en placas que minimiza o elimina el efecto de los valores atípicos. En determinados casos, se calcula un ajuste de puntuación z robusta basada en placas para un cromosoma diana, una región genómica diana o una sección genómica diana (por ejemplo, el cromosoma 21) usando la fórmula a continuación.

$$Z_{robusta_{PLACA}} = \frac{\text{perc.chr21} - \text{Mediana}(\{\text{perc.chr21}\})_{PLACA}}{\text{MAD}(\{\text{perc.chr21}\})_{PLACA}}$$

La fórmula, tal como está escrita, está configurada para calcular una puntuación *z* robusta para el cromosoma 21, en donde *perc.chr21* es el porcentaje de representación del cromosoma 21 (por ejemplo, la representación del recuento de la primera sección genómica), *MAD* representa la mediana de desviación absoluta y *PLACA* representa una o más placas de muestras (por ejemplo, placa o placas de reactivo, placa o placas de preparación de muestras, placa o placas de clasificación). La fórmula generalizada se puede utilizar para obtener puntuaciones *z* robustas para cualquier cromosoma sustituyendo los valores equivalentes por el cromosoma diana elegido, la región genómica diana o la sección genómica diana, en donde se designa la referencia del cromosoma 21 (por ejemplo, *cr21*), en determinados casos.

La mediana de desviación absoluta (*MAD*) a veces se calcula usando la fórmula:

$$MAD = 1,4826 * Mediana(\{|X - Mediana(\{X\})|\})$$

donde, *X* representa cualquier variable aleatoria para la que se calcula la mediana de desviación absoluta, y la constante de normalización 1,4826 representa $1/\text{Inv}[\Phi](3/4)$ y en donde Φ es la función de distribución acumulativa para la distribución Gaussiana estándar (por ejemplo, normal), y $\text{Inv}[\Phi]$ es su inversa (por ejemplo, relacionada con una función cuantil). $\text{Inv}[\Phi]$ se evalúa en $X = \%$ y es igual a $1/1,4826$. En “R code”, la ecuación para calcular la constante de normalización es: $1/\text{qnorm}(.3/4) = 1,4826$. “R code” es un lenguaje de programación de código abierto no patentado que se usa para una diversidad de análisis estadísticos sustancialmente similar al lenguaje de programación S (por ejemplo, R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL www.R-project.org/). La constante de normalización 1,4826 se elige de modo que la mediana de desviación absoluta (por ejemplo, *MAD*) de los datos normalmente distribuidos sea igual, para muestras grandes, a la desviación estándar (por ejemplo, *STDEV*) de los mismos datos, lo que efectivamente pone la *MAD* y *STDEV* en la misma escala. A menudo se utiliza una función cuantil para prescribir una distribución de probabilidad. Una función cuantil de una distribución de probabilidad es la inversa de su integral y, a menudo, especifica el valor en el que la variable aleatoria estará, o por debajo, para una probabilidad dada, en algunos casos.

En determinados casos, una o más etapas de ajuste/procesamiento pueden comprender el ajuste sobrerrepresentación o subrepresentación de secuencias. Como se indica en el presente documento, los procedimientos de amplificación utilizados en algunas etapas de preparación y/o secuenciación a veces generan una sobrerrepresentación o subrepresentación de las secuencias debido al contenido y/o la estructura de la secuencia. La sobrerrepresentación o la subrepresentación de las secuencias a veces reduce la confianza en un resultado proporcionado. El efecto de sobrerrepresentación o la subrepresentación de secuencias se puede minimizar o eliminar ajustando o normalizando una porción o la totalidad de un conjunto de datos con referencia a un valor esperado usando un estimador robusto, en determinados casos. En algunos casos, se calcula un valor esperado para una porción o la totalidad de los cromosomas usando uno o más estimadores elegidos de; un promedio, una mediana, promedio, punto medio, moda, una median absolute deviation (mediana de desviación absoluta - *MAD*), una alternativa a la *MAD* introducida por Rousseeuw y Croux, estimación de remuestreo de tipo *bootstrapping*, desviaciones estándar, puntuaciones *z*, puntuación *z* robusta, ANOVA, análisis de regresión LOESS (por ejemplo, suavizado LOESS, suavizado LOWESS) y similares. Ajustar una porción o la totalidad de un conjunto de datos para reducir o eliminar el efecto de la sobrerrepresentación o subrepresentación de las secuencias puede facilitar la obtención de un resultado y/o reducir la complejidad y/o la dimensionalidad de un conjunto de datos.

Una o más etapas de ajuste/procesamiento pueden comprender el ajuste del contenido de G/C. Como se indica en el presente documento, las secuencias con un alto contenido de G/C a veces se sobrerrepresentan o subrepresentan en un conjunto de datos sin procesar o procesados. En determinados casos, el contenido de G/C para una porción o la totalidad de un conjunto de datos (por ejemplo, bins seleccionados, porciones seleccionadas de cromosomas, cromosomas seleccionados) se ajusta para minimizar o eliminar el sesgo del contenido de G/C ajustando o normalizando una porción o la totalidad de un conjunto de datos con referencia a un valor esperado usando un estimador robusto. En algunos casos, el valor esperado es el contenido de G/C de las lecturas de secuencia de nucleótidos y, en determinados casos, el valor esperado es el contenido de G/C del ácido nucleico de muestra. En algunos casos, se calcula el valor esperado para una porción o la totalidad de los cromosomas usando uno o más estimadores elegidos de; un promedio, mediana, media, moda, punto medio, mediana de desviación absoluta (*MAD*), una alternativa a la *MAD* introducida por Rousseeuw y Croux, estimación de remuestreo de tipo *bootstrapping*, desviación estándar, puntuación *z*, puntuación *z* robusta, ANOVA, análisis de regresión LOESS (por ejemplo, suavizado LOESS, suavizado LOWESS) y similares. Ajustar una porción o la totalidad de un conjunto de datos para reducir o eliminar el efecto del sesgo del contenido de G/C puede facilitar la obtención de un resultado y/o reducir la complejidad y/o la dimensionalidad de un conjunto de datos, en algunos casos.

PERUN

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 determinados casos, 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 determinados casos (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 determinados casos, 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 donde 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 casos, 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 algunos casos, 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 determinados casos, 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 casos, puede usarse una elevación umbral predeterminada para seleccionar tales muestras para la normalización de GC adicional.

En determinados casos, 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 algunos casos, 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 algunos casos, 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 algunos casos, 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 algunos casos, 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 algunos casos, 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 algunos casos. 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 recuento esperado, 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 determinados casos.

Otro procesamiento de datos

En determinados casos, una o más etapas de ajuste/procesamiento pueden comprender el ajuste para secuencias repetitivas. Como se indica en el presente documento, las secuencias repetitivas a menudo son datos poco informativos y/o pueden contribuir a datos con ruido, lo que a veces reduce la confianza en un resultado proporcionado. Se puede usar cualquier método adecuado para reducir el efecto de secuencias repetitivas (por ejemplo, eliminación de secuencias repetitivas) descrito en el presente documento o conocido en la técnica. Se pueden encontrar ejemplos no limitativos de recursos disponibles para la eliminación de secuencias repetitivas en las siguientes publicaciones: URL www.repeatmasker.org/papers.html y www.biomedcentral.com/1471-2105/11/80. El efecto de la presencia de secuencias repetitivas en un resultado proporcionado se puede minimizar o eliminar ajustando o normalizando una porción o la totalidad de un conjunto de datos con referencia a un valor esperado usando un estimador robusto, en determinados casos. En algunos casos, se calcula el valor esperado para una porción o la totalidad de los cromosomas usando uno o más estimadores elegidos de; un promedio, una mediana, moda, punto medio, media, una mediana de desviación absoluta (MAD), una alternativa a la MAD introducida por Rousseeuw y Croux, estimación de remuestreo de tipo *bootstrapping*, desviaciones estándar, puntuaciones z,

puntuación z robusta, ANOVA, análisis de regresión LOESS (por ejemplo, suavizado LOESS, suavizado LOWESS) y similares. Ajustar una porción o la totalidad de un conjunto de datos para reducir o eliminar el efecto de las secuencias repetitivas puede facilitar la obtención de un resultado y/o reducir la complejidad y/o la dimensionalidad de un conjunto de datos.

En algunos casos, una o más etapas de ajuste/procesamiento pueden comprender un ajuste de secuencia indexadora. Como se indica en el presente documento, los cebadores adaptadores utilizados en los casos descritos en el presente documento incluyen con frecuencia secuencias indexadoras. Si todos los índices tienen sustancialmente el mismo rendimiento, la representación cromosómica o alguna otra métrica equivalente relevante para la genómica se distribuiría de la misma manera en prácticamente todas las muestras etiquetadas por diferentes índices. Sin embargo, en la práctica, algunos índices funcionan mejor que otros, lo que a su vez hace que algunos fragmentos sean analizados preferiblemente (por ejemplo, sobreponderados) con respecto a otros fragmentos por un algoritmo. Además, algunos índices pueden conducir a un número menor de lecturas detectadas y/o alineadas, lo que a su vez afecta la resolución de las muestras etiquetadas con estas secuencias indexadoras, en comparación con las muestras etiquetadas con otros índices. Una porción o la totalidad de un conjunto de datos se puede ajustar o normalizar usando un estimador, con respecto a una o más secuencias indexadoras, en determinados casos, y en determinados, el estimador se elige de: un promedio, una mediana, media, moda, punto medio, una mediana de desviación absoluta (MAD), una alternativa a MAD como la introdujeron Rousseeuw y Croux, estimación de remuestreo de tipo *bootstrapping*, desviaciones estándar, puntuaciones z, puntuación z robusta, ANOVA, análisis de regresión LOESS (por ejemplo, suavizado LOESS, suavizado LOWESS) y similares. Ajustar una porción o la totalidad de un conjunto de datos para reducirlo con respecto a una o más secuencias indexadoras puede facilitar la obtención de un resultado y/o reducir la complejidad y/o la dimensionalidad de un conjunto de datos.

Una porción o la totalidad de un conjunto de datos también se puede procesar adicionalmente usando uno o más procedimientos que se describen a continuación.

En algunos casos, una o más etapas de procesamiento pueden comprender una o más etapas de filtrado. El filtrado generalmente elimina las secciones genómicas o bins 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 sobrerepresentadas 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 algunos casos, los bins pueden eliminarse sucesivamente (por ejemplo, uno cada vez para permitir la evaluación del efecto de eliminación de cada bin individual), y en determinados casos todos los bins marcados para eliminación pueden eliminarse al mismo tiempo. En algunos casos, 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 determinados casos, 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 porción de un cromosoma en un múltiplo predeterminado de la varianza del perfil, y en determinados casos, 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 porción de un cromosoma en un múltiplo predeterminado de la varianza del perfil. En algunos casos, 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, microdelección, 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 algunos casos, 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 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 (véase a continuación), repeat masking (enmascaramiento de repetición - RM), normalización de GC y enmascaramiento de repetición (GCRM), cQn y/o combinaciones de los mismos. En algunos casos, 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 GC, en algunos casos, aplica un modelo de LOESS a la relación entre el recuento de fragmentos (por ejemplo, lecturas de secuencia, recuentos) y composición de GC 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.

En determinados casos, la normalización se refiere a la división de uno o más conjuntos de datos por una variable predeterminada. Puede usarse cualquier número adecuado de normalizaciones. En algunos casos, 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 se pueden usar incluyen normalizar los datos de recuento sin procesar para una o más pruebas o secciones genómicas de referencia seleccionadas con respecto al número total de recuentos mapeados en el cromosoma o lo totalidad del genoma en el que se mapean la sección o secciones genómicas seleccionadas; normalizar los datos de recuento sin procesar para uno o más segmentos genómicos seleccionados con respecto a una mediana de recuento de referencia para una o más secciones genómicas o el cromosoma en el que se mapea un segmento o segmentos genómicos seleccionados; normalizar datos de recuento sin procesar con respecto a datos previamente normalizados o derivados de los mismos; y normalizar datos previamente normalizados con respecto a una o más variables de normalización predeterminadas diferentes. 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 algunos casos, 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 determinados casos, una etapa de procesamiento que comprende la normalización incluye la normalización con respecto a una ventana estática y, en algunos casos, una etapa de procesamiento que comprende normalización incluye la normalización con respecto a una ventana móvil o deslizante. Una ventana a menudo es 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 normalización con respecto a una ventana estática a menudo implica el uso de una o más secciones genómicas seleccionadas para la comparación entre un sujeto de prueba y el conjunto de datos del sujeto de referencia en un proceso de normalización. En algunos casos, 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. La normalización con respecto a una ventana móvil o la normalización con respecto a una ventana deslizante a menudo es una normalización realizada en 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 donde 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 determinados casos, 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 donde las ventanas adyacentes tienen una o más secciones genómicas en común. En determinados casos, una pluralidad de secciones genómicas y/o cromosomas de prueba seleccionados pueden analizarse mediante un procedimiento de ventana deslizante.

En algunos casos, la normalización con respecto a una ventana deslizante o móvil puede generar uno o más valores, en donde 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 determinados casos, 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 algunos casos, 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 determinados casos, 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 algunos casos, el análisis de ventana móvil o deslizante se usa para identificar regiones genómicas que contienen microdeleciones y, en determinados casos, el análisis de ventana móvil o deslizante se usa para identificar regiones genómicas que contienen microduplicaciones.

En algunos casos, una etapa de procesamiento comprende una ponderación. La ponderación o realizar una función de ponderación a menudo es 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 algunos casos. 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 $[1/(\text{desviación estándar})^2]$. A veces, una etapa de ponderación se realiza de una manera sustancialmente similar a una etapa de normalización. En algunos casos, 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 determinados casos, 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 algunos casos, 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 algunos casos, 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 algunos casos, 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 omnibus, 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 determinados casos, 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 algunos casos. En determinados casos, 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 algunos casos, las referencias son euploides o no tienen de cualquier otra manera una variación genética o afección médica.

En algunos casos, una etapa de ajuste/procesamiento opcionalmente comprende 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). Generar un perfil a menudo implica emplear una manipulación matemática y/o estadística de datos que facilita la identificación de patrones y/o correlaciones en grandes cantidades de datos. Un perfil a menudo son 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 determinados casos, 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 algunos casos, un perfil es representativo de la totalidad de un conjunto de datos y, en determinados casos, un perfil es representativo de una porción o subconjunto de un conjunto de datos. 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 algunos casos, un punto de datos en un perfil representa los resultados de la manipulación de datos para una sección genómica. En determinados casos, un punto de datos en un perfil representa los resultados de la manipulación de datos para grupos de secciones genómicas. En algunos casos, los grupos de secciones genómicas pueden ser adyacentes entre sí y, en determinados casos, 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 algunos casos, 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 determinados casos, 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 a menudo se presenta como un gráfico, y 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 (por ejemplo, perfil de recuento normalizado o perfil 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 algunos casos. En determinados casos, 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). Un gráfico de perfiles de recuento sin procesar o un gráfico de perfiles sin procesar a menudo es un gráfico de recuentos en cada sección

genómica en una región normalizada para recuentos totales en una región (por ejemplo, genoma, cromosoma, porción de un cromosoma). En algunos casos, puede generarse un perfil usando un procedimiento de ventana estática y, en determinados casos, puede generarse un perfil usando un procedimiento de ventana deslizante.

- 5 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 algunos casos, 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).
- 10 En determinados casos, 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 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 algunos casos, un perfil es indicativo de y/o
- 20 representativo de un fenotipo.

A modo de ejemplo no limitativo, se puede generar un conjunto de datos ajustado/normalizado a partir de datos de lectura de secuencia sin procesar (a) obteniendo recuentos totales para todos los cromosomas, cromosomas seleccionados, secciones genómicas y/o porciones de los mismos para todas las muestras de una o más celdas de flujo, o todas las muestras de una o más placas; (b) ajustando, filtrando y/o eliminando una o más de (i) secciones genómicas no informativas y/o repetitivas (por ejemplo, enmascaramiento de repetición; descrito en el ejemplo 2) (ii) sesgo de contenido de G/C (iii) secuencias sobrerrepresentadas o subrepresentadas, (iv) datos con ruido; y (c) ajustando/normalizando una porción o todos los datos restantes en (b) con respecto a un valor esperado usando un estimador robusto para el cromosoma seleccionado o la ubicación genómica seleccionada, generando así un valor

25 ajustado/normalizado. En determinados casos, los datos en (c) se ajustan opcionalmente con respecto a una o más secuencias indexadoras, uno o más estimadores adicionales, una o más etapas de procesamiento adicionales, similares o combinaciones de los mismos. En algunos casos, el ajuste, filtrado y/o eliminación de uno o más de i) secciones genómicas no informativas y/o repetitivas (por ejemplo, enmascaramiento de repetición) (ii) sesgo de contenido de G/C (iii) secuencias sobrerrepresentadas o subrepresentadas, (iv) los datos con ruido, se pueden realizar en cualquier orden (por ejemplo, (i); (ii); (iii); (iv); (i), (ii); (ii), (i); (iii), (i); (ii), (iii), (i); (i), (iv), (iii); (ii), (i) (iii); (i), (ii), (iii), (iv); (ii), (i), (iii), (v); (ii), (iv), (iii), (i); y similares). En determinados casos, los datos restantes se pueden ajustar en función de una o más condiciones experimentales descritas en el presente documento. En algunos casos, las secuencias ajustadas por un método pueden afectar a una porción de las secuencias sustancialmente ajustadas por completo por un método diferente (por ejemplo, el ajuste de sesgo del contenido de G/C a veces elimina hasta el 50 % de las

30 secuencias eliminadas sustancialmente por completo mediante el enmascaramiento de repetición).

Un conjunto de datos ajustado/normalizado puede generarse mediante una o más manipulaciones de datos de lecturas de secuencia mapeadas contadas. 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). En algunos casos, los

45 conjuntos de datos se ajustan mediante enmascaramiento de repetición para eliminar secciones genómicas no informativas o repetitivas antes del mapeo y, en determinados casos, el genoma de referencia se ajusta mediante enmascaramiento de repetición antes del mapeo. La realización de cualquier procedimiento de enmascaramiento produce sustancialmente los mismos resultados. En determinados casos, los conjuntos de datos se ajustan para el sesgo del contenido de G/C mediante la normalización de G/C basada en bins con respecto a un estimador robusto de la representación de secuencia de G/C esperada para una porción o todos los cromosomas. En algunos casos, un conjunto de datos se ajusta por enmascaramiento de repetición antes del ajuste de contenido de G/C y, en determinados casos, un conjunto de datos se ajusta por contenido de G/C antes del ajuste con enmascaramiento de repetición. Después del ajuste, los recuentos restantes se suman normalmente para generar un conjunto de datos ajustado. En determinados casos, el ajuste del conjunto de datos facilita la clasificación y/u obtención de un resultado.

50 En algunos casos, se genera un perfil de conjunto de datos ajustado a partir de un conjunto de datos ajustado y se utiliza para facilitar la clasificación y/u obtención de un resultado.

Después de contar y ajustar los datos de lectura de secuencia para secuencias repetitivas, el sesgo de contenido de G/C o secuencias repetitivas, y el sesgo de contenido de G/C, los conjuntos de datos se pueden ajustar para una o

60 más secuencias indexadoras, en algunos casos. Las muestras de varios pacientes pueden etiquetarse con diferentes secuencias indexadoras y mezclarse en una celda de flujo. El mapeo de lectura de secuencia entre pacientes e índices es homomórfico (único en ambas direcciones), en algunos casos. Una vez completadas las mediciones de secuenciación, se pueden asignar diferentes fragmentos secuenciados a los pacientes individuales de los que se originan. La separación entre diferentes fragmentos de secuencia a menudo se logra en función de las porciones indexadoras (código de barras) de las secuencias de fragmentos. Sustancialmente todos los fragmentos que llevan el mismo índice (código de barras) se agrupan y se atribuyen al paciente asociado con ese índice. El mismo

65

procedimiento se repite para cada muestra de paciente, en determinados casos. Algunos fragmentos pueden no tener índice o un índice no reconocido (debido a errores experimentales). Los fragmentos que no tienen índice o un índice no reconocido se dejan sin asignar, a menos que el índice no reconocido se parezca a uno de los índices esperados, en cuyo caso también se pueden admitir opcionalmente esos fragmentos. Solo los fragmentos que se asignan a un paciente determinado se alinean con el genoma de referencia y se cuentan para la representación cromosómica de ese paciente en particular. Después del ajuste, los recuentos restantes se suman normalmente para generar un conjunto de datos ajustado. En determinados casos, el ajuste del conjunto de datos facilita la clasificación y/u obtención de un resultado. En algunos casos, se genera un perfil de conjunto de datos ajustado a partir de un conjunto de datos ajustado y se utiliza para facilitar la clasificación y/u obtención de un resultado.

Después de contar los datos de lectura de secuencia, ajustarlos para secuencias repetitivas, sesgo de contenido de G/C o secuencias repetitivas y sesgo de contenido de G/C y/o secuencias indexadoras, los conjuntos de datos se pueden ajustar para minimizar o eliminar el efecto del sesgo de las condiciones experimentales basado en celdas de flujo y/o en placas. En determinados casos, el ajuste del conjunto de datos facilita la clasificación y/u obtención de un resultado. En algunos casos, se genera un perfil de conjunto de datos ajustado a partir de un conjunto de datos ajustado y se utiliza para facilitar la clasificación y/u obtención de un resultado.

Después del ajuste de los conjuntos de datos como se describe en el presente documento, una porción o la totalidad de un conjunto de datos también se puede procesar adicionalmente usando uno o más procedimientos que se describen a continuación. En algunos casos, el procesamiento adicional de una porción o la totalidad de un conjunto de datos comprende generar una puntuación Z como se describe en el presente documento, o como se conoce en la técnica. En determinados casos, se genera una puntuación Z como una puntuación Z robusta que minimiza los efectos de datos falsos o atípicos.

Los conjuntos de datos se pueden normalizar opcionalmente para generar perfiles de recuento normalizados. Un conjunto de datos puede normalizarse normalizando una o más secciones genómicas seleccionadas con respecto a un valor de referencia de normalización adecuado. En algunos casos, 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 determinados casos, 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 algunos casos, 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 determinados casos, el procedimiento de normalización se realiza utilizando un enfoque de ventana estática y, en algunos casos, el procedimiento de normalización se realiza utilizando un enfoque de ventana móvil o deslizante. En determinados casos, se genera un gráfico de perfiles normalizado para facilitar la clasificación y/o la obtención de un resultado. Se puede proporcionar un resultado basado en gráficos de perfiles normalizados.

Los conjuntos de datos se pueden filtrar y normalizar opcionalmente, los conjuntos de datos procesados se pueden manipular aún más mediante uno o más procedimientos de filtrado y/o normalización, en algunos casos. 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 determinados casos. El uno o más procedimientos de filtrado y/o normalización a veces pueden reducir la complejidad y/o dimensionalidad del conjunto de datos, en algunos casos. Puede proporcionarse un resultado basado en un conjunto de datos de complejidad y/o dimensionalidad reducidas.

Los conjuntos de datos se pueden manipular aún más mediante ponderación, en algunos casos. 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 determinados casos, y en algunos casos, 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 los segmentos genómicos seleccionados. En algunos casos, se pondera un conjunto de datos utilizando 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, $[1/(\text{desviación estándar})^2]$). En algunos casos, 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.

Los conjuntos de datos pueden manipularse aún más mediante una o más manipulaciones matemáticas y/o estadísticas (por ejemplo, funciones estadísticas o algoritmos estadísticos), en algunos casos. En determinados casos, 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 algunos casos, 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 conocen en la técnica. En determinados casos, las manipulaciones matemáticas y/o estadísticas incluyen una o más suposiciones que pertenecen a ploidía y/o fracción

fetal. En algunos casos, 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 determinados casos, 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.

Los conjuntos de datos se pueden manipular adicionalmente utilizando uno o más procesos elegidos de 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 algunos casos, 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 algunos casos, 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 determinados casos, 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 algunos casos, el análisis y procesamiento de datos pueden incluir el uso de una o más suposiciones. Puede usarse cualquier 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), las coincidencias idénticas 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 trillizos, cuatrillizos y similares)), el ADN fetal libre de células (por ejemplo, ADNlc) cubre uniformemente todo el genoma, similares y combinaciones de los mismos.

En los casos en donde 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. Un perfil de recuento normalizado suele ser un perfil generado mediante 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 algunos casos, un perfil de recuento normalizado puede presentarse como una representación gráfica.

Tal como se mencionó anteriormente, los datos a veces se transforman de una forma a otra. Los datos transformados, o una transformación, a menudo son 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 algunos casos 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 determinados casos, 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 o delección genómica; 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 algunos casos, 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). Cualquier característica o variable adecuada puede usarse para ajustar y/o reducir la complejidad y/o dimensionalidad del conjunto de datos. Los ejemplos no limitativos de características que pueden elegirse para su uso como característica objetivo para el ajuste/procesamiento de datos incluyen condiciones experimentales basadas en celdas de flujo y/o placas, contenido de GC, secuencias repetitivas, secuencias indexadoras, 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 complejidad de conjuntos de datos y/o reducción de dimensionalidad incluyen; reducción de una pluralidad de lecturas de secuencia a gráficos de perfiles, reducción de una pluralidad de lecturas de secuencia a valores numéricos (por ejemplo, valores normalizados, puntuaciones Z, puntuaciones Z robustas, valores de p, mediana de desviaciones absolutas o alternativas a MAD descritas en el presente documento); reducción de múltiples métodos de análisis a gráficos de probabilidad o puntos únicos; análisis de componentes principales de cantidades derivadas; y similares, o combinaciones de los mismos.

Resultado

El análisis, ajuste y procesamiento de datos puede proporcionar uno o más resultados. Un resultado a menudo es el resultado del ajuste y procesamiento de datos que facilita la determinación de si un sujeto tenía o está en riesgo de tener una variación genética. Un resultado comprende a menudo uno o más valores numéricos generados usando un método de ajuste/procesamiento descrito en el presente documento en el contexto de una o más consideraciones de probabilidad o estimadores. Una consideración de probabilidad incluye, pero sin limitación: medida de variabilidad, nivel de confianza, sensibilidad, especificidad, desviación estándar, coeficiente de variación (CV) y/o nivel de confianza, puntuaciones Z, puntuaciones Z robustas, porcentaje de representación cromosómica, mediana de desviación absoluta, o alternativas a la mediana de desviación absoluta, valores de Chi, valores de Phi, valores de ploidía, fracción fetal, fracción fetal ajustada, relaciones de área, elevación mediana, 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. En algunos casos, un resultado comprende factorizar la fracción de ácido nucleico fetal en el ácido nucleico de muestra (por ejemplo, ajustar recuentos, eliminar muestras o no realizar una identificación). La determinación de la fracción fetal a veces se realiza usando un ensayo cuantificador fetal (FQA), como se describe en los ejemplos del presente documento y se conoce en la técnica (por ejemplo, publicación de solicitud de patente de los Estados Unidos n.º: US 2010-0105049 A1, titulada "PROCESSES AND COMPOSITIONS FOR METHYLATION-BASED ENRICHMENT OF FETAL NUCLEIC ACIDS").

Un resultado a menudo es un fenotipo con un nivel de confianza asociado (por ejemplo, el feto es positivo para la trisomía 21 con un nivel de confianza del 99 %, el sujeto de prueba es negativo para un cáncer asociado con una variación genética con 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. Una puntuación, o identificación, a menudo se genera calculando 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 determinadas realizaciones, se genera un resultado a partir de un conjunto de datos ajustado. En algunos casos, un resultado proporcionado que es determinante de la presencia o ausencia de una variación genética y/o aneuploidía fetal se basa en un recuento de muestra normalizado. En algunos casos, un resultado comprende un perfil. En aquellos casos en los 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 puntuación Z robusta, 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 algunos casos, 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 algunos casos. En determinados casos, se genera un resultado que resulta de una de las etapas de procesamiento descritas en el presente documento, y en algunos casos, 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 cualquier forma adecuada, forma que comprende, sin limitación, 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 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 determinados casos, 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 algunos casos, 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. Un resultado puede describir además cualquier suposición usada en el procesamiento de datos. En determinados casos, un resultado comprende un valor que se encuentra dentro o fuera de un rango predeterminado de valores y el nivel de incertidumbre o de confianza asociado para ese valor que está dentro o fuera del rango. En algunos casos, 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. Un verdadero positivo se refiere a un sujeto diagnosticado correctamente con una variación genética. Un falso positivo se refiere a un sujeto mal identificado con una variación genética. Un verdadero negativo se refiere a un sujeto identificado correctamente sin una variación genética. Un falso negativo se refiere a un sujeto mal identificado sin 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 generalmente es la fracción de positivos previstos que se identifican correctamente como positivos; y (ii) un valor de especificidad, que generalmente es la fracción de negativos previstos correctamente identificados como negativos. Generalmente, la sensibilidad es el número de verdaderos positivos dividido entre el número de verdaderos positivos más el número de falsos negativos, en donde la sensibilidad (sens.) puede estar dentro del rango de $0 < \text{sens.} < 1$. 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. Generalmente, la especificidad es el número de verdaderos negativos dividido entre el número de verdaderos negativos más el número de falsos positivos, en donde la sensibilidad (espec.) puede estar dentro del rango de $0 < \text{espec.} < 1$. 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 determinados casos, uno o más de sensibilidad, especificidad y/o nivel de confianza se expresan como un porcentaje. En algunos casos, 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)). En algunos casos, el coeficiente de variación (CV) 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 determinados casos se expresa como una puntuación Z, un valor de p o los resultados de una prueba de la t. En algunos casos, 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.

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 algunos casos, se selecciona un método que tiene una sensibilidad igual a 1 o el 100 % y, en determinados casos, 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 algunos casos, se selecciona un método que tiene una especificidad equivalente a 1 o el 100 % y, en determinados casos, 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 %).

En algunos casos, un resultado basado en lecturas de secuencia mapeadas contadas o derivaciones de las mismas determina la presencia o ausencia de una o más afecciones, síndromes o anomalías enumerados en la Tabla 1A y 1B. En determinados casos, un resultado generado utilizando uno o más métodos de procesamiento de datos descritos en el presente documento determina la presencia o ausencia de una o más afecciones, síndromes o anomalías enumerados en la Tabla 1A y 1B. En algunos casos, un resultado determinante de la presencia o ausencia de una afección, síndrome o anomalía es, o incluye, la detección de una afección, síndrome o anomalía enumerados en las Tabla 1A y 1B.

En determinados casos, un resultado se basa en una comparación entre: una muestra de prueba y una muestra de referencia; una muestra de prueba y otras muestras; dos o más muestras de prueba; similares; y combinaciones de los mismos. En algunos casos, la comparación entre muestras facilita la obtención de un resultado. En determinadas realizaciones, un resultado se basa en una puntuación Z generada como se describe en el presente documento o como se conoce en la técnica. En algunas realizaciones, se genera una puntuación Z usando un recuento de muestra normalizado. En algunas realizaciones, la puntuación Z generada para facilitar la obtención de un resultado es una puntuación Z robusta generada usando un estimador robusto. En determinadas realizaciones, un resultado se basa en un recuento de muestra normalizado.

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). En algunos casos, 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 determinados casos el informe comprende una visualización de un valor de resultado y un parámetro de confianza asociado. Generalmente, un resultado puede mostrarse en cualquier 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 las figuras y se describen en los ejemplos.

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 algunos casos. 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 algunos. En determinados casos, un profesional sanitario o un individuo cualificado realiza una puntuación o identificación manualmente, usando la observación visual del informe proporcionado. En determinados casos, 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.

Recibir un informe a menudo implica obtener, mediante un medio de comunicación, un texto y/o representación gráfica que comprende un resultado, que 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 cualquier otro método para enviar o recibir datos (por ejemplo, servicio de correo, servicio de mensajería y similares). En algunos casos, 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 información de resultados también se puede obtener de 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 médica 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 determinados casos.

Un profesional sanitario o individuo cualificado puede proporcionar cualquier 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 algunos casos, 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).

Se puede usar software para realizar una o más etapas en el proceso descrito en el presente documento, incluyendo, pero sin limitación; contar, procesar datos, generar un resultado y/o proporcionar una o más recomendaciones basadas en los resultados generados.

Máquinas, software e interfaces

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 algunos casos, un usuario puede introducir un conjunto de datos como información de entrada, un usuario puede descargar uno o más conjuntos de datos mediante cualquier 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).

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 determinados casos, 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, uno o más algoritmos estimadores robustos, 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 algunos casos, los procedimientos pueden implementarse como un solo sistema de usuario ubicado en un solo lugar geográfico. En determinados casos, 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 donde el usuario accede a una página web para introducir y recuperar información. Por consiguiente, en determinados casos, 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 lo tanto, puede usarse cualquier configuración y 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”.

En algunos casos, un aparato puede comprender un sistema basado en web en el que se implementa un producto de programa informático descrito en el presente documento. Un sistema basado en web a veces comprende ordenadores, equipos de telecomunicaciones (por ejemplo, interfaces de comunicaciones, enrutadores, conmutadores de red) y similares suficientes para la funcionalidad basada en web. En determinados casos, un sistema basado en web incluye computación en la nube en red, almacenamiento en la nube en red o computación en la nube en red y almacenamiento en la nube en red. El almacenamiento en la nube en red generalmente es un almacenamiento de datos basado en web en servidores virtuales ubicados en Internet. La computación en la nube en red generalmente es el uso de software y/o hardware basado en la red que tiene lugar en un entorno de red remoto (por ejemplo, software disponible para su uso por algunos ubicado en un servidor remoto). En algunos casos, una o más funciones de un producto de programa informático descrito en el presente documento se implementan en un entorno basado en web.

Un sistema puede incluir una interfaz de comunicaciones en algunos casos. 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 cualquier dispositivo y/o método adecuado incluyendo, pero sin limitación, 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 algunos casos, la salida de un aparato de secuenciación puede servir como datos que pueden introducirse a través de un dispositivo de entrada. En determinados casos, las lecturas de secuencia mapeadas pueden servir como datos que pueden introducirse a través de un dispositivo de entrada. En determinados casos, 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. Como se usa en el presente documento, “*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 adquisición de datos, módulo de procesamiento de datos, módulo de organización de visualización de datos). A menudo, software son instrucciones de programas legibles por ordenador que, cuando se ejecutan por un ordenador, realizan operaciones informáticas. A menudo, un módulo es una unidad funcional autónoma que puede usarse en un sistema de software más grande. Por ejemplo, un módulo de software forma parte de un programa que realiza un procedimiento o tarea particular.

El software a menudo se proporciona en un producto de programa que contiene instrucciones de programa grabadas en un medio legible por ordenador, incluyendo, pero sin limitación, medios magnéticos que incluyen disquetes, discos duros y cintas magnéticas; y medios ópticos que incluyen discos CD-ROM, discos DVD, discos magneto-ópticos, unidades flash, RAM, disquetes, similares, y otros medios de este tipo en los que se pueden grabar las instrucciones de 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 ajusta y/o procesa de manera específica los datos (por ejemplo, un módulo de procesamiento que ajusta y/o procesa los datos recibidos (por ejemplo, filtra, normaliza, proporciona un resultado y/o informe). Obtener y/o recibir información de entrada a menudo implica 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 algunos casos, la información de entrada se modifica antes de procesarse (por ejemplo, se coloca en un formato susceptible al procesamiento (por ejemplo, tabulado)).

En algunos casos, 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, cuyo genoma conocido se ha dividido en secciones genómicas; (c) contar las lecturas de secuencia mapeadas dentro de las secciones genómicas; (d) generar un conjunto de datos ajustado ajustando los recuentos o una derivada de los recuentos para las secciones genómicas obtenidas en (c); y (e) proporcionar un resultado determinante de la presencia o ausencia de una variación genética a partir del perfil de recuento ajustado en (d).

El software puede incluir uno o más algoritmos en determinados casos. 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 geometría 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 algunos casos, 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 algunos casos, 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 agrupación).

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

En determinados casos, los datos simulados (o simulación) pueden ayudar al ajuste y/o procesamiento de datos, por ejemplo, entrenando un algoritmo o someter a prueba un algoritmo. En algunos casos, 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 determinados casos. 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 algunos casos, 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 algunos casos, 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 determinados casos. 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 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 algunos casos, 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 algunos casos, 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 proporcionar un resultado determinante de la presencia o ausencia de una variación genética.

En algunos casos, 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.

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 algunos casos 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 determinados casos, 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 delección; 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 algunos casos, 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 complejidad de conjuntos de datos y/o reducción de dimensionalidad incluyen; reducción de una pluralidad de lecturas de secuencia a gráficos de perfiles, reducción de una pluralidad de lecturas de secuencia a valores numéricos (por ejemplo, valores normalizados, puntuaciones Z, valores de p); reducción de múltiples métodos de análisis a gráficos de probabilidad o puntos únicos; análisis de componentes principales de cantidades derivadas; y similares, o combinaciones de los mismos.

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) normalizar los recuentos para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que comprende muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; y (b) determinar la presencia o ausencia de una aneuploidía fetal basándose en el recuento de muestra normalizado.

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) ajustar las lecturas de secuencia mapeadas contadas en según una variable o característica seleccionada, cuya característica o variable seleccionada minimiza o elimina el efecto de secuencias repetitivas y/o secuencias sobrerrepresentadas o subrepresentadas; (b) normalizar los recuentos restantes después en (a) para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que comprende muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; (c) evaluar la significación estadística de las diferencias entre los recuentos normalizados o una derivada de los recuentos normalizados para el sujeto de prueba y los sujetos de referencia para una o más secciones genómicas seleccionadas; y (d) determinar la presencia o ausencia de una variación genética en el sujeto de prueba basándose en la evaluación en (c).

En determinados aspectos, también 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) normalizar los recuentos para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que comprende muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; y (b) determinar la presencia o ausencia de una aneuploidía fetal basándose en el recuento de muestra normalizado.

En determinados aspectos, también 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) ajustar las lecturas de secuencia mapeadas contadas en según una variable o característica seleccionada, cuya característica o variable seleccionada minimiza o elimina el efecto de secuencias repetitivas y/o secuencias sobrerrepresentadas o subrepresentadas; (b) normalizar los recuentos restantes después en (a) para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que comprende muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; (c) evaluar la significación estadística de las diferencias entre los recuentos normalizados o una derivada de los recuentos normalizados para el sujeto de prueba y los sujetos de referencia para una o más secciones genómicas seleccionadas; y (d) determinar la presencia o ausencia de una variación genética en el sujeto de prueba basándose en la evaluación en (c).

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 libre de células circulante de un sujeto de prueba mapeadas en secciones genómicas de un genoma de referencia, (b) normalizar los recuentos para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que comprende muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; y (c) determinar la presencia o ausencia de una aneuploidía fetal basándose en el recuento de muestra normalizado.

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 extracelular a partir de una muestra de prueba; (b) ajustar las lecturas de secuencia contadas y mapeadas en según una variable o

característica seleccionada, cuya característica o variable seleccionada minimiza o elimina el efecto de secuencias repetitivas y/o secuencias sobrerrepresentadas o subrepresentadas; (c) normalizar los recuentos restantes en (b) para una primera sección genómica, o normalizar una derivada de los recuentos para la primera sección genómica, según un recuento esperado, o una derivada del recuento esperado, obteniendo así un recuento de muestra normalizado, cuyo recuento esperado, o derivada del recuento esperado, se obtiene para un grupo que comprende muestras, referencias o muestras y referencias, expuesto a una o más condiciones experimentales comunes; (d) evaluar la significación estadística de las diferencias entre los recuentos normalizados o una derivada de los recuentos normalizados para el sujeto de prueba y los sujetos de referencia para una o más secciones genómicas seleccionadas; y (e) determinar la presencia o ausencia de una variación genética en el sujeto de prueba basándose en la evaluación en (d).

En determinados casos, 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 recuento esperado configurado para proporcionar recuentos esperados o una derivada de recuentos esperados; (viii) un módulo de representación gráfica configurado para crear gráficos y mostrar una elevación y/o un perfil; (ix) un módulo de resultado configurado para determinar un resultado (por ejemplo, un resultado determinante de la presencia o ausencia de una aneuploidía fetal); (x) un módulo de organización de visualización de datos configurado para indicar la presencia o ausencia de una anomalía cromosómica segmentaria o una aneuploidía fetal o ambas; (xi) un módulo de procesamiento lógico configurado para realizar uno o más mapeo de lecturas de secuencia, contar lecturas de secuencia mapeadas, normalizar recuentos y generar un resultado; o (xii) una combinación de dos o más de los anteriores.

En algunos casos, 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 algunos casos, el módulo de normalización y/o el módulo de recuento esperado están configurados para transferir recuentos normalizados a un módulo de resultados o módulo de representación gráfica.

Módulos

Los módulos a veces forman parte de un aparato, sistema o software y pueden facilitar la transferencia y/o procesamiento de información y datos. En lo sucesivo se describen ejemplos no limitativos de módulos.

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 determinados casos, 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 algunos casos, se requiere un módulo de secuenciación o un aparato que comprende un módulo de secuenciación para proporcionar lecturas de secuenciación. En algunos casos, 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 algunos casos, 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 algunos casos, las lecturas de secuenciación se proporcionan por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunos casos, 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 algunos casos, 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 algunos casos, se imprime o visualiza una secuencia de ácido nucleico proporcionada por un módulo de secuenciación. En algunos casos, 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 algunos casos, 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 algunos casos, un módulo de secuenciación puede transferir lecturas de secuencia a un módulo de mapeo o módulo de recuento.

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 algunos casos, 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 algunos casos, 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 algunos casos, las lecturas de secuenciación se mapean por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunos casos, 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 algunos casos, 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 algunos casos, 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 algunos casos.

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 algunos casos, 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 algunos casos, 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 algunos casos, las lecturas se cuentan con un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunos casos, 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 algunos casos, 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.

5 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 algunos casos, 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 módulo de normalización puede proporcionar una estimación de la variabilidad de los recuentos esperados (por ejemplo, una MAD de los recuentos esperados y/o una MAD de una representación de recuentos esperados). En algunos casos, un módulo de normalización puede proporcionar una MAD de recuentos esperados al derivar múltiples valores de la mediana de recuentos esperados obtenidos de múltiples experimentos (por ejemplo, a veces experimentos diferentes, a veces experimentos expuestos a una o más condiciones experimentales comunes), al derivar un error absoluto (por ejemplo, desviación, variabilidad, desviación estándar, error estándar) de los múltiples valores de la mediana y al determinar una media, promedio o mediana de los errores absolutos calculados. En algunos casos, un módulo de normalización puede proporcionar una MAD de una representación de recuentos esperados al derivar múltiples valores de la mediana de las representaciones de recuentos esperados obtenidas de múltiples experimentos (por ejemplo, a veces experimentos diferentes, a veces experimentos expuestos a una o más condiciones experimentales comunes) y a continuación al derivar un error absoluto (por ejemplo, desviación, variabilidad, desviación estándar, error estándar) de los múltiples valores de la mediana. Un aparato que comprende un módulo de normalización puede comprender al menos un procesador. En algunos casos, 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 algunos casos, los datos normalizados se proporcionan por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunos casos, 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 algunos casos, 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 algunos casos. 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 determinados casos. 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.

Módulo de recuento esperado

Un recuento esperado o una derivada de un recuento esperado (por ejemplo, una representación porcentual) puede proporcionarse por un módulo de recuento esperado (por ejemplo, por un aparato que comprende un módulo de recuento esperado). En algunos casos, se requiere un módulo de recuento esperado para proporcionar recuentos esperados o una derivada de los recuentos esperados obtenidos a partir de lecturas de secuenciación (por ejemplo, recuentos de lecturas de secuencia mapeadas, subconjuntos predeterminados de lecturas de secuencia mapeadas). Un módulo de recuento esperado puede sumar los recuentos de una o más secciones genómicas seleccionadas. A veces, un módulo de recuento esperado aplica una o más manipulaciones matemáticas o estadísticas a lecturas y/o recuentos de secuencia. Un módulo de recuento esperado puede determinar una derivada de un recuento esperado determinando una representación porcentual (por ejemplo, una representación de recuento). Un aparato que comprende un módulo de recuento esperado puede comprender al menos un procesador. En algunos casos, un recuento esperado o una derivada de un recuento esperado se proporciona por un aparato que incluye un procesador (por ejemplo, uno o más procesadores), cuyo procesador puede ejecutar y/o implementar una o más instrucciones (por ejemplo, procesos, rutinas y/o subrutinas) del módulo de recuento esperado. En algunos casos, se proporciona un recuento esperado o una derivada de recuento esperado por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunos casos, un módulo de recuento esperado 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 algunos casos, un recuento esperado o una derivada de un recuento esperado se proporciona 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 recuento esperado puede recibir información y/o datos de un aparato o módulo adecuado. A veces, un módulo de recuento esperado puede recibir datos y/o información de un módulo de secuenciación, un módulo de recuento esperado, un módulo de mapeo, un módulo de normalización o un módulo de recuento. Un módulo de recuento esperado 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 algunos casos. A menudo, un módulo de recuento esperado recibe datos y/o información de otro aparato o módulo, transforma los datos y/o la información y proporciona un recuento esperado o una derivada de un recuento esperado. Un recuento esperado o una derivada de un recuento esperado se pueden transferir de un módulo de recuento esperado a un módulo de comparación, un módulo de recuento esperado, 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 determinados casos.

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 algunos casos, 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 algunos casos, 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 algunos casos, 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 algunos casos, 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 algunos casos, 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 algunos casos, se proporcionan información y/o datos por un aparato que incluye múltiples procesadores, tales como procesadores coordinados y que funcionan en paralelo. En algunos casos, 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, un módulo de recuentos esperado, 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 algunos casos, 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 determinados casos 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 las figuras 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 determinados casos. 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 algunos casos, 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 algunos casos, 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).

Ejemplos

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

Ejemplo 1: Determinación de la presencia o ausencia de una variación genética usando muestras ciegas

Las pruebas de cribado prenatales eficaces para el síndrome de Down a menudo combinan la edad materna con la información de la medición ecográfica de la translucencia nuchal en el primer trimestre y/o las mediciones de varios marcadores de cribado del suero materno obtenidos en el primer y segundo trimestre. Estas pruebas de cribado prenatal a menudo detectan hasta aproximadamente el 90 % de prácticamente todos los casos con una tasa de falsos positivos de aproximadamente el 2 %. Dada la prevalencia del síndrome de Down, 1 de cada 16 mujeres con cribado positivo a las que se les ofrecen pruebas diagnósticas invasivas (por ejemplo, amniocentesis o muestreo de vellosidades coriónicas) tendrán un embarazo afectado y 15 no. Hasta 1 de cada 200 procedimientos invasivos de este tipo están asociados con la pérdida fetal, una consecuencia adversa significativa del diagnóstico prenatal. La importante consecuencia adversa de la pérdida fetal a veces ha llevado a ajustar los puntos de corte de cribado para minimizar la tasa de falsos positivos. En la práctica, son comunes tasas de falsos positivos de aproximadamente el 5 %.

El descubrimiento de que aproximadamente el 3-6 % del ADN libre de células en la sangre materna era de origen fetal impulsó estudios para determinar si el síndrome de Down podía detectarse de forma no invasiva. El síndrome de Down fetal se identificó mediante massively parallel shotgun sequencing (secuenciación de escopeta masivamente paralela - MPSS), una técnica que secuencía las primeras 36 bases de millones de fragmentos de ADN para determinar su origen cromosómico específico. Si un feto tiene un tercer cromosoma 21, el porcentaje de fragmentos del cromosoma 21 es ligeramente superior al esperado. Informes posteriores han ampliado estas observaciones y sugieren que se puede lograr una tasa de detección de al menos aproximadamente el 98 % con una tasa de falsos positivos de aproximadamente el 2 % o menos. Aunque prometedores, estos estudios estuvieron limitados por los siguientes factores; los estudios se realizaron utilizando grupos de pacientes relativamente pequeños (intervalo de 13-86 casos de síndrome de Down y 34-410 muestras de control euploides); la secuenciación de ADN no se realizó en laboratorios certificados por CLIA; y el rendimiento y los tiempos de respuesta no simulaban la práctica clínica.

Los métodos, procesos y aparatos descritos en el presente documento se pueden utilizar para proporcionar un resultado determinante de la presencia o ausencia de una variación genética (por ejemplo, trisomía, síndrome de Down) usando muestras ciegas y sin la necesidad de un conjunto de datos del genoma de referencia para el cual está normalizado el sujeto de prueba los datos, en algunos casos.

Materiales y métodos

Diseño general del estudio

El estudio presentado en el presente documento (véase la URL www.clinicaltrials.gov/NCT00877292) involucró a pacientes incluidas en 27 centros de diagnóstico prenatal en todo el mundo (por ejemplo, en lo sucesivo en el presente documento denominados Sitios de inclusión). Las mujeres con alto riesgo de síndrome de Down en función de la edad materna, los antecedentes familiares o una prueba de cribado en suero y/o ecografía positiva proporcionaron su

consentimiento, muestras de plasma e información demográfica y relacionada con el embarazo. Se obtuvo la aprobación de la Junta de Revisión Institucional (o equivalente) en cada sitio de inclusión. La identificación de pacientes y muestras fue por código de estudio. Las muestras se extrajeron inmediatamente antes de la prueba invasiva, se procesaron en un plazo de 6 horas, se almacenaron a -80°C y se enviaron en hielo seco al centro de coordinación. Dentro de esta cohorte, se desarrolló un estudio anidado de caso-control, con pruebas de ADN ciegas para el síndrome de Down. Siete muestras euploides se emparejó con cada caso, según la edad gestacional (semana más cercana; mismo trimestre), el sitio de inclusión, la raza (autodeclarada) y tiempo en el congelador (en 1 mes). Asumiendo que no hubo resultados falsos negativos, 200 embarazos (casos) con síndrome de Down tuvieron una facultad del 80 % para rechazar el 98 % como el intervalo de confianza inferior (IC). Los casos se distribuyeron por igual entre el primer y segundo trimestre. Para este estudio, el síndrome de Down se definió como 47, XY, +21 o 47, XX, +21; se excluyeron mosaicismos y embarazos gemelares con síndrome de Down. La coordinación del estudio y el almacenamiento de muestras se realizaron en un centro médico académico independiente (por ejemplo, Women & Infants Hospital). Se enviaron muestras congeladas codificadas (4 ml) al Sequenom Center for Molecular Medicine (SCMM, San Diego, CA) para su análisis. El SCMM no tenía conocimiento del cariotipo ni de las pruebas clínicas simuladas, incluida la cuantificación del tiempo de respuesta. Se envió un subconjunto de muestras para su ensayo al Orphan Disease Testing Center en la University of California at Los Angeles (UCLA; Los Angeles, CA), un laboratorio académico independiente con experiencia en secuenciación de ADN. Ambos laboratorios tenían la certificación CLIA y ambos proporcionaron interpretaciones clínicas usando un protocolo escrito estandarizado desarrollado originalmente por el SCMM.

Integridad del estudio

Se dio la más alta prioridad a garantizar la integridad, fiabilidad e independencia de este estudio financiado por la industria. Se creó un Comité de Supervisión de tres personas (véanse los Agradecimientos) y se encargó de evaluar y brindar recomendaciones sobre el diseño, la realización, el análisis y la interpretación del estudio. El protocolo del estudio incluía inspecciones del sitio de inclusión, aislamiento de los sitios de inclusión del patrocinador del estudio, pruebas de confirmación por parte de un laboratorio académico independiente, cegamiento de los resultados de las pruebas de diagnóstico en múltiples niveles, sin acceso remoto por ordenador a los datos de resultados, acceso a todos los datos sin procesar por parte del sitio de prueba académico, transferencia inmediata de archivos de secuenciación y resultados de interpretación al centro de coordinación, y uso de sumas de verificación de archivos para identificar cambios posteriores. El SCMM proporcionó al laboratorio independiente equipo, formación, software interpretativo y protocolos operativos estándar similares.

La prueba desarrollada en el laboratorio

Como se ha señalado anteriormente, se utilizó MPSS para secuenciar el ADN libre de células. En resumen, los fragmentos de ADN libre de células circulante se aíslan del plasma materno y se cuantifican con un ensayo que determina la contribución fetal (fracción fetal). El aislado restante se usó para generar bibliotecas de secuenciación, normalizadas y multiplexadas para permitir que se ejecutaran cuatro muestras en un solo carril de celda de flujo (por ejemplo, ocho carriles por celda de flujo). Las bibliotecas de ADN se cuantificaron mediante una plataforma de microfluidos (Caliper Life Sciences, Hopkinton, MA) y se generaron agrupaciones mediante la plataforma cBot (Illumina, Inc, San Diego, CA). Las celdas de flujo se secuenciaron en la plataforma Illumina HiSeq 2000 y se analizaron los datos resultantes con el software Illumina. La interpretación por ordenador proporcionó una estimación robusta de las desviaciones estándar (por ejemplo, D.E.) por encima o por debajo de la estimación central (puntuación z); las puntuaciones z de 3 o más se consideraron compatibles con el síndrome de Down. El director del laboratorio CLIA primario (SCMM) revisó los resultados, inició peticiones para analizar segundas alícuotas y proporcionó una interpretación de "registro" final para todos los embarazos analizados. El director del laboratorio CLIA independiente (UCLA) hizo lo mismo pero sin la capacidad de solicitar alícuotas de segunda muestra. Cada laboratorio solo tenía acceso a sus propios resultados.

Análisis estadístico

El estudio se detuvo cuando un análisis intermedio mostrara que más de 3 de 16 casos o 6 de 112 controles estaban mal clasificados. Aunque se trataba de un estudio emparejado, se planificó que el análisis no estuviera igualado. Se examinaron las diferencias entre grupos y asociaciones usando la prueba χ^2 , la prueba de la t , análisis de varianza (ANOVA) y regresión lineal (después de las transformaciones apropiadas) usando SASTM Analytics Pro (Cary, NC; anteriormente conocido como Statistical Analysis System) y True Epistat (Richardson, TX). Los intervalos de confianza (IC) de las proporciones se calcularon usando una distribución binomial. Los valores de p fueron de dos colas y la significación estaba en el nivel de 0,05.

Resultados

Población de muestras

Entre abril de 2009 y febrero de 2011, 27 sitios de inclusión (véase la Tabla 1 a continuación) identificaron mujeres embarazadas aptas, obtuvieron el consentimiento informado y recogieron muestras. Entre 4664 incluidas, se

produjeron 218 embarazos únicos con síndrome de Down y 3930 embarazos únicos euploides. La Fig. 1 proporciona detalles sobre los resultados fetales, el estado de la muestra de plasma y las razones por las que se excluyó a 279 mujeres (6 %). Ninguna de las muestras se incluyó en publicaciones o estudios anteriores. Un total de 4385 mujeres (94 %) tuvieron un embarazo único, al menos dos muestras de plasma adecuadas y resultados de pruebas de diagnóstico. De estas, el 97 % tenían entre 11 y 20 semanas de gestación, inclusive; el 34 % estaban en el primer trimestre. Los cariotipos fetales (o equivalentes) estaban disponibles para todas menos 51 mujeres incluidas. Para 116 mujeres, las muestras de plasma no se consideraron adecuadas para la prueba (por ejemplo, descongeladas durante el tránsito, más de 6 horas antes de ser congeladas, solo una alícuota y volumen insuficiente). Se excluyeron 112 mujeres adicionales debido a gestaciones múltiples o muerte fetal existente. Entre los 4385 embarazos únicos viables, el 34 % se obtuvieron al final del primer trimestre y el 66 % al principio del segundo trimestre. Se seleccionaron un total de 212 casos de síndrome de Down para la prueba. Para cada caso, se eligieron siete embarazos euploides emparejados (por ejemplo, 1484; relación de 7:1 de euploides con respecto a casos de síndrome de Down). Entre los otros 237 resultados se encontraban aneuploidías autosómicas adicionales, aneuploidías de cromosomas sexuales, mosaicismos y otras anomalías cromosómicas. Más tarde se descubrió que un control era trisomía 18, pero se incluyó como un control "euploide".

Tabla 1 Centros clínicos incluidos en el estudio, junto con la inclusión relacionada y la información de resultados

Sitio de inclusión	Ubicación	Investigador clínico	Embarazo único Síndrome de Down	Cariotipo normal	Otro	Pacientes incluidas
North York General Hospital	Toronto, Canadá	Wendy S. Meschino, Méd.	41	651	86	778
Istituto G. Gaslini	Genoa, Italia	Pierangela De Biasio, Méd.	27	492	35	554
Hospital Clinic Barcelona	Barcelona, España	Antoni Borrell, Méd., Dr.	24	291	44	359
Centrum Lekarske Genetiky	Ceske Budejovice, República Checa	David Cutka, Méd.	14	362	19	395
Hospital Italiano	República Buenos Aires,	Lucas Otano, Méd., Dr.	13	68	14	95
Dalhousie University	Argentina	Michiel Van den Hof, Méd.	12	115	18	145
Rotunda Hospital	Halifax, Canadá	Fergal Malone, Méd.	12	70	12	94
Semmelweis University	Dublín, Irlanda	Csaba Papp, Méd., Dr.	10	64	9	83
IMALAB s.r.o. Medical Laboratories	Budapest, Hungría	Jaroslav Loucky, RNDr	9	238	8	255
	Zlin, República Checa	Maria Laura Igarzabal, Méd.	8	224	49	281
CEMIC	Buenos Aires, Argentina	Kristi Borowski, Méd.	8	135	30	173
University of Iowa	Iowa City, IA	Barbara O'Brien, Méd.	6	99	21	126
		Béla Veszprémi, Méd., Dr.				
Women & Infants Hospital	Providence, RI	Dr.	4	172	31	207
University of Pecs	Pecs, Hungría	Joseph Biggio, Méd.	4	169	20	193
University of Alabama at Birmingham	Birmingham, AL	Zeev Weiner, Méd.	4	133	10	147
	Haifa, Israel	John Williams, Méd.	3	192	28	223
Rambam Medical Center	Los Angeles, CA	Jeffrey Dungan, MD	3	88	11	102
Cedars Sinai PDC	Chicago, IL	Jacquelyn Roberson, MD	3	74	14	91
Northwestern University	Detroit, MI	Devereux N. Sailer, Jr, MD	3	21	8	32
Henry Ford Hospital	Charlottesville, VA	Sylvie Langlois, MD	2	67	14	83
University of Virginia	Vancouver, Canadá	Nancy Rose, Méd.	2	67	9	78
University of British Columbia	Salt Lake City, UT	Louise Wilkins-Haug, Méd.	2	21	8	31
Intermountain Healthcare	Boston, MA	Anthony Johnson, DO	2	20	0	22
Brigham and Women's Hospital	Houston, TX	Maurice J. Mahoney, Méd.,	1	31	9	41
Baylor College of Medicine	New Haven, CT	JD	1	7	4	12
Yale University	Providence, RI	Marshall Carpenter, Méd.	0	52	5	57
New Beginnings Perinatal Consultores	Calgary, Canadá	Jo-Ann Johnson, Méd.	0	7	0	7
University of Calgary	Sydney, Australia	Vitomir Tasevski, Dr.	218	3930	516	4664
Royal North Shore Hospital						
All						

La Tabla 2 a continuación compara la información demográfica y relacionada con el embarazo entre casos y controles. La coincidencia fue exitosa. La mediana de edad fue de aproximadamente 37 años en ambos grupos; todas tenían 18 años o más. Las indicaciones para las pruebas diagnósticas diferían, siendo más probable que los casos tuvieran una anomalía ecográfica o múltiples indicaciones. Las muestras se recogieron, se procesaron y se congelaron, en promedio, en 1 hora; todas en 6 horas. Los resultados se basaron en el cariotipado, excepto en dos casos del primer trimestre (reacción en cadena de la polimerasa cuantitativa en uno e hibridación fluorescente *in situ* en el otro, de productos de la concepción después de la terminación de un feto viable con graves anomalías ecográficas).

Tabla 2 Información demográfica y relacionada con el embarazo para las muestras seleccionadas de síndrome de Down y euploides emparejadas ensayadas

Característica	Síndrome de Down	Euploide	P
Número de muestras	212	1484	
Edad materna en años (promedio, D.E.)	37,0, 5,0	36,6, 5,1	0,36
Edad materna 35 años o más (N, %)	160 (75 %)	1036 (70 %)	0,12
Edad gestacional (promedio, intervalo)	15,3 (9,2-21,3)	15,0 (8,1-21,5)	0,21
Edad gestacional en el primer/segundo trimestre	50 %/50 %	50 %/50 %	1,0
Peso materno en libras (promedio, D.E.)	149 (30)	152 (33)	0,33
Sangrado (%)	17 %	15 %	0,44
Raza materna (N, %)			
Caucásica	188 (89 %)	1316 (89 %)	
Negra	5 (2 %)	35 (2 %)	
Asiática	15 (7 %)	105 (7 %)	
Desconocida	4 (2 %)	28 (2 %)	
Caucásica hispana (N, %)	39 (18 %)	303 (20 %)	0,92
Judía ashkenazi (N, %)	3 (1 %)	42 (3 %)	0,13
Indicación principal de inclusión (N, %)	48 (23 %)	327 (22 %)	<0,001
Cribado positiva en la prueba del primer trimestre	11 (5 %)	118 (8 %)	
Cribado positiva en la prueba del segundo trimestre	38 (18 %)	192 (13 %)	
Cribado positivo por prueba integrada	51 (24 %)	130 (9 %)	
Anomalía de ultrasonido identificada	24 (12 %)	543 (37 %)	
Edad materna avanzada	39 (18 %)	112 (8 %)	
Dos o más indicaciones	0 (0 %)	44 (3 %)	
Antecedentes familiares de aneuploidía	1 (<1 %)	18 (1 %)	
Otro o No indicado			0,79
Procedimiento diagnóstico (N, %)	114 (54 %)	787 (53 %)	
Amniocentesis	97 (46 %)	697 (47 %)	
Muestreo de vellosidades coriónicas	1 (<1 %)	0 (0 %)	
Examen de los productos de la concepción			<0,001
Prueba diagnóstica (N, %)	95 (46 %)	805 (54 %)	
Cariotipo solo	115 (53 %)	679 (45 %)	
Cariotipo y otros	2 (<1 %)	0 (0 %)	
QF-PCR o FISH solo	8 (4 %)	45 (3 %)	0,60
Hemólisis de moderada a grave (N, %)	1,1 (0,1-6)	1,2 (0,1-6)	0,63
Procesamiento de muestras en horas (media, intervalo)			

Contribución fetal al ADN libre circulante

Antes de la MPSS, el ADN extraído se ensayó para determinar la proporción de ADN libre de origen fetal en el plasma materno (fracción fetal). Casi todos (1687/1696; 99,5 %) tuvieron una fracción fetal final dentro de los límites aceptables (4-50 %); la media geométrica fue del 13,4 %. Se eligió el punto de corte más bajo para minimizar los resultados de falsos negativos. El punto de corte superior se eligió para alertar al director del laboratorio de que esto representa un evento raro. Nueve tenían niveles inaceptables; seis por debajo del umbral y tres por encima. Dado que el éxito de la MPSS en la identificación del síndrome de Down depende en gran medida de la fracción fetal, se exploraron 16 posibles covariables (véanse las Figs. 4-19, ejemplo 2) (tiempo de procesamiento, hemólisis, región geográfica, indicación de pruebas de diagnóstico, sitio de inclusión, edad gestacional, edad materna, peso materno, sangrado vaginal, raza materna, etnia caucásica, sexo del feto, tiempo de almacenamiento en el congelador y efecto de la fracción fetal en la concentración de bibliotecas de ADN, número de secuencias coincidentes y resultado fetal).

Se observó una fuerte asociación negativa de la fracción fetal con el peso materno en mujeres de casos y controles (véase la Fig. 11, ejemplo 2), estando los pesos de 100, 150 y 250 libras asociados con fracciones fetales predichas del 17,8 %, 13,2 % y 7,3 %, respectivamente. No se encontró asociación para la edad gestacional, la raza materna o la indicación para la prueba. Otras asociaciones fueron pequeñas y generalmente no significativas.

Prueba de secuenciación de escopeta masivamente paralela para el síndrome de Down

Las pruebas se realizaron durante 9 semanas (enero a marzo de 2011) por 30 científicos, técnicos/tecnólogos moleculares con preparación en los protocolos de ensayo e instrumentación relacionada. Los rangos de referencia históricos debían usarse para la interpretación, 9 con un requisito de revisión en tiempo real de nuevos datos. La revisión de las primeras celdas de flujo por parte del director del laboratorio (antes del registro) reveló que era necesario realizar ajustes en los datos de referencia (véase el ejemplo 2 y las Figs. 20-22). Después de que se generaron los datos de seis celdas de flujo, el Comité de Supervisión evaluó los resultados según los criterios provisionales y se tomó la decisión confidencial de permitir que continuaran las pruebas. Al finalizar la prueba, pero antes de revelar la ocultación, el SCMM solicitó una segunda alícuota para 85 de las 90 pruebas fallidas entre las 1696 incluidas (5,3 %; 95 % de IC, 4,3-6,5; véase el ejemplo 2). El segundo resultado se usó para la interpretación final.

Las muestras de síndrome de Down mostraron una relación positiva clara y significativa con la fracción fetal; 208 de las muestras están por encima del punto de corte y cuatro por debajo. Cuatro muestras de síndrome de Down tenían puntuaciones z por debajo del límite de 3; todas tenían fracciones fetales de <7 %. (por ejemplo, 7 %, 7 %, 5 % y 4 %). Existía una fuerte asociación positiva entre la fracción fetal y la puntuación z para los casos (después de la transformación logarítmica, pendiente = 0,676, $P < 0,001$) pero no para los controles (pendiente = 0,0022, $P = 0,50$). Una de las muestras de síndrome de Down de fracción fetal baja tenía una puntuación z inicial de 5,9 con un fallo de calidad límite; la puntuación z de la muestra repetida fue de 2,9 (un valor límite compatible con el resultado positivo inicial). La combinación de la información de la muestra repetida con una puntuación de 5,9 en la muestra inicial (por ejemplo, un fallo límite), permitió al director del laboratorio tomar la decisión correcta. Todas las demás interpretaciones clínicas coincidieron con la interpretación por ordenador. Por lo tanto, los resultados registrados identificaron correctamente 209 de 212 fetos con síndrome de Down (tasa de detección del 98,6 %; 95 % de IC, 95,9-99,7).

La interpretación clínica de todas las muestras de síndrome de Down y euploides usadas en el estudio es la siguiente: Entre los embarazos euploides, 1471 fueron negativos, 3 positivos y 13 fracasaron también en la segunda alícuota. Entre los embarazos con síndrome de Down, 209 fueron positivos y 3 negativos. Entre las 1471 muestras euploides, 3 tenían puntuaciones $z > 3$ en un intervalo de fracciones fetales y se clasificaron incorrectamente como síndrome de Down, lo que arrojó una tasa de falsos positivos del 0,2 % (95 % de IC, $<0,1-0,6$). Para 13 mujeres (13/1696 o 0,8 %; 95 % de IC, 0,4-1,3), no se proporcionó interpretación debido a fallos en el control de calidad en las muestras iniciales y repetidas (seis tenían fracciones fetales <4 %, una >50 %), aunque los resultados de sus pruebas estaban disponibles y por lo general eran "normales" (véase la Fig. 2B). Los resultados de laboratorio, el manejo de muestras y los resultados del embarazo para los embarazos mal clasificados se revisaron exhaustivamente para detectar posibles errores; no se identificó ninguno (véase la Tabla 3, ejemplo 2). Se realizó el análisis de las primeras 15 covariables frente a la puntuación z (véanse las Figs. 7-10, ejemplo 2). Existía una fuerte asociación negativa para el peso materno entre los casos; esta asociación fue más débil en los controles. Hubo una asociación positiva pequeña, pero significativa, con la edad gestacional en los casos (véase la Fig. 7, ejemplo 2), con puntuaciones z regresivas a las 11 y 19 semanas de gestación de 7,2 y 9,9, respectivamente. Otras asociaciones fueron pequeñas y por lo general no significativas.

Confirmación por parte de un laboratorio independiente del rendimiento de las pruebas

Un laboratorio universitario independiente (por ejemplo, UCLA) realizó la generación de grupos, la secuenciación de ADN y la interpretación de un subconjunto de 605 alícuotas de muestras iniciales procesadas y analizadas originalmente por el SCMM. Este subconjunto se seleccionó al azar por el centro de coordinación de todos los grupos completos de 92 muestras de pacientes (por ejemplo, placas). Se analizaron con éxito un total de 578 muestras en ambos sitios (96 %). Los resultados de la MPSS interpretados por ordenador se expresan como puntuación z , con valores de SCMM. Un total de 77 embarazos con síndrome de Down y 501 euploides fueron analizados con éxito en ambos sitios. Las 27 muestras que fallaron en la prueba inicial en uno o ambos sitios no se incluyen. Se usó un punto de corte de puntuación z de 3. Entre estas muestras, sólo se produjo un desacuerdo. UCLA clasificó erróneamente una muestra euploide (puntuación $z = 3,46$), pero el SCMM la clasificó correctamente (puntuación $z = 2,02$). Ambos grupos clasificaron erróneamente una muestra con síndrome de Down. Las correlaciones fueron altas entre los 77 embarazos con síndrome de Down y los 501 euploides (por ejemplo, $R = 0,80$ y $0,83$, respectivamente). En este subconjunto de 578, las tasas de detección, falso positivo y fallo inicial para el SCMM fueron del 98,7 %, 0,0 % y el 4,4 %, respectivamente. Las tasas correspondientes para UCLA fueron del 98,7 %, 0,2 % y el 3,9 % (véase la Tabla 3, ejemplo 2). En otro subconjunto de 56 incluidas, cada laboratorio analizó muestras de plasma de 4 ml por duplicado. Una muestra euploide falló en ambos sitios debido a la baja fracción fetal. Dos muestras euploides adicionales fallaron en la secuenciación en UCLA; su protocolo no permitía volver a realizar la prueba. Las tasas de fallo en el SCMM y UCLA fueron del 1,8 % y el 5,3 %, respectivamente. Entre las 53 muestras restantes, los dos sitios coincidieron en

todos los parámetros de calidad y resultados interpretativos (ejemplo 2). En ambos laboratorios, las tasas de detección y falsos positivos fueron del 100 % y el 0 %, respectivamente.

Análisis post hoc

El gran tamaño de la muestra brindó la oportunidad de investigar métodos alternativos para interpretar los resultados de la MPSS. Después del registro, pero antes de que el laboratorio revelara la ocultación, el laboratorio de SCMM ajustó los resultados porcentuales del cromosoma 21 para el contenido de GC, un proceso que demostró mejorar el rendimiento de la MPSS, y también se filtró con respecto a The Repeat Mask (URL www.repeatmasker.org/PreMaskedGenomes.html) y los resultados se enviaron al centro de coordinación para determinar si los algoritmos interpretativos alternativos podrían funcionar mejor, ser más fuertes, o ambos. El análisis mostró que los resultados del control variaban según la celda de flujo o la placa (tres celdas de flujo que se procesan por lotes) (ANOVA, $F = 13,5$, $P < 0,001$), pero la D.E. era constante (ANOVA, $F = 1,2$, $P = 0,23$), lo que permitió la conversión de los resultados ajustados por GC a múltiplos de la mediana de la placa. Los múltiplos de los valores de la mediana de la placa en los embarazos con síndrome de Down y euploides estaban completamente separados, excepto por un resultado falso negativo persistente (véase el ejemplo 2). El ajuste de las puntuaciones z específicas de la celda de flujo también mejoró el rendimiento, quedando dos falsos negativos y un falso positivo (véase el ejemplo 2). Los análisis *post hoc* no estaban disponibles en el momento en que se realizaron las interpretaciones clínicas.

Implicaciones clínicas

Se analizaron dos mil ciento dieciséis muestras iniciales de pacientes (1696 informadas aquí y otras 420 muestras de pacientes) con un rendimiento de 235 pacientes por semana usando dos plataformas HiSeq 2000. El tiempo de respuesta (por ejemplo, descongelación de la muestra para registro) mejoró durante las 9 semanas de prueba, alcanzando el objetivo de 10 días para 18 de las 20 celdas de flujo finales (véase el ejemplo 2). Esto no incluye el 5 % de las muestras que requirieron una segunda alícuota, aunque el tiempo de respuesta para las muestras que requirieron una segunda alícuota no se duplicó porque los fallos a menudo se detectaban al principio del proceso de prueba.

Para evaluar la utilidad, un modelo simple (véase el ejemplo 2) compara los protocolos de diagnóstico actuales para el síndrome de Down con uno que inserta la MPSS entre la identificación del embarazo de alto riesgo y el diagnóstico invasivo. Se asumen 100.000 mujeres con alto riesgo de síndrome de Down, con un embarazo afectado por cada 32 embarazos normales, costes de pruebas de diagnóstico de 1000 \$ por paciente (véase el ejemplo 2) y una tasa de pérdida fetal relacionada con el procedimiento de 1 en 200. La aceptación completa de pruebas invasivas por parte de mujeres de alto riesgo detectaría 3000 casos a un coste de 100 millones de dólares y 500 pérdidas relacionadas con el procedimiento. La aceptación completa de la prueba MPSS por parte de todas las mujeres de alto riesgo, seguida de pruebas invasivas en aquellas con resultados positivos de MPSS (junto con aquellas que no pasaron la prueba), detectaría 2958 casos (42 perdidos) a un coste de 3,9 millones de dólares y 20 pérdidas. La diferencia en los costes financieros de los dos protocolos podría ayudar a compensar los costes de las pruebas MPSS. Asignar un valor en dólares a las 480 pérdidas relacionadas con el procedimiento potencialmente evitables es difícil, pero son una consideración igualmente importante. Si la tasa de pérdidas relacionadas con el procedimiento fuera inferior a 1 en 200, el número absoluto de pérdidas disminuiría, pero la reducción proporcional seguiría siendo la misma.

Análisis

Se han informado un total de 350 embarazos con síndrome de Down y 2061 embarazos de control, incluidos los informados en el presente documento. El total informado de embarazos con síndrome de Down y de control documenta una sensibilidad y especificidad del 99,0 % (por ejemplo, 95 % de IC, 98,2-99,8 %, $I^2 = 0$ %; véase la Tabla 5, ejemplo 2), que proporciona evidencia definitiva de la validez clínica de una prueba para el síndrome de Down basada en MPSS. Un resultado positivo a veces aumentaba el riesgo de síndrome de Down 490 veces (por ejemplo, 98,6 % de detección/0,2 % de tasa de falsos positivos), y un resultado negativo a veces reducía el riesgo 72 veces (por ejemplo, 99,8 %/1,4 %). La prueba tuvo éxito en 992 de cada 1000 mujeres. Aunque el 5,3 % de las pruebas iniciales fallaron en los controles de calidad, el 82 % de estos se resolvieron después de ensayar las segundas alícuotas. Los fallos restantes de las pruebas a menudo se asociaron con una fracción fetal baja, que a veces se puede resolver repitiendo el muestreo una o dos semanas más tarde en el embarazo. El rendimiento de la MPSS se confirmó por un laboratorio independiente (por ejemplo, véase la Tabla 5 en el ejemplo 2) usando muestras de plasma originales y preparaciones de ADN plasmático.

El estudio actual manejó un gran número de muestras (recogida, procesamiento, congelación y envío) en 27 sitios de inclusión; simulación de la práctica clínica esperada. Estos hallazgos respaldan el rendimiento de la MPSS en un amplio rango de edad gestacional, entre diversos grupos raciales/étnicos, para todas las edades maternas y para todas las indicaciones de pruebas de diagnóstico (véase el ejemplo 2). El rendimiento no se ve afectado por el sangrado vaginal o la hemólisis de la muestra y es resistente a un tiempo de procesamiento de muestras de hasta 6 horas. Debido al ya descrito efecto de dilución del aumento del volumen de sangre, los 15 fallos en la prueba son más comunes en mujeres más con más peso. Puede estar justificado tener en cuenta la fracción fetal en la interpretación. En general, la mayoría de las mujeres con resultados de cribado falsos positivos evitarán las pruebas

invasivas, mientras que casi todos los embarazos afectados serán diagnosticados con confianza por medios invasivos convencionales. El presente estudio apoya ofrecer la MPSS a mujeres identificadas con alto riesgo de síndrome de Down, teniendo en cuenta la complejidad de la prueba y los recursos requeridos. Si las pruebas se realizaran al menos dos veces por semana, el tiempo de respuesta para el 95 % de los resultados de las pacientes sería comparable con el actualmente disponible para el análisis citogenético de las células del líquido amniótico y el muestreo de vellosidades coriónicas. La disponibilidad de la MPSS también podría justificar la reducción de los puntos de corte de cribado de suero/ultrasonido, lo que daría como resultado una mayor detección del síndrome de Down. Este estudio documenta, por primera vez, una variabilidad inherente entre celdas de flujo. Tener en cuenta estos cambios mejora el rendimiento clínico. La mejor manera de realizar tales ajustes necesita más estudio.

Los análisis *post hoc* dieron como resultado una reducción de los resultados falsos negativos y falsos positivos, principalmente debido a los ajustes del contenido de GC. Esto constituye una fuerte evidencia de que el rendimiento de la MPSS será mejor cuando las pruebas se introduzcan en la práctica. Este estudio también proporciona evidencia de que la MPSS se puede trasladar de la investigación a un entorno clínico con una respuesta y un rendimiento razonables. Determinados problemas de implementación merecen atención. Sería útil tubo de recogida que permita el almacenamiento y envío a temperatura ambiente sin afectar a los niveles de ADN libre de células. Actualmente, las muestras deben procesarse, congelarse y enviarse en hielo seco, de manera similar al protocolo seguido en este estudio. Dado que se trata de un estudio observacional, se justifica un proyecto de demostración que muestre la eficacia en entornos clínicos. Es necesario desarrollar y validar materiales educativos tanto para pacientes como para proveedores para ayudar a garantizar una toma de decisiones informada. Otras preocupaciones incluyen el reembolso y el desarrollo de pautas profesionales relevantes. Algunos han sugerido que las pruebas de ADN fetal plantean nuevas cuestiones éticas. En el entorno recomendado de la prueba MPSS de mujeres de alto riesgo, muchas de estas cuestiones no son relevantes.

Un objetivo principal en el campo del cribado prenatal ha sido reducir la necesidad de procedimientos invasivos. Las pruebas MPSS aún no pueden considerarse diagnósticas. Sin embargo, ofrecer las pruebas MPSS a mujeres que ya tienen un alto riesgo de síndrome de Down puede reducir las pérdidas relacionadas con el procedimiento hasta en un 96 %, manteniendo al mismo tiempo una alta detección. Todavía se necesita la confirmación mediante pruebas invasivas. Este estudio, junto con informes anteriores, documenta un alto rendimiento, pero se amplía la evidencia al realizar las pruebas en un laboratorio con certificado CLIA, tener segundas alícuotas disponibles para fallos iniciales, monitorizar el tiempo de respuesta, evaluar la variabilidad de operador a operador y de máquina a máquina, validar un subconjunto de resultados de muestras en un laboratorio clínico académico independiente, e integrar a un genetista médico/director de laboratorio en el proceso de elaboración de informes. Este informe no aborda otras anomalías cromosómicas 13 o eventos tales como embarazos gemelares. A medida que la tecnología avanza, estarán disponibles dichas mejoras. Aunque aún deben abordarse algunos problemas de implementación, la evidencia justifica la introducción de esta prueba sobre una base clínica para las mujeres con alto riesgo de síndrome de Down, antes de las pruebas diagnósticas invasivas.

Ejemplo 2: Determinación de la presencia o ausencia de una variación genética usando muestras ciegas: materiales, métodos y resultados adicionales.

Integridad del estudio

El Comité de Supervisión del estudio se creó en febrero de 2009 para ayudar a garantizar la independencia e integridad continuas del estudio. La composición del comité fue diseñada para representar a la comunidad académica de obstetricia y genética, con experiencia en los aspectos clínicos y de laboratorio de las pruebas prenatales y los métodos de genética molecular. El Comité se reunió con los Coinvestigadores principales (Co-PI) del estudio, ya sea en persona o por teléfono, un promedio de tres veces al año durante 2009 y 2010, completó su misión y realizó su última conferencia telefónica con el fin de la inclusión activa al estudio en febrero de 2011. Los miembros del comité optaron por no firmar acuerdos de confidencialidad con el patrocinador del estudio (Sequenom) para no tener conocimiento de métodos patentados o resultados y no interactuar directamente con el personal de Sequenom durante el curso del estudio. El aporte del Comité de Supervisión fue esencial para implementar 1) métodos seguros en la codificación y selección de muestras para las pruebas, 2) la verificación provisional de los resultados de la prueba, y 3) reglas para mantener la separación entre el patrocinador del estudio y el centro de coordinación y las actividades del sitio de inclusión.

Las inspecciones de cada sitio de inclusión por parte de un co-PI o coordinador del estudio incluyeron una visita *in situ* para revisar y evaluar el cumplimiento de los procedimientos, examinar el espacio de trabajo y los recursos, validar los datos enviados y responder preguntas sobre los objetivos, métodos y plazos del estudio. Se generaron resúmenes de cada inspección, firmados por el IP del estudio en particular y el IP del sitio de inclusión, y se enviaron al patrocinador del estudio copias que no contenían identificadores ni datos de pacientes. Los sitios de inclusión no se comunicaron directamente con el patrocinador del estudio y un laboratorio independiente analizó una proporción de las muestras.

También se implementaron procedimientos para garantizar que los datos sin procesar no pudieran modificarse sin detección, y que todos los resultados sin procesar pudieran volver a analizarse en un laboratorio independiente. La ocultación de los resultados de las pruebas de diagnóstico se logró en dos niveles. En el centro de coordinación, las

muestras y la información demográfica se almacenaron en Rhode Island, mientras que los datos de resultados se almacenaron en una segunda sucursal del centro de coordinación (por ejemplo, en Maine), para combinarlos con los datos demográficos en el momento adecuado. Esta información no era accesible desde ubicaciones remotas ya que el servidor no estaba conectado a Internet.

5 Centro de coordinación

El Woman & Infants Hospital (WIH) actuó como centro de coordinación y tuvo la responsabilidad global del estudio. Las responsabilidades incluían implementar y adherirse al diseño del estudio, reclutar y establecer comunicaciones con los sitios de inclusión, mantener la seguridad de la base de datos y el sitio web del estudio, recopilar y verificar los datos de los pacientes, mantener el banco de muestras de plasma procesado y organizar y utilizar el Comité de Supervisión. El centro estaba ubicado en dos sitios, uno en Standish, ME, en donde los datos computarizados se mantuvieron bajo la supervisión de un Co-PI y un coordinador del estudio, y otro en Providence, RI, en donde se recibieron muestras de los sitios de inclusión, se almacenaron a -80 °C y se enviaron a los laboratorios de prueba según fuera necesario, y en donde se ubicaba el apoyo administrativo y de suministro para los sitios de inclusión. El estudio se administró por el WIH según las directrices federales. Se firmó un acuerdo de confidencialidad entre el WIH y el patrocinador del estudio, lo que permitió a los Co-PI acceder a datos provisionales y resultados de investigación durante todo el estudio.

20 Sitios de inclusión

Se buscaron preferiblemente sitios que ofrecieran servicios a un gran número de pacientes, cribado integrado o pruebas diagnósticas del primer trimestre. Los 27 sitios de inclusión participantes (véase la Tabla 1, ejemplo 1) proporcionaron pruebas diagnósticas para el síndrome de Down (u otras aneuploidías autosómicas) al final del primer y/o principios del segundo trimestre. Todos tenían la capacidad de recoger, procesar, almacenar y enviar muestras de plasma según un estricto protocolo. Los sitios obtuvieron la aprobación de la junta de revisión institucional (o equivalente) y obtuvieron el consentimiento informado de cada mujer que se incluyó en el estudio.

30 Sitios de laboratorio

El Sequenom Center for Molecular Medicine en San Diego (SCMM-SD) tiene certificación CLIA como laboratorio de genética molecular de alta complejidad. El laboratorio tiene dos secuenciadores de próxima generación Illumina HiSeq 2000, ambos usados en este estudio. El Orphan Disease Testing Center en la University of California, Los Angeles School of Medicine (UCLA), también un laboratorio de genética de alta complejidad con certificación CLIA, tenía una plataforma Illumina HiSeq 2000 durante este estudio. UCLA colaboró con el SCMM-SD en la realización de secuenciación masivamente paralela de muestras de estudio ciegas y proporcionó interpretaciones clínicas según un protocolo escrito estandarizado, actualizado para su uso en la plataforma Illumina HiSeq 2000, creada en el SCMM-SD.

40 Población de estudio

La información sobre las mujeres embarazadas que tenían programadas pruebas de diagnóstico se revisó en cada sitio de inclusión para identificar aquellas con un alto riesgo de aneuploidía según los criterios del estudio, y cuyos fetos tenían 21 semanas y 6 días de gestación o menos. El alto riesgo se definió como una prueba de cribado positiva para el síndrome de Down u otra trisomía mediante pruebas de suero y/o ultrasonido, edad materna de 38 años o más en el momento del parto (durante la primera parte del estudio se fijó en 40 años o más), o antecedentes familiares de aneuploidía. Las mujeres que eran aptas fueron informadas sobre el estudio por asesores genéticos o médicos y firmaron un consentimiento informado si eligieron participar. La firma de cada mujer y el formulario de consentimiento completo se almacenaron localmente. Se obtuvo información demográfica y relacionada con el embarazo seleccionada en un formulario estandarizado, junto con al menos dos (y hasta cinco) tubos de sangre venosa con tapón morado de 10 ml, extraídos antes del procedimiento de diagnóstico. Las participantes fueron identificadas únicamente por un código de estudio en los formularios de datos y en los tubos de plasma procesados. Los embarazos con gestaciones múltiples y muertes fetales existentes fueron aptos, siempre que se planificaran pruebas de diagnóstico para todos los fetos.

55 Análisis de potencia

El estudio pretendía determinar si la práctica existente debería cambiar. Por lo tanto, se necesitaba un alto nivel de confianza para estimar tanto la tasa de detección (proporción de embarazos con síndrome de Down con una prueba positiva o sensibilidad) como la tasa de falsos positivos (proporción de embarazos no afectados con una prueba positiva o 1-especificidad). Bajo el supuesto de que no hay falsos negativos, se deben incluir suficientes casos para tener al menos un 80 % de potencia para encontrar una tasa de detección significativamente superior al 98 %. El análisis de 200 casos proporcionaría una potencia del 90 % para rechazar este límite inferior. Para cada uno de estos casos, se seleccionarían siete embarazos euploides (controles) para garantizar una confianza razonable en la tasa de falsos positivos.

Recopilación de muestras/datos

Las muestras de plasma se extrajeron antes de la amniocentesis o del muestreo de vellosidades coriónicas y se procesaron según el protocolo de Ehrich *et al.*, (Am.J.Obstet.Gynecol.(2011) 204:205.e1-11). Brevemente, se centrifugaron tubos de plasma de 10 ml (que contenían EDTA, tapón morada) a 2500 x g durante 10 minutos a 4 °C, el plasma se reunió en un tubo de centrifuga de 50 ml y se centrifugó a 15.500 x g durante 10 minutos a 4 °C. A continuación, el plasma se transfirió a dos o más tubos cónicos de 15 ml, 4 ml por tubo, conteniendo el último tubo cualquier volumen residual. Estos tubos se colocaron en un congelador a -70 °C o menos para un almacenamiento a largo plazo en el sitio de inclusión o a -20 °C durante no más de 24 horas antes del envío en hielo seco para la entrega de 1 a 2 días al centro de coordinación. Si se almacenaron a -80 °C, las muestras se enviaron en lotes en hielo seco, generalmente mensualmente, para entrega de 1 a 2 días al centro de coordinación. Todos los tubos de plasma se identificaron mediante una etiqueta de código de barras preimpresa con la identificación del estudio específica del sitio adherida. Se usó Quick International Courier, Inc. para envíos internacionales para garantizar un seguimiento adecuado, el mantenimiento de hielo seco en los paquetes y la entrega.

Se usó un formulario estandarizado de varias partes para la recopilación de datos e incluyó una etiqueta del estudio con código de barras preimpresa, fecha de recopilación, edad gestacional, edad materna, peso, raza y etnia, indicación para el procedimiento, número de fetos, sexo del feto, fecha y hora de extracción de la muestra, número de tubos extraídos, hora de recepción en el laboratorio y hora de colocación en el congelador. Se conservó una copia en el sitio, mientras que la otra se envió con las muestras al centro de coordinación. Para obtener información del cariotipo, se generó un formulario de solicitud electrónico para cada mujer, en donde cada formulario de solicitud incluía: fecha del procedimiento, edad gestacional, procedimiento (por ejemplo, amniocentesis, CVS), prueba de diagnóstico (por ejemplo, cariotipo, qfPCR), el resultado de la prueba interpretado (así como el sexo del feto), y espacio suficiente para incluir resultados para fetos adicionales y comentarios. Tanto para los tubos de plasma procesados como para los formularios de datos, las participantes fueron identificadas solo por un código de estudio.

Selección de muestras para análisis

Los criterios de selección incluyeron el acceso a una muestra completa procesada de 4 ml, la edad de la mujer de al menos 18 años y la ausencia de falta de datos importantes o falta limitada de datos importantes. Los últimos casos incluidos del final del primer trimestre (<14 semanas de gestación) y principios del segundo trimestre (15-22 semanas de gestación) no se incluyeron porque el objetivo de 100 casos por trimestre se había alcanzado con un margen razonable. El emparejamiento se basó en la edad gestacional, la raza materna, el origen étnico de la madre, el sitio de inclusión y el tiempo en el congelador. Las muestras se enviaron en hielo seco para su procesamiento y análisis, solo después de que la prueba desarrollada en el laboratorio (LDT) hubiera pasado por la validación interna final, se hubiera presentado una publicación y el consentimiento del Comité de Supervisión. En determinadas circunstancias (por ejemplo, alícuota rota, extracción fallida), se podría solicitar una segunda alícuota. Se realizó un seguimiento del número de segundas alícuotas e indicaciones de envío.

Pruebas de laboratorio

Preparación de bibliotecas

El ADN circulante libre de células (clc) extraído se usó para la preparación de bibliotecas sin fragmentación adicional ni selección de tamaño; el ADNclc generalmente se fragmenta de forma natural con una longitud promedio de aproximadamente 160 pares de bases. Se almacenaron 55 µl de eluyente de ADN a 4 °C en tubos Eppendorf de baja unión después de la extracción hasta que comenzó la preparación de la biblioteca. Los tiempos de almacenamiento variaron de 24 a 72 horas. La preparación de bibliotecas se realizó según las especificaciones del fabricante (Illumina), con algunas modificaciones como se indica en el presente documento. Las enzimas y los tampones se obtuvieron en Enzymatics, MA (mezcla de reparación de extremos-LC; mezcla de dNTP (25 mM cada uno); Exo(-) Klenow polimerasa; tampón azul 10x; dATP 100 mM; ADN ligasa T4; tampón de ligadura rápida 2x) y New England Biolabs, MA (Phusion PCR MM). Los oligonucleótidos adaptadores, los oligonucleótidos indexadores y los cebadores de PCR se obtuvieron en Illumina Inc, CA.

La preparación de la biblioteca se inició tomando 40 µl de ADNclc para la reparación de extremos, conservando 15 µl para el control de calidad (QC) del ensayo cuantificador fetal (FQA). La reparación de extremos de la muestra se realizó con una concentración final de tampón de reparación de extremos 1X, 24,5 pM de cada dNTP y 1 pl de mezcla de enzimas de reparación de extremos. La reacción de reparación final se realizó a temperatura ambiente durante 30 minutos y los productos se limpiaron con columnas Qiagen Qiaquick, eluyendo en 36 µl de tampón de elución (EB). La monoadenilación 3' de la muestra de extremos reparados se realizó mezclando la muestra de extremos reparados con una concentración final de tampón azul 1x, dATP 192 µM, y 5LI de Exo(-) Klenow polimerasa. La reacción se incubó a 37 °C durante 30 minutos y se limpió con columnas Qiagen MinElute, eluyendo los productos en 14 µl de EB. Los adaptadores se ligaron a los fragmentos mediante incubación durante 10 minutos a temperatura ambiente con tampón de ligadura rápida 1x, oligos adaptadores de índice PE 48,3 nM y 600LI de ADN ligasa T4. La reacción de ligadura se limpió con columnas QiaQuick y la muestra se eluyó en 23 µl de EB. La muestra modificada con adaptador se enriqueció mediante amplificación con una polimerasa de alta fidelidad. Los 23 µl de eluyente de cada muestra se

mezclaron con cebadores 1x Phusion MM, Illumina PE 1.0 y 2.0 y 1 de 12 cebadores indexadores para un volumen total de reacción de PCR de 50 µl. Los métodos y procesos descritos en el presente documento no se limitan al uso de 12 cebadores indexadores. Se puede usar cualquier cantidad de cebadores indexadores adicionales con los métodos y procesos descritos en el presente documento, dependiendo de la plataforma y/o la disponibilidad del fabricante.

5 Cuanto mayor sea el número de cebadores indexadores, mayor será el número de muestras que se pueden ejecutar en un carril de celda de flujo. Los métodos y procesos descritos en el presente documento utilizaron cebadores indexadores disponibles comercialmente en el momento del estudio.

La muestra se amplificó en un tubo de PCR de 0,65 ml usando un termociclador AB GeneAmp PCR System 9700. Las condiciones de PCR utilizadas para la amplificación incluyeron una desnaturalización inicial a 98 °C durante 30 segundos, 15 ciclos de desnaturalización a 98 °C durante 10 segundos, hibridación a 65 °C durante 30 segundos y extensión a 72 °C durante 30 segundos. Una extensión final a 72 °C durante 5 minutos estuvo seguida por una retención a 4 °C. Los productos de PCR se limpiaron con columnas MinElute y las bibliotecas se eluyeron en 17 µl de EB.

Control de calidad de la biblioteca de secuenciación (LabChip GX)

Las bibliotecas se cuantificaron mediante separación electroforética en una plataforma de microfluído. Cada biblioteca se diluyó 1:100 y se analizó por triplicado usando el instrumento Caliper LabChip GX con HT DNA 1K LabChip, v2 y el kit de reactivos HiSens (Caliper Life Sciences, Hopkinton, MA). Las concentraciones se calcularon mediante el software Caliper LabChip GX v2.2 mediante análisis de frotis de 200-400 pb.

Agrupación y secuenciación

La agrupación y la secuenciación se realizaron según los protocolos estándar de Illumina. Las bibliotecas individuales se normalizaron a una concentración de 2 nM y a continuación se agruparon en formato de 4 plex a una concentración de carga de celda de flujo final de 1,2 pM por muestra o 4,8 pM por carril de celda de flujo. Se usaron el instrumento cBOT y los kits de reactivos v4 Single-Read cBOT. Se realizaron 36 ciclos de secuenciación multiplexada de lectura única en el HiSeq 2000 usando kits de reactivos de secuenciación HiSeq v1 y kits de cebadores de secuenciación multiplex complementarios. El análisis de imágenes y la identificación de bases se realizaron con el software RTA1.7/HCS1.1 de Illumina. Las secuencias se alinearon con el genoma de referencia humano UCSC hg19 (sin repetición enmascarada) usando CASAVA versión 1.6. La agrupación y la secuenciación también se pueden realizar usando 8 plex, 12 plex, 16 plex, 24 plex, 48 plex, 96 plex o más, según la disponibilidad de cebadores indexadores únicos.

Análisis de datos

Para la clasificación de muestras como cromosoma 21 trisómico frente a disómico, se utilizó un método similar al descrito en Chiu *et al.*, (BMJ (2011) 342:c7401) y Ehrich *et al.*, (Am.J.Obstet.Gynecol.(2011) 204:205.e1-11), cuyo contenido completo se incorpora en el presente documento por referencia en su totalidad. A diferencia de los métodos usados para estos estudios, la clasificación aplicada en el presente documento se realizó “en línea” para simular la práctica clínica. Se solicitaron muestras tan pronto como se procesó una celda de flujo. Esta versión “en línea” de las predicciones de clasificación usó todos los datos asociados con una celda de flujo para establecer una representación cromosómica estandarizada (por ejemplo, una puntuación z robusta de celda de flujo o puntuación z robusta de FC), usando estimaciones robustas de la ubicación y la escala de la representación cromosómica. Representando chr_i la representación cromosómica del cromosoma i ,

$$chr_i = \frac{recuentos_i}{\sum_{j=1}^{22} recuentos_j}$$

en donde $recuentos_j$ es el número de lecturas alineadas en el cromosoma j , la ecuación de la puntuación z robusta del cromosoma de FC para la muestra N asociada con el cromosoma i es

$$z_N = \frac{chr_{i_N} - Mediana(chr_i)}{MAD(chr_i)}$$

Se usó una forma normalizada de la mediana de desviación absoluta (MAD) para una estimación robusta de la escala,

$$MAD(X) = \frac{1}{\Phi^{-1}(3/4)} \cdot Mediana(|X - Mediana(X)|),$$

con la constante multiplicativa elegida para aproximar la desviación estándar de una variable aleatoria normalmente distribuida. Las muestras se denominaron trisómicas con respecto al cromosoma 21 si $z_N > 3$ y disómicas en otro caso.

Filtrado de regiones repetidas y normalización de GC

En el genoma humano, las secuencias genómicas repetidas que se pueden inferir con los métodos de detección actuales representan hasta la mitad de todo el genoma. Estas regiones repetitivas pueden adoptar la forma de repeticiones simples o repeticiones en tándem (por ejemplo, ADN satélite, minisatélite, microsatélite que se encuentra principalmente en los centrómeros y telómeros de los cromosomas), o duplicaciones segmentarias y repeticiones intercaladas (por ejemplo, SINE, LINE, transposones de ADN). El tamaño de dichas duplicaciones puede variar desde unos pocos pares de bases (pb) hasta cientos de pb y hasta 10-300 kilopares de bases. Se cree que la naturaleza repetitiva de estas regiones es una fuente de variación en la etapa de amplificación por PCR que está presente en algunas de las técnicas de secuenciación de próxima generación, por ejemplo, la secuenciación de escopeta masivamente paralela.

Para evaluar el impacto de las lecturas mapeadas en tales regiones repetitivas en la precisión de la clasificación, todas las muestras se analizaron con o sin dichas lecturas incluidas en la tabulación de la representación cromosómica. Las muestras se analizaron con o sin el beneficio de eliminar la contribución de secuencias genómicas repetidas. Para un procesamiento computacional eficiente, el genoma de referencia usado para la alineación de las lecturas cortas no era una versión con “enmascaramiento de repetición”, sino una que incluía tales regiones repetitivas. Después de la alineación, se utilizó un procedimiento de filtrado basado en la información contenida en la Biblioteca de repetición 20090604 (URL www.repeatmasker.org). Para la clasificación con reconocimiento de enmascaramiento de repetición, solo se consideraron para la estimación de la representación cromosómica las lecturas que no se solapaban con las regiones repetidas.

El diferente contenido de GC de las secuencias genómicas a veces conduce a una eficiencia de amplificación diferente durante las etapas de la PCR, lo que a su vez puede conducir a veces a un muestreo sesgado del material genómico original. Para compensar este posible sesgo de amplificación, los recuentos de cada bin de 50 Kb se resumieron y se normalizaron aún más con respecto al contenido de GC específico del bin mediante el uso de una técnica LOESS similar a la descrita en Alkan *et al.* (Nat. Genet.(2009) 41:1061-1067) Los recuentos filtrados normalizados con respecto al sesgo de GC estimado se usaron a continuación para la determinación de la representación cromosómica.

Los procedimientos de normalización de recuento y filtrado de lectura descritos en el presente documento no se usaron para la clasificación “en línea” de la ploidía del cromosoma 21, sino que se usaron como parte de un análisis posterior, y los conjuntos de datos para todas las muestras se entregaron por el SCMM al centro de coordinación antes de revelar la ocultación. La representación cromosómica calculada después de aplicar tanto el filtrado con respecto a la máscara de repetición como los procedimientos de normalización de GC se denominan en este estudio “representación cromosómica ajustada por GC”, las puntuaciones z calculadas a partir de dicha representación cromosómica se denominan “puntuaciones z ajustadas por GC”.

El laboratorio del SCMM-SD realizó todas las etapas para las 1640 muestras. El laboratorio de UCLA recibió preparaciones de bibliotecas para aproximadamente el 40 % de estas muestras y a continuación completó el protocolo de prueba. Para un conjunto de muestras (por ejemplo, 1 placa; 3 celdas de flujo; aproximadamente 96 muestras) que contenían siete casos de síndrome de Down y controles, se enviaron muestras de plasma procesadas de 4 ml separadas a los laboratorios del SCMM-SD y UCLA y la LDT completa se realizó por duplicado. Para cualquier muestra con resultados de prueba de ambos laboratorios, el resultado del SCMM-SD se consideró el resultado principal.

Resultados y análisis

Los datos gráficos y en tablas presentados en el presente documento para las Figs. 4 a 19 incluyen el análisis de covariables de las fracciones fetales (porcentaje de ADN circulante libre derivado del feto) para los 212 embarazos con síndrome de Down y 1484 embarazos euploides. Para mejorar la visibilidad de los datos, los datos categóricos se “difuminaron” a la izquierda y a la derecha de la marca de verificación etiquetada. Todos los embarazos estudiados eran viables en el momento del muestreo, y todos fueron embarazos únicos verificados con resultados de pruebas de diagnóstico disponibles (por ejemplo, cariotipo). El difuminado a menudo es una fluctuación aleatoria o un ligero cambio de puntos de datos para evitar el exceso de trazado. La coordenada del eje X se varió ligeramente para permitir la visualización de puntos individuales para esa categoría, sin cambiar la vista general de la gráfica. Dado que los resultados de la prueba de fracción fetal estaban disponibles antes de la secuenciación, se usaron para determinar la idoneidad de las muestras. Las fracciones fetales aceptables estaban entre el 4 % y el 50 %, inclusive (líneas discontinuas finas horizontales en los gráficos). En la práctica clínica, las muestras fuera de este intervalo pueden considerarse inaceptables para la secuenciación. La mediana de la fracción fetal global del 14,0 % (media geométrica del 13,4 %, media aritmética del 15,0 %) se muestra en las Figs. 1 a 3 como una línea continua horizontal fina. Si la fracción fetal es inferior al 4 %, se hace difícil resolver la pequeña diferencia entre el ADN circulante de embarazos con síndrome de Down y los euploides. Los niveles más altos indican problemas potenciales con el manejo de muestras. La distribución de las fracciones fetales está sesgada a la derecha. Por esta razón, la presentación y análisis es

después de una transformación logarítmica. Para las covariables exploradas mediante análisis de regresión, solo se muestra la línea de regresión si los resultados no alcanzan la significación estadística. De lo contrario, también se muestran los límites de predicción del 95 %.

- 5 La fracción fetal se analizó según el tiempo transcurrido entre la extracción de la muestra y el almacenamiento en el congelador. Usando los resultados del análisis de embarazos euploides, las fracciones fetales esperadas durante 1, 2, 3, 4 y 5 horas de congelación serían del 13,5 %, 13,2 %, 12,8 %, 12,5 % y 12,2 %, respectivamente.

10 El sitio de inclusión evaluó el estado de hemólisis de la muestra antes de la congelación. Se usó un esquema estándar de ninguna, leve, moderada y grave. Ninguna y leve se agruparon posteriormente en una categoría “No”, con moderada y grave agrupadas en una categoría “Sí”. No hubo diferencia significativa en la fracción fetal para aquellos con hemólisis (media = 13,2 % y 13,6 % para No y Sí, respectivamente, $t = -0,46$, $p = 0,64$). Para los embarazos con síndrome de Down hubo poca o ninguna diferencia para aquellos con hemólisis (media = 15,4 % y 15,0 %, respectivamente, $t = 0,14$, $p = 0,89$).

15 No hubo una relación significativa para el porcentaje de fracción fetal (embarazos euploides), estratificado por región geográfica; (fracciones fetales medias del 13,9 %, 13,1 %, 12,8 % y el 13,4 %, de izquierda a derecha, ANOVA $F = 1,93$, $p = 0,12$) o entre los embarazos con síndrome de Down (fracciones fetales medias del 17,4 %, 15,0 %, 14,5 % y el 15,9 %, de izquierda a derecha, ANOVA $F = 1,45$, $p = 0,23$).

20 No hubo una asociación significativa para el porcentaje de fracción fetal estratificado por indicación de prueba diagnóstica; (fracciones fetales medias del 13,0 %, 13,2 %, 13,4 %, 12,7 %, 13,1 %, 14,1 %, 15,6 % y el 13,3 %, de izquierda a derecha, ANOVA $F = 0,61$, $p = 0,75$) o entre los embarazos con síndrome de Down, de nuevo sin mostrar asociación (fracciones fetales medias del 14,9 %, 15,0 %, 15,6 %, 15,3 %, 14,8 %, NA, 13,0 % y el 15,7 %, de izquierda a derecha, ANOVA $F = 0,11$, $p = 0,99$).

25 Para el porcentaje de fracción fetal estratificado por sitios de inclusión con al menos 50 muestras, existe una diferencia significativa (las fracciones fetales medias varían del 10,2 % al 18,7 %, ANOVA $F = 5,59$, $p < 0,0001$) y para el mismo análisis entre los embarazos con síndrome de Down no hay una diferencia significativa (las fracciones fetales medias varían del 12,7 % al 16,9 %, ANOVA $F = 0,35$, $p = 0,97$). Esto no se explica por los diferentes pesos maternos (véase la Figs. B8), ya que el peso promedio en los cinco sitios de inclusión con las fracciones fetales más altas fue de 151 libras en comparación con 150 libras en los seis sitios con las fracciones fetales más bajas.

30 Fig. 1: El eje x muestra la edad gestacional en el momento de la extracción de la muestra. El panel superior (embarazos euploides) muestra la fracción fetal por edad gestacional. La regresión lineal no encontró una relación significativa (línea discontinua gruesa, $p = 0,23$, pendiente = -0,0024). Un análisis de los embarazos con síndrome de Down (panel inferior) encontró un resultado similar ($p = 0,10$, pendiente = 0,0084).

35 Fig. 2: El eje x muestra la edad materna en la fecha estimada de parto. El panel superior (embarazos euploides) muestra la fracción fetal por edad materna. La regresión lineal no encontró una relación significativa (línea discontinua gruesa, $p = 0,23$, pendiente = -0,0013). Un análisis de los embarazos con síndrome de Down (panel inferior) encontró un resultado similar ($p = 0,26$, pendiente = -0,0031).

40 Fig. 3: El eje x muestra el peso materno en libras en el momento de la extracción de la muestra. El panel superior (embarazos euploides) muestra la fracción fetal por peso materno de embarazos euploides. La regresión lineal encontró una relación significativa (línea discontinua gruesa, con límites de predicción del 95 % mostrados por líneas discontinuas finas, $p < 0,0001$, pendiente = -0,0026). Se encontró un resultado similar (panel inferior) para los embarazos con síndrome de Down ($p = 0,0002$, pendiente = -0,0017). Usando los resultados euploides como ejemplo, se espera que las mujeres que pesan 100, 150, 200, 250 y 300 libras tengan fracciones fetales promedio del 17,8 %, 13,2 %, 9,8 %, 7,3 % y el 5,4 %, respectivamente.

45 Hubo una disminución leve, pero significativa, en la fracción fetal para aquellas (embarazos euploides) que notificaron sangrado vaginal (media = 13,3 % y 12,3 % para No y Sí, respectivamente, $t = 2,04$, $p = 0,04$). Para el mismo análisis entre los embarazos con síndrome de Down hubo un aumento significativo para aquellas que notificaron sangrado (media = 14,7 % y 17,6 %, respectivamente, $t = -2,07$, $p = 0,04$).

50 No hubo diferencia en la fracción fetal entre los fetos euploides de sexo masculino y de sexo femenino (media del 13,4 % y 12,9 %, respectivamente, $t = 1,68$, $p = 0,094$) o entre los embarazos con síndrome de Down (media = 15,2 % y 15,3 %, respectivamente, $t = -0,05$, $p = 0,96$).

55 Los embarazos con síndrome de Down tienen una fracción fetal más alta que es estadísticamente significativa (media del 15,2 % frente al 13,2 %, $t = -4,11$, $p < 0,0001$) que los embarazos euploides. Si se utilizara como prueba de cribado para el síndrome de Down, con tasas de falsos positivos del 5 % y el 10 %, las tasas de detección correspondientes serían del 9,0 % y el 17,5 %, respectivamente. Estas corresponden a una razón de probabilidad acumulada de aproximadamente 1,8.

El análisis de covariables de la fracción fetal reveló que el peso materno era un factor significativo en la determinación de la variación genética. Con pesos promedio de 100 y 250 libras, las fracciones fetales esperadas son del 17,8 % y el 7,3 %, respectivamente. El efecto del peso materno puede explicar los efectos pequeños pero significativos encontrados para la fracción fetal frente a la raza y el origen étnico materno. El tiempo desde la extracción de la muestra hasta el almacenamiento en el congelador también tiene un efecto significativo en la fracción fetal, con tiempos más prolongados que dan como resultado fracciones fetales ligeramente más bajas. Sin embargo, el efecto observado para la extracción de muestras con respecto al almacenamiento en el congelador es sustancialmente menor que para el peso materno. Las asociaciones restantes son generalmente pequeñas y, por lo general, no significativas.

Los datos presentados gráficamente en las Figs. 4 a 6 resumen las relaciones entre la representación del cromosoma 21 (por ejemplo, porcentaje de cromosoma 21) y la variabilidad del ensayo. Por lo general, las muestras de cuatro pacientes se duplicaron por cuadruplexación en un solo carril de celda de flujo (por ejemplo, 8 carriles equivalen a 32 pacientes). Sin embargo, por lo general solo se analizaron 30 muestras de pacientes, y las posiciones adicionales tenían controles. Se procesaron juntas 92 pacientes en placas de 96 pocillos. Cada placa se procesó en 3 celdas de flujo (por ejemplo, 1 placa de muestra se procesó en 3 celdas de flujo cuando se utilizó cuadruplexación y 4 cebadores indexadores por carril). Generalmente, se agruparon 7 placas de datos para formar un lote. Cada lote contenía las muestras asignadas en orden aleatorio. Por lo tanto, los casos y controles dentro de un lote no necesariamente se procesaron en la misma placa de muestra o celda de flujo. Procesar conjuntamente los casos y los controles a veces puede subestimar la varianza total en los análisis emparejados. En las Figs. 4 a 6 se muestran los 212 resultados de síndrome de Down y todos menos 13 de los 1484 resultados euploides. En los casos en los que una muestra falló inicialmente, pero el segundo resultado fue exitoso, se muestra el segundo resultado. No se muestran aquellas muestras que no lograron producir un resultado útil en la muestra repetida. Todos los embarazos estudiados eran viables en el momento del muestreo, y todos fueron embarazos únicos verificados con resultados de pruebas de diagnóstico disponibles (por ejemplo, análisis de cariotipo).

La Fig. 4 muestra los resultados del % de C21 por celda de flujo. El porcentaje de lecturas emparejadas del cromosoma 21 dividido por el total de lecturas autosómicas se representa tanto para euploides (círculos pequeños) como para síndrome de Down (círculos más grandes) por el número de células de flujo (eje x). Cada celda de flujo puede analizar 32 muestras (en cuádruplex), lo que da como resultado de 28 a 30 muestras de pacientes junto con muestras de control (no todas las muestras de pacientes procesadas en cada celda de flujo están incluidas en este informe). Generalmente, se muestran de 20 a 25 embarazos euploides y de 2 a 7 con síndrome de Down cada vez. En algunos casos (por ejemplo, una celda de flujo con repeticiones), los números son mucho más pequeños. En general, 76 celdas de flujo contenían datos relevantes para el estudio actual, incluida la prueba de alícuotas adicionales. Las celdas de flujo se numeraron consecutivamente y las celdas de flujo faltantes se usaron para otros estudios, incluidas las pruebas en el laboratorio independiente. Se pueden ver cambios entre las celdas de flujo en el nivel medio. Además, existe una clara tendencia a que las primeras celdas de flujo estén por encima de la media euploide del 1,355 %, mientras que las últimas celdas de flujo tienden a ser más bajas. No hay diferencia en las desviaciones estándar de los resultados euploides entre las celdas de flujo. Se dibuja una línea de referencia al 1,355 %, la fracción fetal promedio global para las muestras euploides. Se puede observar la variabilidad entre celdas de flujo en los niveles medios (ANOVA, $F = 4,93$, $p < 0,001$), pero la desviación estándar es constante ($F = 1,1$, $p = 0,31$).

La Fig. 5 contiene los mismos datos que la Fig. 4, pero los datos están estratificados por placa en lugar de celda de flujo. El procesamiento se realiza en placas de 96 pocillos. Las muestras procesadas de una placa se procesan a continuación en tres celdas de flujo. La línea de referencia está en el 1,355 %. Se puede observar la variabilidad entre placas en los niveles medios (ANOVA, $F = 13,5$, $p < 0,001$), pero la desviación estándar es constante ($F = 1,2$, $p = 0,23$). En esta figura pueden observarse las mismas tendencias que eran evidentes en la Fig. 4. La reducción en la varianza global es algo menor cuando se tienen en cuenta las diferencias entre placas en comparación con las de entre celdas de flujo. Sin embargo, una vez que se tienen en cuenta las diferencias de las placas, no hay un efecto significativo para las diferencias de las celdas de flujo. Como se observa en la Fig. 4, no hay diferencia en las desviaciones estándar de los resultados euploides entre placas.

La Fig. 6 contiene los mismos datos que las Figs. 4 y 5, pero los datos están estratificados según el instrumento Illumina que se usó para la secuenciación. Se procesaron 42 y 34 placas en el Número 2 y el Número 3, respectivamente. La línea de referencia está en el 1,355 %. No hay diferencia en el porcentaje del cromosoma 21 por instrumento en embarazos euploides (medias de 1,355 y 1,354, respectivamente, $t = 2,0$, $p = 0,16$) o con síndrome de Down (medias de 1,436 y 1,438, respectivamente, $t = 0,32$, $p = 0,57$). No hay una diferencia sistemática en los resultados de % de C21 de las dos máquinas.

Se resumieron quince covariables potenciales para los 212 resultados de síndrome de Down y todos menos 13 de los 1.484 resultados euploides en comparación con la puntuación z del cromosoma 21 informado clínicamente. Todos los embarazos estudiados eran viables en el momento del muestreo, y todos fueron embarazos únicos verificados con resultados de pruebas de diagnóstico disponibles (por ejemplo, análisis de cariotipo). Una muestra de síndrome de Down tenía una puntuación z ligeramente superior a 25, pero se representó en 24,9. El intervalo de muestras euploides está entre -3 y +3. Entre los casos, se usó un nivel de corte de 3. La distribución de las puntuaciones z está sesgada hacia la derecha en los casos, pero es gaussiana en los controles. Los datos, sin embargo, todavía se

representaron en una escala lineal. El análisis de regresión en los casos fue después de una transformación logarítmica.

Todas las muestras seleccionadas para la prueba se procesaron y se almacenaron en el congelador en las seis horas posteriores a la recogida. Para la puntuación z del cromosoma 21 por tiempo desde la extracción de la muestra hasta el almacenamiento en el congelador, la regresión lineal no encuentra una relación significativa para los embarazos euploides o con síndrome de Down ($p = 0,90$, pendiente = $-0,0025$; y $p = 0,50$, pendiente = $-0,20$, respectivamente).

El sitio de inclusión evaluó el estado de hemólisis antes de la congelación. No hubo diferencias significativas en la puntuación z después de la estratificación por estado de hemólisis para ninguno de los grupos ($t = -0,01$, $p = 0,99$ y $t = -0,12$, $p = 0,90$ para embarazos euploides y con síndrome de Down, respectivamente).

No hubo una relación significativa para las puntuaciones z estratificadas por región geográfica para embarazos euploides (puntuaciones z medias de $-0,22$, $-0,14$, $-0,12$ y $-0,01$, de izquierda a derecha, ANOVA $F = 1,84$, $p = 0,14$) o entre los embarazos con síndrome de Down (puntuaciones z medias de $10,1$, $9,9$, $8,9$ y $10,2$, de izquierda a derecha, ANOVA $F = 1,00$, $p = 0,39$).

Hubo un efecto leve pero significativo para las puntuaciones z estratificadas por indicación de prueba de diagnóstico para embarazos euploides (puntuaciones z medias de $-0,15$, $-0,14$, $-0,24$, $-0,05$, $-0,11$, $0,20$, $-0,52$ y $-0,20$, de izquierda a derecha, ANOVA $F = 2,02$, $p = 0,049$), pero ningún efecto significativo para los embarazos con síndrome de Down (puntuaciones z medias de $8,9$, $9,1$, $9,7$, $9,8$, $10,0$, n/a, $10,7$ y $9,5$, de izquierda a derecha, ANOVA $F = 0,25$, $p = 0,96$).

Para la puntuación z estratificada por sitio de inclusión y sitios con al menos 50 muestras, no hay ningún efecto para los embarazos euploides (las puntuaciones z promedio varían de $-0,21$ a $0,02$, ANOVA $F = 0,57$, $p = 0,84$) o embarazos con síndrome de Down (las puntuaciones z promedio varían de $6,90$ a $12,34$, ANOVA $F = 1,45$, $p = 0,16$).

Fig. 7: El eje x muestra la edad gestacional en el momento de la extracción de la muestra. El panel superior (embarazos euploides) muestra la puntuación z por edad gestacional. La regresión lineal no encontró una relación significativa ($p = 0,79$, pendiente = $0,0023$). Un análisis de embarazos con síndrome de Down (véase el panel inferior) encontró una asociación positiva significativa con la edad gestacional ($p = 0,0023$, pendiente = $0,017$ en el logaritmo de la puntuación z).

Fig. 8: El eje x muestra la edad materna en la fecha estimada de parto. El panel superior (embarazos euploides) muestra la puntuación z por edad materna. La regresión lineal no encontró una relación significativa (línea discontinua gruesa, $p = 0,62$, pendiente = $-0,0023$). Un análisis de los embarazos con síndrome de Down (panel inferior) encontró un resultado similar ($p = 0,14$, pendiente = $-0,0046$).

Fig. 9: El eje x muestra el peso materno en libras en el momento de la extracción de la muestra. El panel superior (embarazos euploides) muestra la puntuación z por peso materno para muestras de embarazos euploides. La regresión lineal encontró una pendiente negativa significativa (línea discontinua gruesa, con límites de predicción del 95 % mostrados por líneas discontinuas finas, $p = 0,029$, pendiente = $-0,0016$). Se observa un efecto similar, pero mucho mayor, para los embarazos con síndrome de Down (panel inferior, $p = 0,0003$, pendiente = $-0,038$). Este último efecto probablemente se deba al efecto del peso materno sobre la fracción fetal (véase la Fig. 11).

No hubo diferencias significativas en las puntuaciones z según el estado de sangrado vaginal informado para los embarazos euploides (media = $-0,14$ y $-0,09$, para No y Sí, respectivamente, $t = -0,65$, $p = 0,52$). Para el mismo análisis entre los embarazos con síndrome de Down hubo un aumento significativo para aquellas que notificaron sangrado (media = $9,03$ y $11,70$, respectivamente, $t = -3,14$, $p = 0,0019$).

No hay un efecto significativo para la puntuación z estratificada por raza materna para embarazos euploides (puntuaciones z medias de $-0,14$, $-0,15$, $0,28$ y $-0,21$, de izquierda a derecha; ANOVA $F = 2,44$, $p = 0,063$) o embarazos con síndrome de Down (puntuaciones z medias de $9,55$, $8,90$, $9,63$ y $10,24$, de izquierda a derecha, ANOVA $F = 0,12$, $p = 0,95$).

No hay un efecto significativo para la puntuación z estratificada por etnia caucásica para embarazos euploides (puntuaciones z medias de $-0,16$, $-0,06$ y $0,00$, de izquierda a derecha, ANOVA $F = 1,70$, $p = 0,18$) o embarazos con síndrome de Down (puntuaciones z medias de $9,5$, $9,4$ y $11,9$, de izquierda a derecha, ANOVA $F = 0,38$, $p = 0,68$).

No hay diferencia en las puntuaciones z estratificadas por sexo del feto entre el sexo masculino y el sexo femenino para embarazos euploides (media = $-0,13$ y media = $-0,13$, respectivamente, $t = -0,04$, $p = 0,97$) o para embarazos con síndrome de Down (media = $9,25$ y media = $9,80$, respectivamente, $t = -0,85$, $p = 0,39$).

Para las puntuaciones z por tiempo de almacenamiento en el congelador, la regresión lineal no encontró una pendiente significativa para los embarazos con euploides (línea discontinua gruesa, $p = 0,72$, pendiente = $0,000057$) o síndrome de Down (panel inferior, $p = 0,25$, pendiente = $-0,0022$).

Fig. 10: El panel superior (embarazos euploides) muestra la puntuación z frente a la concentración de bibliotecas de ADN. La regresión lineal muestra una pendiente positiva estadísticamente significativa (línea discontinua gruesa, con límites de predicción del 95 % mostrados por líneas discontinuas finas, $p < 0,0001$, pendiente = 0,0034). Se observa un efecto similar pero no significativo para los embarazos con síndrome de Down (panel inferior, $p = 0,82$, pendiente = 0,0024).

La regresión lineal para la puntuación z por millones de secuencias de ADN emparejadas encuentra una pendiente positiva no significativa para embarazos euploides (línea discontinua gruesa, $p = 0,47$, pendiente = 0,0072) y para embarazos con síndrome de Down (panel inferior, $p = 0,94$, pendiente = 0,0099).

Como se señaló para el análisis de covariables de la fracción fetal, el análisis de covariables de las puntuaciones z del cromosoma 21 reveló que el peso materno también fue un factor significativo en la determinación de la variación genética, pero el efecto observado fue mayor entre los embarazos con síndrome de Down. La edad gestacional también tiene una asociación positiva significativa en algunos casos. Sin embargo, el efecto observado con la edad gestacional es significativamente menor que el observado con el peso materno. Las asociaciones restantes son generalmente pequeñas y, por lo general, no significativas.

La Tabla 3 a continuación proporciona información detallada adicional sobre seis muestras originalmente clasificadas incorrectamente por las pruebas de MPSS. En tres casos, los sujetos que se confirmaron con síndrome de Down se clasificaron inicialmente como sin síndrome de Down (véase los números de ID de muestra 162, 167 y 371), y en tres casos, los sujetos que se confirmaron como niños sanos se clasificaron inicialmente como que tenían síndrome de Down.

Tabla 3: Información detallada sobre seis clasificaciones erróneas por prueba de MPSS

	ID = 162	ID = 167	ID = 371	ID = 22	ID = 221	ID = 249
Puntuación z de T21	+0,83	+1,50	+1,57	+3,82	+4,72	+3,56
Interpretación de la MPSS	Sin SD	Sin SD	Sin SD	SD	SD	SD
Cariotipo	47, XX +21	47, XY +21	47, XY +21	46, XY	46, XX	46, XX
Confirmación	Cariotipo confirmado	Confirmado en la autopsia	Confirmado por el proveedor	“Niño sano” confirmado	“Niña sana” confirmada	“Niña sana” confirmada
Edad gestacional (sem.)	Falso neg. 9,2	Falso neg. 14,6	Falso neg. 13,0	Falso pos. 12,1	Falso pos. 10,0	Falso pos. 13,6
Edad materna (años)	42	43	40	41	33	39
Peso materno (lb)	200	165	182	125	174	185
Raza/Etnia	Blanca	Blanca	Blanca	Blanca, hispana	Blanca	Blanca
Sangrado	No	No	No	Sí	Sí	No
Razón de referencia	Edad mat. y aneuploidía hx	Edad mat. y cribado integrado	Cribado del primer trimestre	Edad materna de 38 años o más	Edad mat. y aneuploidía hx	Cribado del primer trimestre
Tiempo de procesamiento (h)	1	3	3	1	1	1
Volumen de muestra (ml)	4,0	4,0	3,8	3,9	4,0	4,0
Hemólisis	Ligera	NR	Ninguna	Ninguna	Ligera	Ligera
Fracción fetal (%)	4	7	5	19	24	11
Nota	1ª muestra fallida - bajo nivel de ADN fetal			1ª muestra fallida - alto nivel de ADN fetal		

Se analizó el turn-around time (tiempo de respuesta total - TAT) en días por celda de flujo para todo el proceso de secuenciación de escopeta masivamente paralela. Para el primer tercio de las celdas de flujo procesadas, el tiempo de respuesta total (TAT) estuvo dominado por el tiempo de interpretación por ordenador debido a las modificaciones realizadas en el algoritmo antes del registro clínico descrito en la publicación de los inventores. El proceso de registro clínico mejoró con el tiempo. Fue necesario volver a secuenciar completamente dos celdas de flujo (aproximadamente dos tercios del estudio) y esto dio como resultado un aumento del TAT. Durante las últimas 20 celdas de flujo, el TAT estuvo dentro del objetivo de 10 días para 18 (90 %). Los TAT en un entorno clínico real pueden ser algo mejores, en función de dos posibles mejoras: en el estudio actual, las muestras no se procesaron durante el fin de semana y un médico dedicado no siempre estuvo disponible para el registro en un día determinado. Se repitieron aproximadamente el 5 % de las muestras, casi duplicando el TAT para esas muestras.

La tasa de éxito/fracaso para identificar muestras euploides y con síndrome de Down dio como resultado una tasa de interpretación exitosa (92 %), así como razones de fallos de la prueba entre las 212 muestras de embarazos con síndrome de Down. Las pruebas repetidas de una nueva alícuota de estas 17 mujeres dieron como resultado que el 100 % de las muestras tuvieran una interpretación exitosa. El análisis se repitió para los 1484 embarazos euploides analizados. Un total de 13 muestras se consideraron fallos de la prueba, incluso después del análisis de una segunda alícuota. En general, la tasa de éxito en la realización de la MPSS fue del 99,2 %, y el 5 % de las muestras iniciales necesitaron una segunda alícuota.

La Tabla 4 que se presenta a continuación proporciona información detallada adicional sobre una comparación de las interpretaciones finales de la MPSS para 79 muestras de síndrome de Down y 526 muestras euploides analizadas en los laboratorios del SCMM y UCLA. Se prepararon bibliotecas mixtas para 605 muestras en el Sequenom Center for Molecular Medicine (SCMM), se ensayaron, se congelaron y a continuación se enviaron al laboratorio independiente de UCLA para volver a ensayarlas. Las tasas de detección y falsos positivos en el SCMM (98,7 % y 0 %, respectivamente) fueron ligeramente mejores, pero no significativamente, que las de UCLA (97,5 % y 0,2 %, respectivamente). Sin embargo, las tasas de fallo fueron levemente, pero no significativamente, más bajas en UCLA que en el SCMM (0 % y 2,5 % en síndrome de Down; 3,9 % y 4,4 % en embarazos euploides, respectivamente).

Tabla 4. Comparación de las interpretaciones finales de la MPSS para 79 muestras de síndrome de Down y 526 muestras euploides ensayadas en dos laboratorios.

		SCMM					
UCLA	Síndrome de Down			Euploide			Totales
	Verdadero positivo	Falso negativo	Fallo	Verdadero negativo	Falso positivo	Fallo	
Síndrome de Down							
Pos. verdadero	76	0	1				77
Falso neg.	0	1	1				2
Fallo de la prueba	0	0	0				0
Euploide							
Neg. verdadero				500	0	4	504
Falso pos.				1	0	0	1
Fallo de la prueba				2	0	19	21
Totales	76	1	2	503	0	23	605

Se analizó el impacto del ajuste de las puntuaciones de representación del porcentaje del cromosoma 21 para las condiciones experimentales basadas en el contenido de GC y en placas. El ajuste de GC redujo la presencia de valores atípicos altos (y bajos) entre los embarazos euploides, al tiempo que redujo la dispersión de los datos. Sin ningún ajuste (eje x), un límite del 1,38 % da como resultado cuatro falsos negativos y tres falsos positivos. Con el ajuste de GC, dos de los cuatro falsos negativos y los tres falsos positivos se resuelven usando el mismo punto de corte del 1,38 %. Sin embargo, se encuentran en la línea de corte uno de los resultados falsos negativos y un nuevo resultado falso positivo. La interpretación del cuarto falso negativo restante no ha cambiado. Al añadir el ajuste de la placa para crear el MoM, los tres falsos positivos y tres de los cuatro falsos negativos se resuelven potencialmente con cualquier punto de corte que se encuentre dentro del rectángulo horizontal de la zona de color gris.

Para 1.471 casos euploides y 212 casos de síndrome de Down, el uso de puntuaciones z del cromosoma 21 ajustadas para el contenido de GC y la variabilidad de la celda de flujo conduce a la resolución de dos falsos negativos y los tres falsos positivos originales mediante el valor de corte 3 de la puntuación z (equivalente al algoritmo de identificación “en línea”). Sin embargo, se genera un nuevo falso positivo.

La Tabla 5 que se presenta a continuación compara el protocolo de este estudio y los resultados con estudios publicados previamente que también usaron secuenciación masivamente paralela de plasma materno para cribar el síndrome de Down.

Características	Estudio actual	Ehrich 2011	Chiu 2011	Sehnert 2011
Multiplexación	4 plex	4 plex	2 plex ¹	NR
Síndrome de Down (N)	212	39	86	13
Euploide/sin síndrome de Down	1.484	410	146	34
Plataforma Illumina	HiSeq 2000	GAllx	GAllx	X
Realizado en un laboratorio CLIA	Sí	No	No	No
¿Práctica simulada?	Sí	No	No	No
Celdas de flujo	76	>15	>16	NR
Población de estudio	N Amer., S Amer.,	EE. UU.	Hong Kong, Países	EE. UU.

	Europa, Australia		Bajos, RU	
Edad gestacional en semanas (media, intervalo)	15 (8-22)	16 (8-36)	13 (NR)	15 (10-28)
1 ^{er} /2 ^o trimestre (%)	50/50	NR	88/12	58/42
Fallos (n/N, %)	13/1696 (<1)	18/467 (3,9)	11/764 (1,4)	0/47
Tasa de detección (%)	209/ 212 (98,6)	39/39 (100)	86/86 (100)	13/13 (100)
Tasa de falsos positivos (%)	3/1471 (0,2)	1/410 (0,2)	3/146 (2,1)	0/34 (0)
Rendimiento (muestras/semana)	250	NR	NR	NR
Volumen requerido	>3,5 ml	>3,5 ml	>2 ml	~4 ml ³
2 ^a muestra disponible	Sí	No	No	Sí
Fracción fetal estimada	Todos	Todos	Solo sexo masculino	NR
Tiempo de respuesta ² (días)	8,8 ⁴	10 ⁵	NR	NR

¹ El informe también incluye 8 plex, pero solo se muestran los resultados para 2 plex

² desde el inicio del procesamiento hasta la finalización de la secuenciación (no incluye alineación ni registro)

³ Los autores afirman que “el plasma de un solo tubo de sangre [10 ml] fue suficiente para la secuenciación”

⁴ Media de las últimas 20 celdas de flujo [32 muestras cada vez]

⁵ Los autores afirman que “cada lote [96 muestras] requirió aproximadamente 10 días desde la extracción del ADN hasta el resultado final de la secuenciación”

Ejemplo 3: Detección de microdeleciones utilizando ADN circulante libre de células

El campo del diagnóstico prenatal ha avanzado gracias a la implementación de técnicas que permiten la caracterización molecular del ADN fetal circulante libre de células (clc) aislado del plasma materno. Usando metodologías de secuenciación de próxima generación, se ha demostrado que se pueden detectar anomalías cromosómicas. La detección de la trisomía 21 se ha validado tanto analíticamente como en estudios clínicos a gran escala. Es probable que en un futuro cercano se realice una validación similar de las trisomías 13 y 18, las aneuploidías sexuales y otras anomalías cromosómicas raras.

Una faceta de las anomalías genéticas que aún no se ha abordado a fondo usando el ADN fetal clc como analito son las variaciones del número de copias (CNV) subcromosómicas.

Aproximadamente el 12 % de las personas con retraso en el desarrollo/discapacidad intelectual (RD/DI), trastorno del espectro autista (TEA) o anomalías congénitas múltiples (ACM) sin explicación han sido diagnosticadas con una CNV clínicamente relevante.

Un ejemplo de tal afección clínicamente relevante es el síndrome de delección de 22q11.2, un trastorno compuesto por múltiples afecciones que incluyen el síndrome de DiGeorge, el síndrome velocardiofacial y el síndrome de anomalía facial conotruncal. Si bien la manifestación exacta de estas afecciones varía ligeramente, cada una se ha relacionado con una eliminación heterocigota de una región rica en genes de aproximadamente 3 millones de pares de bases (pb) en el cromosoma 22, que se ha demostrado que es propensa a altos niveles de duplicaciones y microdeleciones debidas a la presencia de elementos repetitivos que permiten la recombinación homóloga. El síndrome de delección del cromosoma 22q11.2 afecta aproximadamente a 1 de cada 4000 nacidos vivos y se caracteriza por defectos cardíacos frecuentes, paladar hendido, retrasos en el desarrollo y problemas de aprendizaje.

En el presente documento se describen los resultados de las investigaciones realizadas para determinar la viabilidad técnica de detectar una CNV subcromosómica mediante la secuenciación del ADNclc del plasma materno. Se examinó el plasma materno de dos mujeres, cada una con un feto confirmado por análisis de cariotipo como afectado por el síndrome de delección de 22q11.2 y 14 mujeres con bajo riesgo de aneuploidías fetales como controles. El ADNclc de cada muestra se secuenció usando dos carriles individuales en un instrumento HiSeq2000, lo que dio como resultado una cobertura genómica de aproximadamente 4x. Se detectó una disminución estadísticamente significativa en la representación de una región de 3 millones de pb en el cromosoma 22 correspondiente al área afectada conocida en los dos casos verificados, en comparación con los controles, lo que confirma la viabilidad técnica de detectar una CNV subcromosómica por secuenciación del ADNclc del plasma materno.

Materiales y métodos

Adquisición de muestras

Las muestras se extrajeron bajo dos protocolos clínicos separados aprobados por la Investigational Review Board (IRB) ID 20091396 y Compass IRB 00462). Las dos muestras de sangre afectadas se extrajeron antes de un procedimiento invasivo. La presencia de una microdelección de 22q 11.2 se confirmó en estas muestras mediante

análisis de cariotipo en material obtenido por amniocentesis no transplacentaria. Las 14 muestras de control se extrajeron sin un procedimiento invasivo posterior, por lo que no se disponía de información de cariotipo para las muestras de control. Todos los sujetos dieron su consentimiento informado por escrito antes de someterse a cualquier procedimiento relacionado con el estudio, incluida la venopunción para la extracción de 30 a 50 ml de sangre completa en vacutainers de 10 ml secados por aspersión con EDTA-K2 (Becton Dickinson, Franklin Lakes, NJ). Las muestras se refrigeraron o se almacenaron en hielo húmedo hasta su procesamiento. En las 6 horas posteriores a la extracción de sangre, se centrifugó sangre completa materna usando un rotor basculante Eppendorf 581 OR plus a 4 °C y 2500 g durante 10 minutos y se recogió el plasma (por ejemplo, aproximadamente 4 ml). El plasma se centrifugó por segunda vez usando un rotor de ángulo fijo Eppendorf 581 OR plus a 4 °C y 15.000 g durante 10 minutos. Después del segundo centrifugado, el plasma se eliminó del sedimento que se formó en el fondo del tubo, se distribuyó en alícuotas con código de barras de plasma de 4 ml y se almacenó inmediatamente congelado a -80 °C hasta la extracción del ADN.

Extracción de ácido nucleico

El ADNcll se extrajo del plasma materno usando el kit de ácido nucleico circulante QIAamp según el protocolo del fabricante (Qiagen) y se eluyó en 55 µl de tampón AVE (Qiagen).

Ensayo cuantificador fetal

La calidad y cantidad relativas de ADNcll se evaluaron mediante un Fetal Quantifier Assay (ensayo cuantificador fetal - FQA), según métodos conocidos en la técnica. El FQA usa las diferencias en la metilación del ADN entre el ADNcll materno y fetal como base para la cuantificación. El análisis FQA se realizó en cada una de las 16 muestras analizadas como se ha descrito previamente en Ehrich *et al.* y Palomaki *et al.* (Genet Med. (2011)13(11):913-20 y Genetics in Medicine(2012)14:296-305).

Preparación de la biblioteca de secuenciación

Las bibliotecas se crearon usando una versión modificada del protocolo recomendado por el fabricante para la preparación de bibliotecas TruSeq (Illumina). Se usó ADNcll extraído (por ejemplo, aproximadamente 40 µl) como molde para la preparación de bibliotecas. Todas las bibliotecas se crearon con un proceso semiautomatizado que empleó instrumentación de manejo de líquidos (Caliper Zephyr; Caliper LifeSciences) con una etapa de limpieza basada en perlas magnéticas (Beckman Coulter) después de los procesos bioquímicos de reparación de extremos, ligadura y PCR. Dado que se ha caracterizado bien que el ADNcll existe en el plasma materno dentro de un intervalo pequeño de tamaños de fragmentos, no se realizó ninguna selección de tamaño ni en el ADNcll extraído ni en las bibliotecas preparadas. La distribución de tamaño y la cantidad de cada biblioteca se midió mediante electroforesis capilar (Caliper LabChip GX; Caliper) y cada biblioteca se normalizó a una concentración estándar de aproximadamente 2 nM antes de la agrupación usando un instrumento CBot (Illumina). Cada muestra se sometió a 36 ciclos de secuenciación por síntesis usando dos carriles de una celda de flujo HiSeq2000 v3 (Illumina).

Análisis de datos

El análisis de los datos de secuenciación se realizó como se describe en Palomaki *et al.* (Genet Med. (2011)13(11):913-20 y Genetics in Medicine(2012)14:296-305). Brevemente, todos los archivos de salida (por ejemplo, archivos .bcl) del instrumento HiSeq2000 se convirtieron al formato fastq y se alinearon con la versión de febrero de 2009 del genoma humano (hg19) mediante CASAVA v1.7 (Illumina). Todas las lecturas que se superpusieron con regiones repetitivas del genoma se eliminaron después de la alineación según la información contenida en la Biblioteca de repetición 20090604 (Localizador de recursos universal (URL) en Internet repeatmasker.org) para minimizar el efecto de las secuencias repetidas en los cálculos posteriores. Para fines de análisis, cada cromosoma se dividió en bins distintos de 50 kb y se sumó el número de lecturas mapeadas en cada uno de estos bins. Las lecturas dentro de cada bin se normalizaron con respecto al contenido de GC específico del bin usando un método LOESS, como se conoce en la técnica, para minimizar el efecto del sesgo del contenido de G/C en los cálculos posteriores. A continuación, los recuentos de lectura normalizados de GC con enmascaramiento de repetición por bin se usaron para el cálculo de la significación estadística y la cobertura.

La significación estadística se determinó calculando una puntuación z para la fracción de lecturas autosómicas alineadas totales mapeadas en la región de interés en relación con el número total de lecturas autosómicas alineadas. Las puntuaciones z se calcularon usando un método robusto mediante el cual se calculó una puntuación z para una muestra determinada mediante el uso de la fórmula $Z_{\text{Muestra}} = (\text{Fracción}_{\text{Muestra}} - \text{Mediana de fracción}_{\text{Población}}) / \text{Mediana de desviación absoluta}_{\text{Población}}$. La cobertura se calculó mediante la fórmula $\text{Cobertura} = LN/G$, en donde L es la longitud de lectura (36 pb), N es el número de lecturas normalizadas de GC con enmascaramiento de repetición, y G es el tamaño del genoma haploide con enmascaramiento de repetición.

Resultados

La secuenciación de próxima generación se realizó sobre el ADNcll aislado del plasma de las 16 mujeres embarazadas, de las cuales dos se confirmó por análisis de cariotipo después de la amniocentesis que tenían un feto

afectado por el síndrome de delección del cromosoma 22q11.2. La información del cariotipo de los fetos de las 14 muestras de control no estaba disponible. Se recogió plasma de las dos muestras afectadas a una edad gestacional similar (19 y 20 semanas) en comparación con las muestras de control (mediana = 20 semanas; véase la Tabla 6 a continuación). Antes de la secuenciación, se midió la contribución fetal al ADNcl total como se conoce en la técnica. Todas las muestras contenían más del 10 % de ADN fetal con una mediana de contribución del 18 %; las dos muestras que portaban la microdelección fetal contenían el 17 y el 18 % de ADN fetal (véase la Tabla 6 a continuación).

Tabla 6.

ID de la muestra	Grupo de muestra	Vol. de plasma (ml)	Fracción fetal	Edad gestacional (Semanas)	Lecturas de norm. de GC totales	Cobertura genómica	Fracción de norm. de GC en la región afectada
12800	Bajo riesgo	4	0,42	19	202890379	4,43	0,000755
12801	Microdelección	4	0,17	20	188214827	4,11	0,000732
12802	Bajo riesgo	3,9	0,24	24	164976211	3,60	0,000752
12803	Bajo riesgo	4	0,13	12	190397481	4,16	0,000753
12804	Bajo riesgo	4	0,16	24	175708269	3,84	0,000747
12805	Bajo riesgo	4	0,35	17	192035852	4,19	0,000755
12806	Bajo riesgo	3,9	0,13	12	189438328	4,14	0,000757
12807	Bajo riesgo	3,9	0,18	20	185562643	4,05	0,000755
12808	Microdelección	4	0,18	19	146700048	3,20	0,000726
12809	Bajo riesgo	4	0,54	21	154878242	3,38	0,000750
12810	Bajo riesgo	4	0,15	16	188121991	4,11	0,000768
12811	Bajo riesgo	4	0,16	24	172366695	3,76	0,000757
12812	Bajo riesgo	4	0,10	12	180005977	3,93	0,000751
12813	Bajo riesgo	4	0,23	25	151510852	3,31	0,000752
12814	Bajo riesgo	4	0,20	20	143687629	3,14	0,000752
12815	Bajo riesgo	3,9	0,18	12	177482109	3,88	0,000754

Cada muestra se secuenció usando dos carriles de una celda de flujo HiSeq2000, lo que dio como resultado una cobertura genómica de aproximadamente 3,1x a aproximadamente 4,4x (véase la Tabla 6 anterior). Las lecturas se agruparon usando un tamaño de bin de 50 kb y los bins se visualizaron en el cromosoma 22 para determinar las muestras de microdelección afectadas con el fin de identificar la ubicación de la microdelección de las muestras afectadas. Ambas muestras que portaban la microdelección de 22q11.2 confirmada exhibieron una reducción de la representación en esta área genómica (véase la Fig. 47). Se calcularon las puntuaciones z para cada muestra en relación con la mediana de todas las muestras para la región afectada en el cromosoma 22. Los valores correspondientes al plasma de mujeres de bajo riesgo se muestran en color negro, mientras que los valores que representan casos conocidos del síndrome de delección de 22q11.2 se muestran en color gris. La línea discontinua en -3 representa una puntuación z que es 3 veces la mediana de desviación absoluta más baja que la mediana de representación de esta región en todas las muestras analizadas y es el punto de corte de clasificación usado tradicionalmente en la detección de aneuploidía fetal.

Debido a que la ubicación exacta de la delección genómica puede variar ligeramente de un caso a otro, se elige ensayar un área de 3 millones de pares de bases ubicada entre Cr22: 19000000-22000000 (véase la Tabla 6 anterior). Se usó un método análogo al usado para la detección de aneuploidías cromosómicas para calcular la fracción de todas las lecturas autosómicas que se mapearon en la región diana. Las muestras de control contenían el 0,075 % de las lecturas ubicadas en 22q11, mientras que las muestras afectadas con la microdelección fetal conocida solo mostraban el 0,073 % de las lecturas en esta región. Para ensayar la significación estadística de esta diferencia, se calculó una puntuación z para cada muestra usando un método robusto. Ambas muestras afectadas mostraron puntuaciones z inferiores a -3 (por ejemplo, -5,4 y -7,1, respectivamente), mientras que todas las muestras de control de bajo riesgo tenían una puntuación z superior a -3 (véase la Fig. 47). Una de las muestras de bajo riesgo mostró una puntuación z superior a +3. La región genómica de 22q11 se ha asociado previamente con inestabilidad genómica y este resultado podría indicar una duplicación potencial que se ha informado anteriormente, sin embargo, debido a que la información del cariotipo no estaba disponible para las muestras de bajo riesgo, no está claro si el resultado observado está relacionado a una CNV fetal.

Análisis

Los avances recientes en el campo del diagnóstico prenatal no invasivo han permitido detectar aneuploidías fetales mediante la secuenciación del ADNcl presente en el plasma materno. Usando un enfoque similar al usado para la detección de aneuploidía, los resultados presentados en el presente confirman la viabilidad de detectar de forma no invasiva las CNV a nivel subcromosómico en un feto en desarrollo mediante la secuenciación del ADNcl correspondiente en el plasma materno. Los datos presentados en el presente documento, aunque con un pequeño

número de casos, muestran que pueden detectarse de forma fiable regiones más pequeñas que un solo cromosoma en el plasma materno, en este caso una delección de 22q11.2. Peters *et al.* (2011) informaron una delección de 4,2 Mb en el cromosoma 12 que se detectó usando una metodología similar. Peters *et al.* examinaron un solo caso de una microdelección fetal detectada a una edad gestacional tardía (35 semanas) y lo comparó con siete muestras que se sabía que eran diploides para los cromosomas 12 y 14. Por el contrario, los resultados presentados en el presente documento, que se obtuvieron antes de la publicación del estudio mencionado anteriormente, examinaron muestras afectadas a una edad gestacional más temprana (19 y 20 semanas), usando el doble de muestras afectadas y no afectadas, y detectaron una microdelección un 28 % menor (3 Mb) que la descrita previamente. Además, los resultados presentados en el presente documento utilizaron una cobertura genómica 4x para detectar con éxito la eliminación fetal de 3 Mb, que es un aumento en la cobertura de aproximadamente 20 veces con respecto a la detección de aneuploidía estándar actual. Las delecciones más pequeñas, potencialmente de hasta 0,5 Mb, o las muestras que contienen menos ADNclc fetal pueden requerir una cobertura aún mayor.

Ejemplo 4: Automatización de la preparación de bibliotecas, aumento del nivel de multiplexación y bioinformática

A continuación se proporcionan implementaciones de un conjunto de cambios en el proceso que condujeron a un aumento del triple en el rendimiento y una reducción de 4 veces en el tiempo de manipulación manteniendo la precisión clínica. Los tres cambios principales de este ensayo modificado incluyen: mayores niveles de multiplexación (de 4 plex a 12 plex), preparación de bibliotecas de secuenciación automatizada y la implementación de nuevos métodos bioinformáticos. Los resultados confirman que el protocolo produce un flujo de trabajo más simplificado que permite un mayor rendimiento al tiempo que mantiene una alta sensibilidad y especificidad para la detección de las trisomías 21, 18 y 13.

Material y métodos

Adquisición de muestras y procesamiento de sangre.

Las muestras para la evaluación inicial del ensayo de alto rendimiento (desarrollo de preparación de bibliotecas y verificación del ensayo) se recogieron bajo tres protocolos clínicos separados aprobados por la Investigational Review Board (IRB) (BioMed IRB 301-01, Western IRB 20091396, y Compass IRB 00462). Todos los sujetos dieron su consentimiento informado por escrito antes de someterse a cualquier procedimiento relacionado con el estudio, incluida la venopunción para la extracción de hasta 20 ml de sangre completa en vacutainers de 10 ml secados por aspersión con EDTA-K2 (tubos con EDTA; Becton Dickinson, Franklin Lakes, NJ) y 30 ml de sangre completa en vacutainers de 10 ml con Cell-Free DNA BCT (tubos con BCT; Streck, Omaha, NE). Las muestras extraídas en tubos con EDTA se refrigeraron o se almacenaron en hielo húmedo y se procesaron en plasma en las 6 horas posteriores a la extracción de sangre. Las muestras extraídas en tubos con BCT se almacenaron a temperatura ambiente y se procesaron en plasma en las 72 horas posteriores a la extracción de sangre. La sangre completa materna en tubos con EDTA se centrifugó (rotor basculante Eppendorf 581 OR plus), se enfrió (4 °C) a 2500 g durante 10 minutos y se recogió el plasma. El plasma con EDTA se centrifugó por segunda vez (rotor de ángulo fijo Eppendorf 581 OR plus) a 4 °C y 15.500 g durante 10 minutos. Después del segundo centrifugado, el plasma con EDTA se eliminó del sedimento que se formó en el fondo del tubo, se distribuyó en alícuotas con código de barras de plasma de 4 ml y se almacenó inmediatamente congelado a <-70 °C hasta la extracción del ADN. La sangre completa materna en tubos con BCT se centrifugó (rotor basculante Eppendorf 581 OR plus), se calentó (25 °C) a 1600 g durante 15 minutos y se recogió el plasma. El plasma con BCT se centrifugó por segunda vez (rotor basculante Eppendorf 581 OR plus) a 25 °C y 2.500 g durante 10 minutos. Después del segundo centrifugado, el plasma con BCT se eliminó del sedimento que se formó en el fondo del tubo, se distribuyó en alícuotas con código de barras de plasma de 4 ml y se almacenó inmediatamente congelado a <-70 °C hasta la extracción del ADN.

Las muestras para el desarrollo de multiplexación y la evaluación clínica se extrajeron como se ha descrito anteriormente (Palomaki GE, *et al.* (2012) Genet. Med. 14: 296-305 y Palomaki GE, *et al.* (2011)). Brevemente, se extrajo sangre completa de las pacientes incluidas antes de un procedimiento invasivo. Todas las muestras se extrajeron de mujeres embarazadas con mayor riesgo de aneuploidía fetal en su primer o segundo trimestre de gestación como parte de una colaboración internacional (ClinicalTrials.gov NCT00877292). Se obtuvo la aprobación del IRB (o equivalente) para esta colaboración en cada uno de los 27 sitios de extracción. Algunos datos generados en formato 4 plex y usados en el presente documento se han presentado previamente en el presente documento, sin embargo, todos los datos de la secuenciación 12 plex se generaron usando las mismas bibliotecas ahora secuenciadas independientemente en formato 12 plex. Además, para la confirmación independiente del método de alto rendimiento, se procesó una alícuota de plasma de cada una de las 1269 pacientes. Cada una de estas pacientes contribuyó con una alícuota de plasma distinta a los estudios publicados previamente y se conoció el cariotipo fetal. Solo se usaron muestras de embarazos únicos confirmados como trisomías simples 21, 18 y 13 o de controles euploides. El ADN circulante libre de células se extrajo del plasma materno usando el kit de ácido nucleico circulante QIAamp (Qiagen) como se describe en el presente documento. La cantidad de ADNclc se evaluó para cada muestra mediante el ensayo cuantificador fetal (FQA). Se usó ADNclc extraído (40 µl) como molde para toda la preparación de bibliotecas. Se prepararon bibliotecas para la experimentación por multiplexación aumentada inicial (12 plex) usando métodos descritos previamente. Brevemente, se extrajo el ADNclc y se prepararon bibliotecas de secuenciación usando oligonucleótidos (Illumina), enzimas (Enzymatics) y procesos de purificación manual entre cada reacción

enzimática usando métodos basados en columnas (Qiagen). Todas las bibliotecas recién creadas usadas en este estudio se crearon en formato de placa de 96 pocillos usando una versión modificada del protocolo del fabricante para la preparación de bibliotecas TruSeq (Illumina) y un proceso semiautomático que utilizó instrumentación de manipulación de líquidos (Caliper Zephyr; Caliper LifeSciences) con una etapa de limpieza basada en perlas magnéticas (AM Pure XP; Beckman Coulter) después de los procesos bioquímicos de reparación de extremos, ligadura y PCR. Dado que se ha caracterizado bien que el ADNcl existe en el plasma materno dentro de un intervalo pequeño de tamaños de fragmentos, no se realizó ninguna selección de tamaño ni en el ADNcl extraído ni en las bibliotecas preparadas. La evaluación de la distribución de tamaño y cuantificación de bibliotecas se realizó como se ha descrito previamente en el presente documento. Se agruparon y se secuenciaron doce bibliotecas de secuenciación isomolar en el mismo carril (12 plex) de una celda de flujo Illumina v3 en un Illumina HiSeq2000. La secuenciación por síntesis se realizó durante 36 ciclos seguidos de 7 ciclos para leer el índice de cada muestra. Se prepararon bibliotecas de secuenciación a partir de ADNcl agrupado aislado del plasma de dos voluntarios varones adultos diagnosticados con trisomía 21 o mujeres euploides no embarazadas. Las bibliotecas se cuantificaron y se mezclaron en dos concentraciones (4 % de trisomía 21 y 13 % de trisomía 21) para aproximar la contribución del ADNcl fetal en el plasma materno. El rendimiento de la biblioteca se probó antes de la implementación de estos controles en el estudio de evaluación clínica.

Análisis de datos

Todos los archivos de salida BCL (identificación de bases) de HiSeq2000 se convirtieron al formato FASTQ y se alinearon con la versión de febrero de 2009 del genoma humano (hg19). Dado que las bibliotecas para el desarrollo de multiplexación se prepararon de forma manual con la versión anterior de bioquímica, los métodos de análisis se aplicaron como se ha descrito previamente (Palomaki *et al.*, 2012 y en el presente documento). Para todos los estudios posteriores, las lecturas se alinearon con hg19, lo que permitió solo emparejamientos perfectos dentro de la secuencia semilla usando Bowtie 2 (Langmead B, Salzberg SL (2012) Nat. Methods 9: 357-359). Para fines de análisis, las lecturas mapeadas en cada cromosoma se cuantificaron usando histogramas estándar que comprendían segmentos genómicos de 50 kpb de longitud adyacentes, no superpuestos. Después del agrupamiento, la selección de los segmentos genómicos de 50 kpb incluidos se determinó usando un método de validación cruzada descrito previamente (Brunker AT (1992) Nature 355: 472-475). Las regiones se excluyeron de un análisis posterior en función de exhibir una alta varianza entre muestras, baja capacidad de mapeo (Derrien T, *et al.* (2012) PLoS One 7: e30377), o alto porcentaje de elementos repetitivos (Biblioteca de repetición 20090604; <http://www.repeatmasker.org>). Finalmente, las lecturas alineadas correspondientes a los segmentos genómicos restantes de 50 kpb se normalizaron para tener en cuenta el sesgo de GC (Alkan C, *et al.* (2009) Nat Genet 41: 1061-1067) y se usan para calcular la fracción de lecturas alineadas derivadas de cada cromosoma. Se calculó una puntuación z robusta como se describe usando la fórmula $Z_{\text{Cromosoma}} = (\text{Muestra de Fracción cromosómica} - \text{Celda de flujo Mediana de fracción cromosómica}) / \text{Mediana de desviación absoluta}$. La mediana de la fracción cromosómica se calculó específicamente para cada celda de flujo, mientras que la mediana de desviación absoluta (MAD) fue un valor constante derivado de una MAD estática.

Resultados

Algunos estudios clínicos que usan MPSS para la detección no invasiva de aneuploidías fetales han mostrado un intervalo de tasa de detección del 92-100 % mientras mantienen una tasa de falsos positivos de menos del 1 %. El objetivo era mantener o mejorar este rendimiento al mismo tiempo que se simplifica el protocolo y se aumenta el rendimiento de las muestras. Las mejoras se centraron en tres aspectos: I) optimizar la preparación de bibliotecas para permitir un rendimiento robusto y un mayor rendimiento, II) aumentar el número de muestras indexadas molecularmente de forma individual agrupadas en un solo carril de celda de flujo (nivel de multiplexación), y III) mejorar los métodos analíticos para la clasificación de aneuploidía.

La preparación de bibliotecas de secuenciación tradicional requiere mucho trabajo, mucho tiempo y es sensible a la variabilidad entre operadores. Para aliviar estos problemas, se desarrolló un proceso semiautomatizado que usa una plataforma de manejo de líquidos de 96 canales. La bioquímica de preparación de bibliotecas TruSeq se optimizó para la baja abundancia de ADNcl recuperado de 4 ml de plasma (10-20 ng), lo que supuso una reducción de 50 veces con respecto a la cantidad de entrada recomendada de 1 pg para el kit de preparación de bibliotecas TruSeq. Además, los procedimientos de purificación manual se reemplazaron por un proceso automatizado de purificación de micropelotas AM Pure XP optimizado para la velocidad, reproducibilidad y recuperación de ADNcl. Comparación de un conjunto de 287 bibliotecas preparadas usando este método con bibliotecas producidas usando el método manual como se describe (en el presente documento y Palomaki *et al.* 2011 y Palomaki *et al.* 2012) reveló un aumento en la mediana de concentración de bibliotecas de 124 a 225 nM después de la estandarización del volumen de elución (Fig. 11A). El proceso semiautomatizado combinado produjo 96 bibliotecas en 5 horas, requiriendo solo un técnico y 1,5 horas de mano de obra. Esto dio como resultado un aumento de 4 veces en el rendimiento que coincidió con una disminución de 4 veces en la mano de obra sin sacrificar el rendimiento o la calidad de las bibliotecas. Se prepararon 93 bibliotecas (83 muestras euploides confirmadas y 10 muestras de trisomía 21 confirmadas; Tabla 7) usando este método, se secuenciaron, se analizaron y demostraron un rendimiento de clasificación preciso en este pequeño conjunto de datos (Fig. 11B; Tabla 8).

Las bibliotecas preparadas y secuenciadas en 4 plex durante un estudio anterior se secuenciaron en 12 plex para determinar la viabilidad de una mayor multiplexación. Las celdas de flujo Illumina v3 y la bioquímica de secuenciación, en combinación con las mejoras del software HCS, produjeron un aumento de 2,23 veces (de 72 a 161 millones) en el recuento total de lecturas por carril. Se secuenciaron y se analizaron 1900 bibliotecas en 12 plex, incluidas 1629 muestras euploides, 205 muestras de trisomía 21, 54 muestras de trisomía 18 y 12 muestras de trisomía 13 (Tabla 7) y se compararon las puntuaciones z para los cromosomas 21, 18 y 13 con respecto a resultados en 4 plex (Fig. 12). Dado que estudios previos habían indicado un aumento en el rendimiento del ensayo usando un punto de corte de la puntuación z elevado, la clasificación se basó en $z = 3,95$ para los cromosomas 18 y 13. La clasificación para el cromosoma 21 permaneció en $z = 3$. Usando estos puntos de corte de clasificación, hubo un total de 7 resultados de clasificación discordantes entre la secuenciación de 4 plex y 12 plex. Para el cromosoma 21, dos muestras previamente clasificadas incorrectamente (1 falso positivo, 1 falso negativo) se clasificaron correctamente, mientras que no se detectó un verdadero positivo anotado previamente. Cuatro muestras se clasificaron erróneamente como muestras de falso positivo para el cromosoma 18 cuando previamente se habían clasificado correctamente; cada una de estas bibliotecas estaba altamente sesgada por GC. Todas las muestras fueron concordantes para la clasificación de la trisomía 13. Al secuenciar en 12 plex, se detectó el 99,3 % de las muestras aneuploides (204/205 trisomía 21, 54/54 trisomía 18 y 11/12 trisomía 13) con una tasa de falsos positivos del 0 % (0/1900), 0,26 % (5/1900) y 0,16 % (3/1900) para las trisomías 21, 18 y 13, respectivamente (Tabla 8). En general, estos datos sugieren que el rendimiento del ensayo cuando se ejecuta con multiplexación de 12 plex es similar a los resultados obtenidos previamente.

Se realizó un estudio de verificación usando el método optimizado de preparación de bibliotecas junto con la secuenciación de 12 plex (configuración de ensayo de alto rendimiento) para garantizar la integridad del proceso. Se analizaron los resultados de la secuenciación de un total de 2856 muestras, de las cuales 1269 tenían cariotipo conocido. Estas 1269 muestras clínicas estaban compuestas por 1093 muestras de euploidía, 134 de trisomía 21, 36 de trisomía 18 y 6 de trisomía 13 (Tabla 7). La mediana de la fracción de ADN fetal para las muestras fue de 0,14 (intervalo: 0,04-0,46). La mediana de concentración de bibliotecas de las bibliotecas fue de 28,21 nM (intervalo: 7,53-42,19 nM), lo que da como resultado un rendimiento total similar a otros métodos descritos en el presente documento. Finalmente, la mediana del número de lecturas autosómicas alineadas por muestra fue de 16.291.390 (intervalo: 8.825.886-35.259.563).

La comparación inicial de los datos generados a partir de las 1269 muestras con cariotipo fetal conocido con una alícuota de plasma distinta previamente secuenciada del mismo sujeto reveló una disminución en la distancia discriminadora (diferencia entre el percentil 95 de las muestras euploides y el percentil 5 de las muestras con trisomía 21) de 4,9 a 3,09 cuando se analizó usando métodos previamente establecidos que normalizan el contenido de GC y eliminan las lecturas superpuestas con regiones repetidas (por ejemplo, GCRM). Para mitigar este efecto concomitante con la disminución del tiempo de análisis global, se desarrolló un nuevo algoritmo bioinformático específico para los datos de ensayo de alto rendimiento. Estos métodos basan los cálculos para la clasificación solo en aquellos segmentos genómicos de 50 kpb con representación estable entre individuos. Cuando se aplicó al mismo conjunto de datos de alto rendimiento, la distancia discriminadora entre las muestras euploides y de trisomía 21 aumentó a 6,49. En general, los nuevos enfoques bioinformáticos dan como resultado un aumento en la distancia discriminadora entre las muestras euploides y de trisomía 21 en relación con los métodos descritos previamente.

Los resultados del ensayo de alto rendimiento se analizaron usando los nuevos métodos de análisis para 67 muestras de control y 1269 muestras de pacientes. Se secuenciaron 33 bibliotecas preparadas a partir de plasma euploide combinado (0 % de biblioteca T21), 17 bibliotecas de control que contenían el 4 % de ADN de trisomía 21, y 17 bibliotecas de control que contenían el 13 % de ADN de trisomía 21. En todos los casos, las muestras euploides combinadas tenían una puntuación z inferior a 3, mientras que las muestras de control con trisomía 21 al 4 % y al 13 % tenían una puntuación z superior a 3. A continuación, se comparó la precisión de clasificación de las 1269 muestras de pacientes con información de cariotipo conocida. Basándose en los límites de clasificación descritos anteriormente (puntuación $z = 3$ para el cromosoma 21, puntuación $z = 3,95$ para los cromosomas 18 y 13), todas las aneuploidías fetales confirmadas (134 de trisomía 21, 36 de trisomía 18, 6 de trisomía 13) se detectaron con una tasa de falso positivo del 0,08 %, 0 % y el 0,08 % para las trisomías 21, 18 y 13, respectivamente (Fig. 13; Tabla 8). Hubo una correlación positiva entre la fracción fetal y la magnitud de la puntuación z , mientras que no hay correlación entre estas métricas para las muestras euploides.

Previamente se secuenciaron distintas muestras de plasma de cada uno de los 1269 donantes y, por lo tanto, sirven como comparación del rendimiento. Para garantizar una evaluación comparable, las puntuaciones z de los estudios anteriores se calcularon usando valores de GCRM y un tamaño de población (para cálculos de mediana y MAD) de 96 muestras, equivalente al número de muestras usado para cálculos de la mediana mediante análisis de alto rendimiento. La comparación de los dos estudios reveló la clasificación correcta de una muestra de trisomía 21 de falso negativo informada previamente y una muestra de trisomía 21 de falso positivo informada previamente; sin embargo, hubo un falso positivo adicional durante este estudio (Fig. 14). No hubo muestras discordantes al comparar la clasificación de trisomía 13 y la clasificación correcta de una sola muestra de trisomía 18 con una puntuación z anterior ligeramente inferior a 3,95. La evaluación de las puntuaciones z con datos emparejados para las muestras aneuploides reveló una diferencia media de 2,19 para la trisomía 21, 1,56 para la trisomía 18 y 1,64 para la trisomía 13, lo que refleja un aumento en la puntuación z para las muestras afectadas usando los métodos de alto rendimiento.

Hubo un aumento estadísticamente significativo en la puntuación z para las muestras de trisomía 21 y trisomía 18 confirmadas usando el ensayo de alto rendimiento ($p = 4,24 \times 10^{-12}$ y $p = 0,0002$, respectivamente; prueba de Wilcoxon de datos emparejados) en relación con el estudio anterior, pero no hubo diferencias significativas en las puntuaciones z para las muestras de trisomía 13 confirmadas ($p = 0,31$; prueba de Wilcoxon de datos emparejados). No hubo diferencias estadísticamente significativas en las puntuaciones z del cromosoma 21, el cromosoma 18 o el cromosoma 13 para muestras no aneuploides ($p = 0,06$, $p = 0,90$, $p = 0,82$, respectivamente; prueba de Wilcoxon de datos emparejados). Este aumento significativo en las puntuaciones z aneuploides sin afectar significativamente a las muestras euploides indica aún más una expansión de la distancia analítica entre las muestras euploides y aneuploides para los cromosomas 21 y 18 cuando se usa la configuración de ensayo de alto rendimiento y los nuevos métodos bioinformáticos.

Análisis

El desarrollo presentado aquí fue precedido por actividades de investigación y seguido por estudios adicionales de verificación y validación realizados en un laboratorio con certificación CLIA. En total, todo el proceso de traer una nueva prueba de laboratorio desde la investigación hasta la validación estuvo respaldado por datos de más de 5000 muestras analizadas. En este estudio, se secuenciaron más de 3400 muestras durante la investigación, optimización y desarrollo. A continuación, se realizó un estudio de evaluación clínica utilizando 1269 muestras, de las cuales se detectaron las 176 muestras aneuploides manteniendo al mismo tiempo una tasa de falsos positivos del 0,08 % o menos para cada trisomía.

Se desarrolló un ensayo que permite un aumento de 4 veces en el rendimiento de preparación de bibliotecas y se combinó con un aumento de 3 veces en la multiplexación de muestras para permitir un procesamiento de muestras de ADNcl de alto rendimiento. Al usar estos métodos en combinación con datos analíticos mejorados, se mejoró la sensibilidad y la especificidad para la detección de aneuploidía no invasiva y se redujeron los requisitos técnicos y de instrumentos. En general, estos datos sugieren que el ensayo de alto rendimiento desarrollado es técnicamente sólido y clínicamente preciso, lo que permite la detección de todas las aneuploidías fetales analizadas (176/176) con una baja tasa de falsos positivos (0,08 %).

Tabla 7. Resumen de los tipos de muestra utilizados para cada uno de los estudios realizados.

Descripción del estudio	Número de muestras por cariotipo				
	Desconocido	Euploide	Trisomía 21	Trisomía 13	Trisomía 18
Optimización de bibliotecas	0	83	10	0	0
Secuenciación de 12 plex	0	1629	205	12	54
Verificación	1587	1093	134	6	36

Tabla 8. Resumen de los resultados de análisis para cada uno de los estudios realizados.

Sens. = sensibilidad; Espec. = especificidad; NA = No aplicable

Descripción del estudio	Resultados de análisis por cromosoma						Método de análisis
	Espec. del cr21	Sens. del cr21	Espec. del cr13	Sens. del cr13	Espec. del cr18	Sens. del cr18	
Optimización de bibliotecas	100	100	NA	NA	NA	NA	GCRM
Secuenciación de 12 plex	100	99,5	99,84	91,7	99,74	100	GCRM
Verificación	99,92	100	99,92	100	100	100	Nuevo

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.

La tecnología descrita de manera ilustrativa en el presente documento puede ponerse en práctica adecuadamente en ausencia de cualquier elemento no desvelado específicamente en el presente documento. Determinadas realizaciones de la tecnología se exponen en la(s) reivindicación/reivindicaciones que sigue(n).

REIVINDICACIONES

1. Un método implementado por ordenador para detectar la presencia o ausencia de una aneuploidía fetal, que comprende:
 - (a) obtener recuentos de lecturas de secuencia de nucleótidos mapeadas en secciones genómicas de referencia, en donde las lecturas de secuencia de nucleótidos se obtienen de un grupo de muestras de prueba, comprendiendo cada una ácido nucleico libre de células circulante de una mujer embarazada, en donde las muestras de prueba se secuencian bajo una o más condiciones experimentales comunes;
 - (b) para cada muestra de prueba, filtrar secciones genómicas basándose en uno o más datos redundantes, datos no informativos, datos con ruido, secciones genómicas con secuencias sobrerrepresentadas y secciones genómicas con secuencias subrepresentadas; y/o calcular una medida de error para los recuentos de lecturas de secuencia mapeadas en algunas o todas de las secciones genómicas y eliminar los recuentos de lecturas de secuencia para determinadas secciones genómicas según un umbral de la medida de error, proporcionando así secciones genómicas filtradas;
 - (c) para cada muestra de prueba, normalizar los recuentos para las secciones genómicas filtradas, o normalizar una derivada de los recuentos para las secciones genómicas filtradas, ajustando los recuentos para la variabilidad inducida por las condiciones experimentales, en donde el ajuste comprende restar los recuentos esperados para los recuentos para las secciones genómicas, generando así un valor de resta, y dividir el valor de la resta por una estimación de la variabilidad, cuyos recuentos esperados y estimación de la variabilidad se basan en la variabilidad inducida por las condiciones experimentales de los recuentos de las secciones genómicas filtradas, obteniendo así recuentos normalizados para secciones genómicas filtradas; y
 - (d) detectar la presencia o ausencia de una aneuploidía fetal para cada muestra de prueba basándose en los recuentos normalizados para las secciones genómicas filtradas.
2. El método de la reivindicación 1, en donde cada muestra de prueba es de plasma sanguíneo de la mujer embarazada.
3. El método de la reivindicación 1 o 2, en donde la aneuploidía fetal es trisomía 13, trisomía 18 o trisomía 21.
4. El método de una cualquiera de las reivindicaciones 1 a 3, en donde la medida de error en (b) es un factor R.
5. El método de una cualquiera de las reivindicaciones 1 a 4, en donde el recuento esperado es una mediana de recuento y la estimación de la variabilidad es una mediana de desviación absoluta (MAD) del recuento esperado.
6. El método de una cualquiera de las reivindicaciones 1 a 4, en donde el recuento esperado es una media recortada o truncada, una media Winsorizada o una estimación de remuestreo de tipo *bootstrapping*.
7. El método de una cualquiera de las reivindicaciones 1 a 6, en donde los recuentos se normalizan adicionalmente mediante el contenido de GC, la normalización basada en bins, GC LOESS, PERUN, GCRM o combinaciones de los mismos.
8. El método de una cualquiera de las reivindicaciones 1 a 7, en donde la una o más condiciones experimentales comunes se eligen de una unidad de celda de flujo común, celdas de flujo comunes a un recipiente, celdas de flujo comunes a un lote o proceso de fabricación, una unidad de placa de reactivo común, placas de reactivo comunes a un recipiente y placas de reactivo comunes a un lote o proceso de fabricación.
9. El método de una cualquiera de las reivindicaciones 1 a 8, en donde la normalización de los recuentos comprende determinar una representación porcentual.
10. El método de una cualquiera de las reivindicaciones 1 a 9, en donde el recuento normalizado es una puntuación z.
11. El método de una cualquiera de las reivindicaciones 1 a 10, en donde el recuento normalizado es una puntuación z robusta.
12. El método de una cualquiera de las reivindicaciones 1 a 11, en donde la derivada de los recuentos para la sección genómica es una representación porcentual de la sección genómica.
13. El método de una cualquiera de las reivindicaciones 4 a 12, en donde la mediana es una mediana de una representación porcentual.
14. El método de una cualquiera de las reivindicaciones 9 a 13, en donde la representación porcentual es una representación cromosómica.

15. El método de una cualquiera de las reivindicaciones 1 a 14, que comprende secuenciar el ácido nucleico mediante un módulo de secuenciación, proporcionando así las lecturas de secuencia de ácido nucleico y mapeando las lecturas de secuencia de ácido nucleico a las secciones genómicas de un genoma de referencia.
- 5

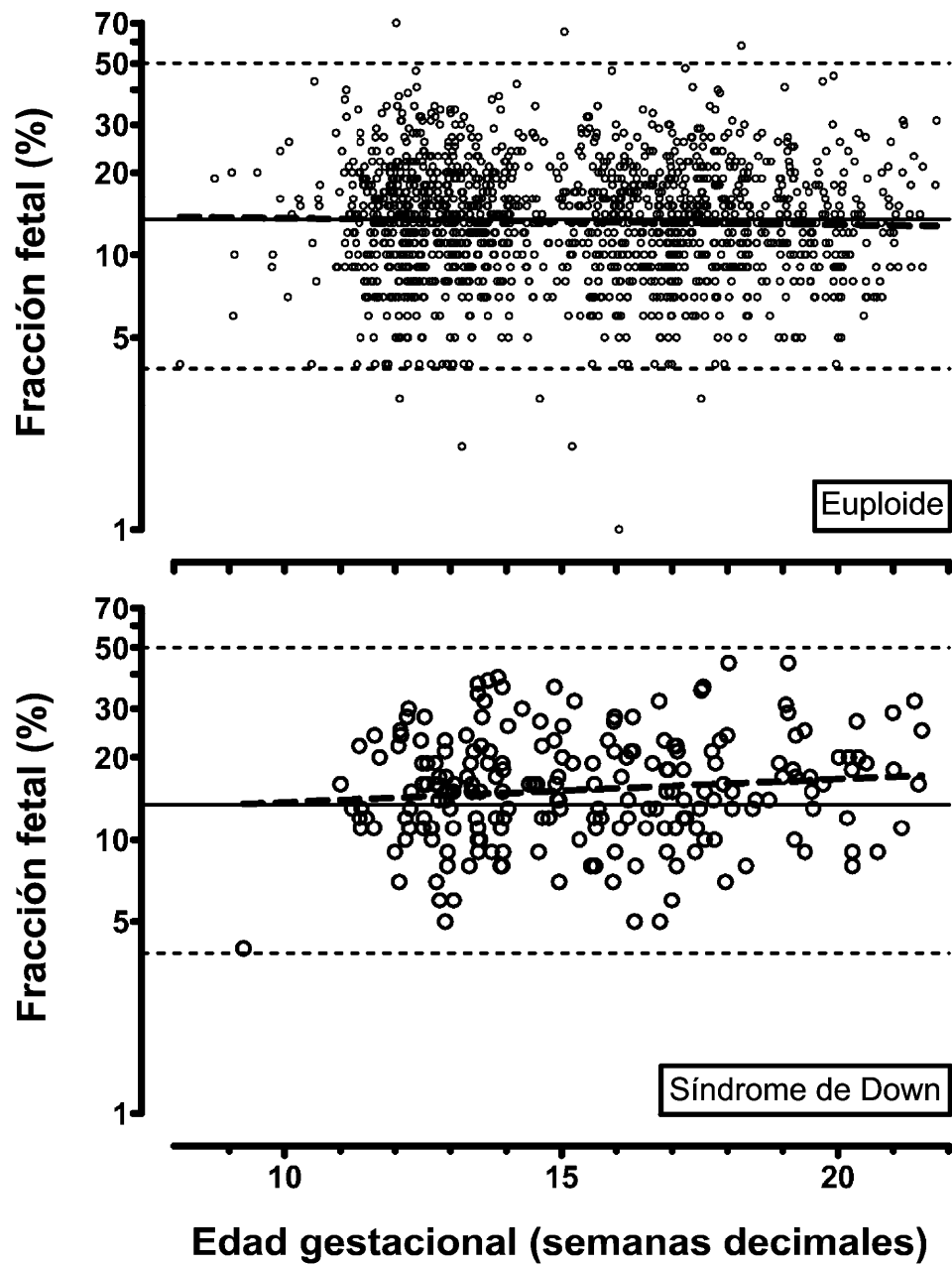


Fig. 1

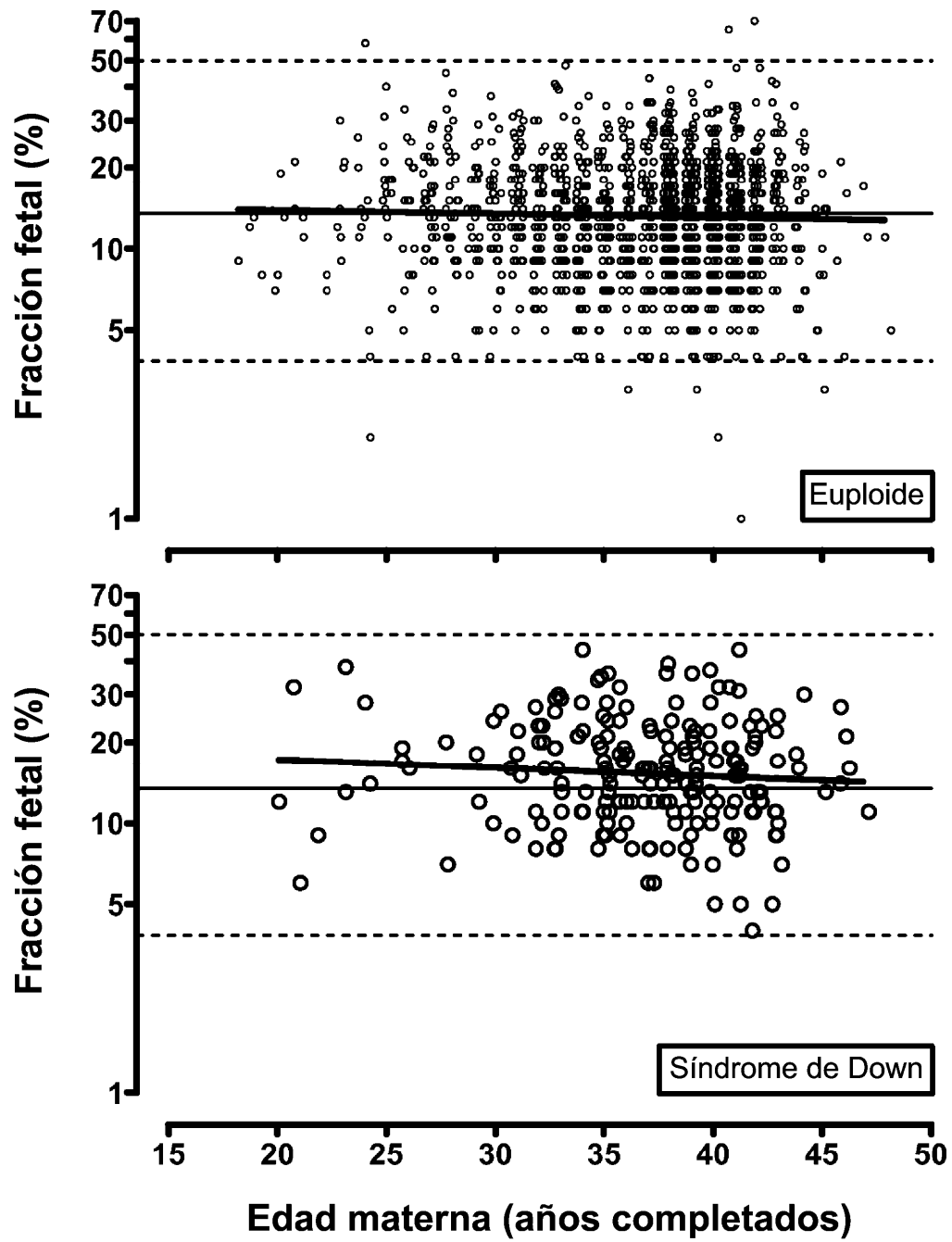


Fig. 2

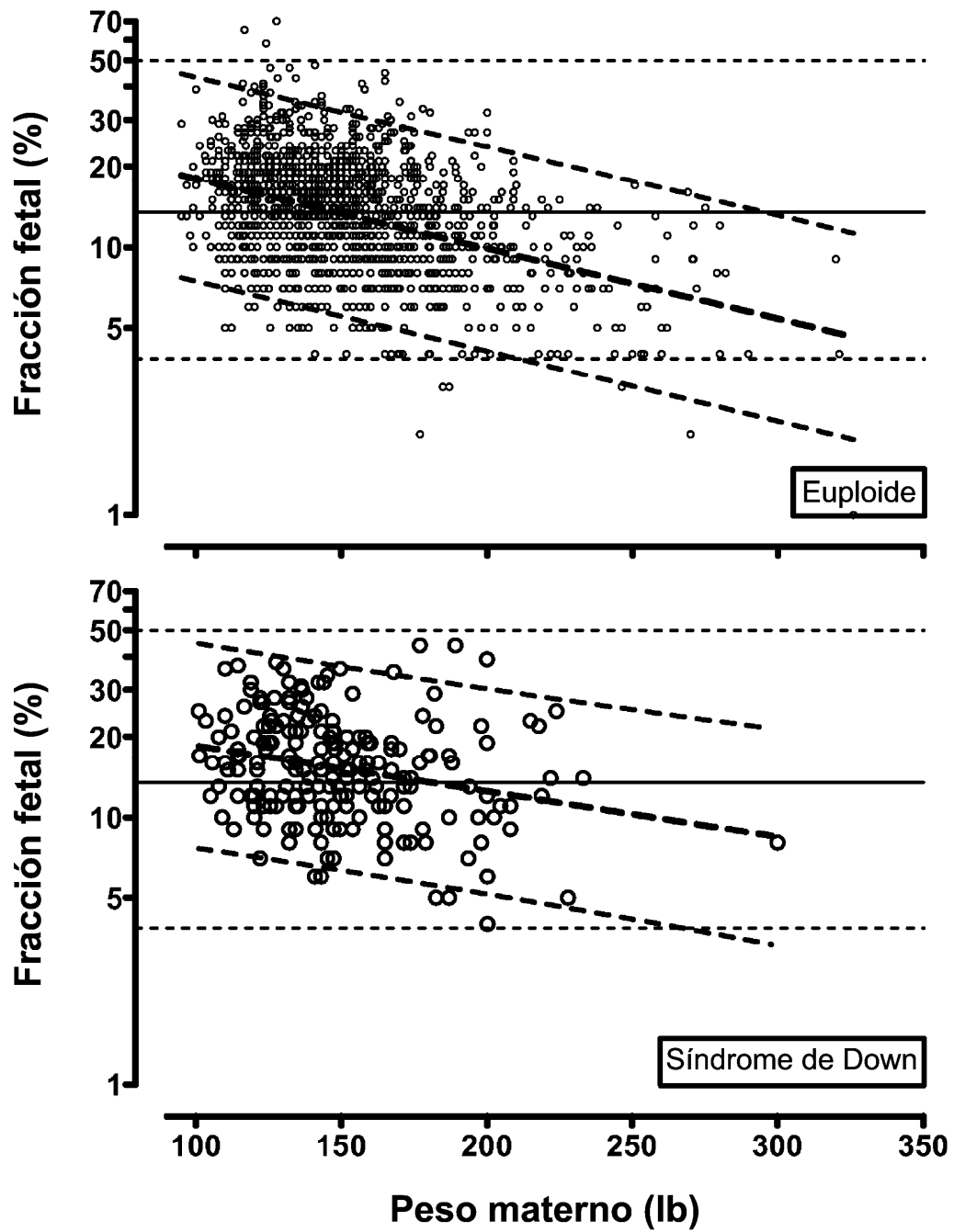


Fig. 3

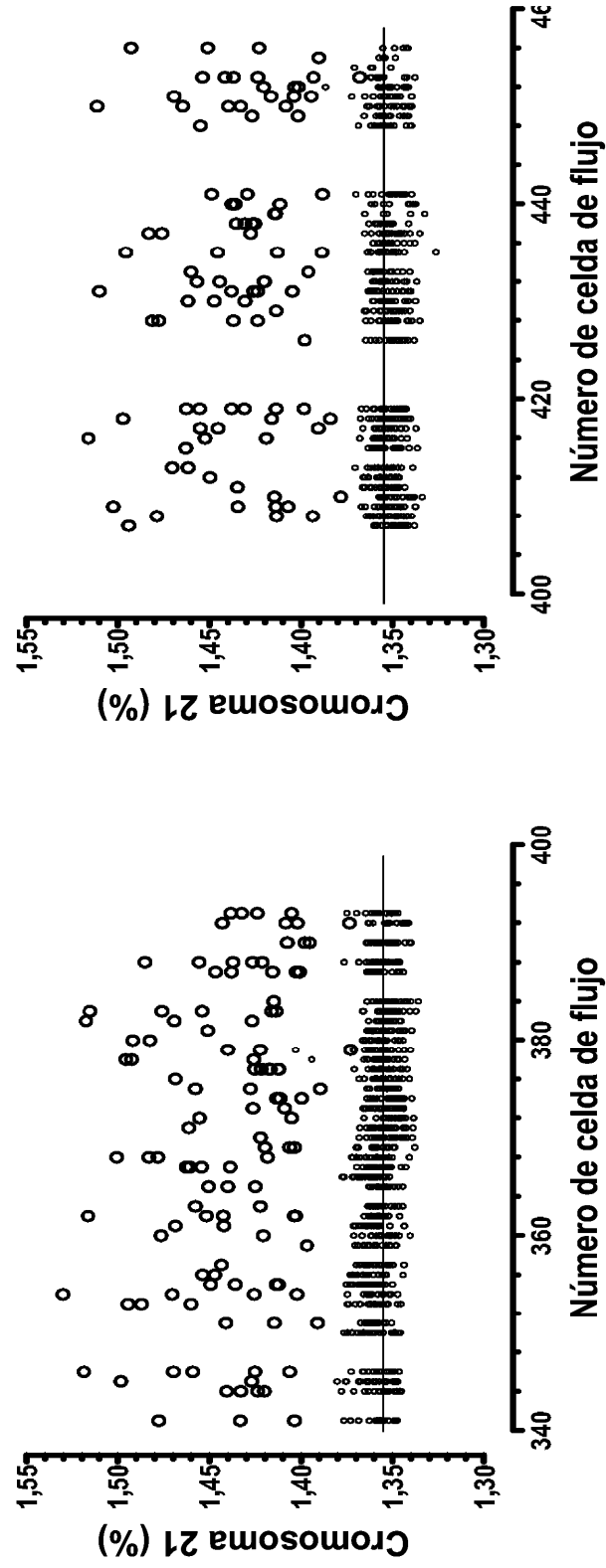


Fig. 4

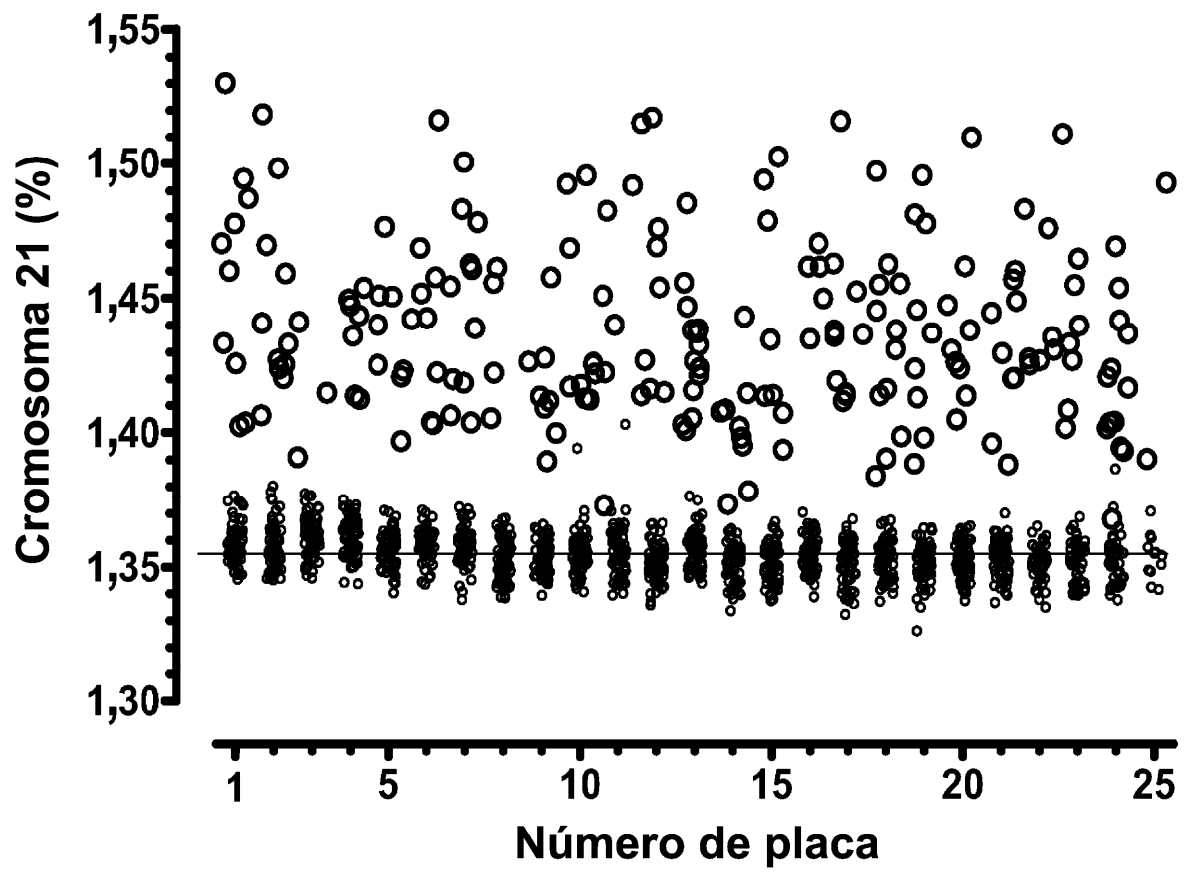


Fig. 5

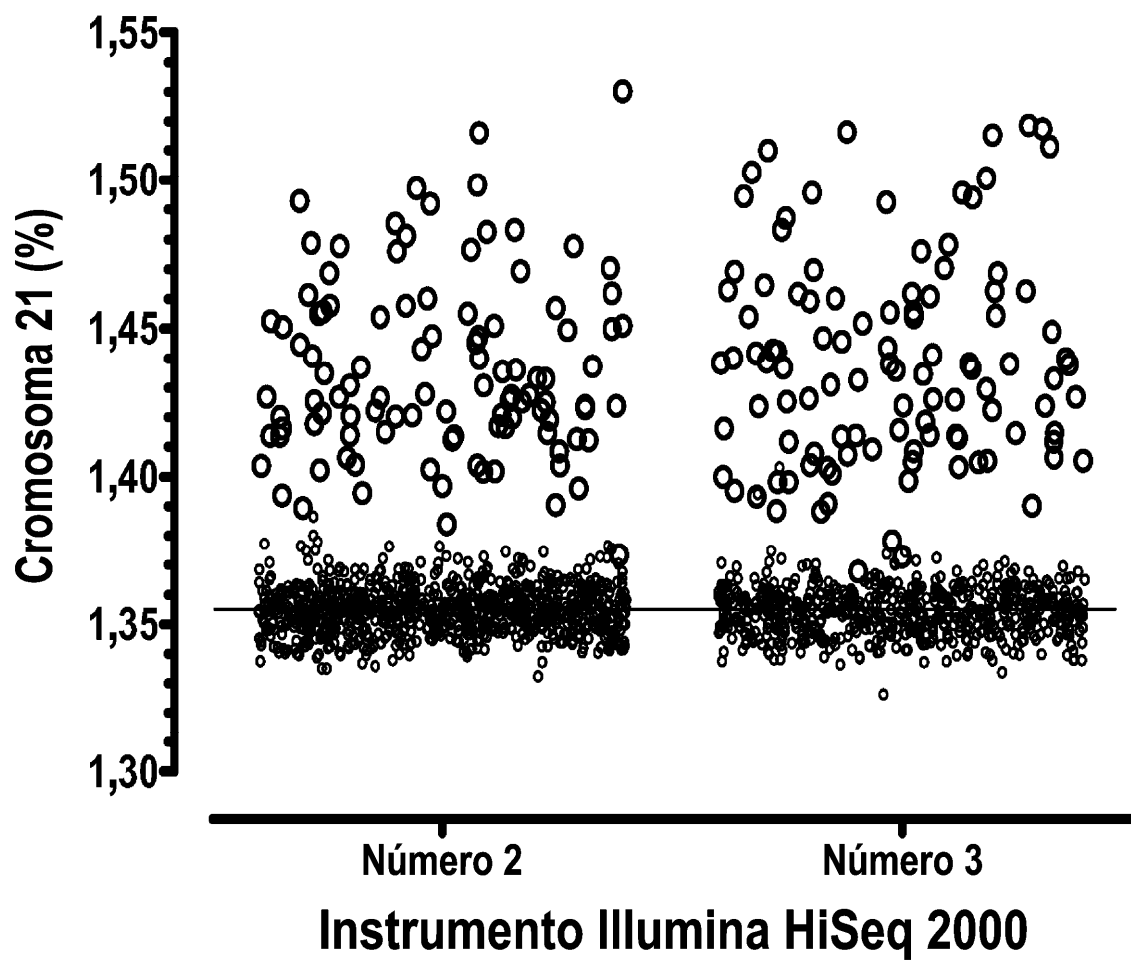


Fig. 6

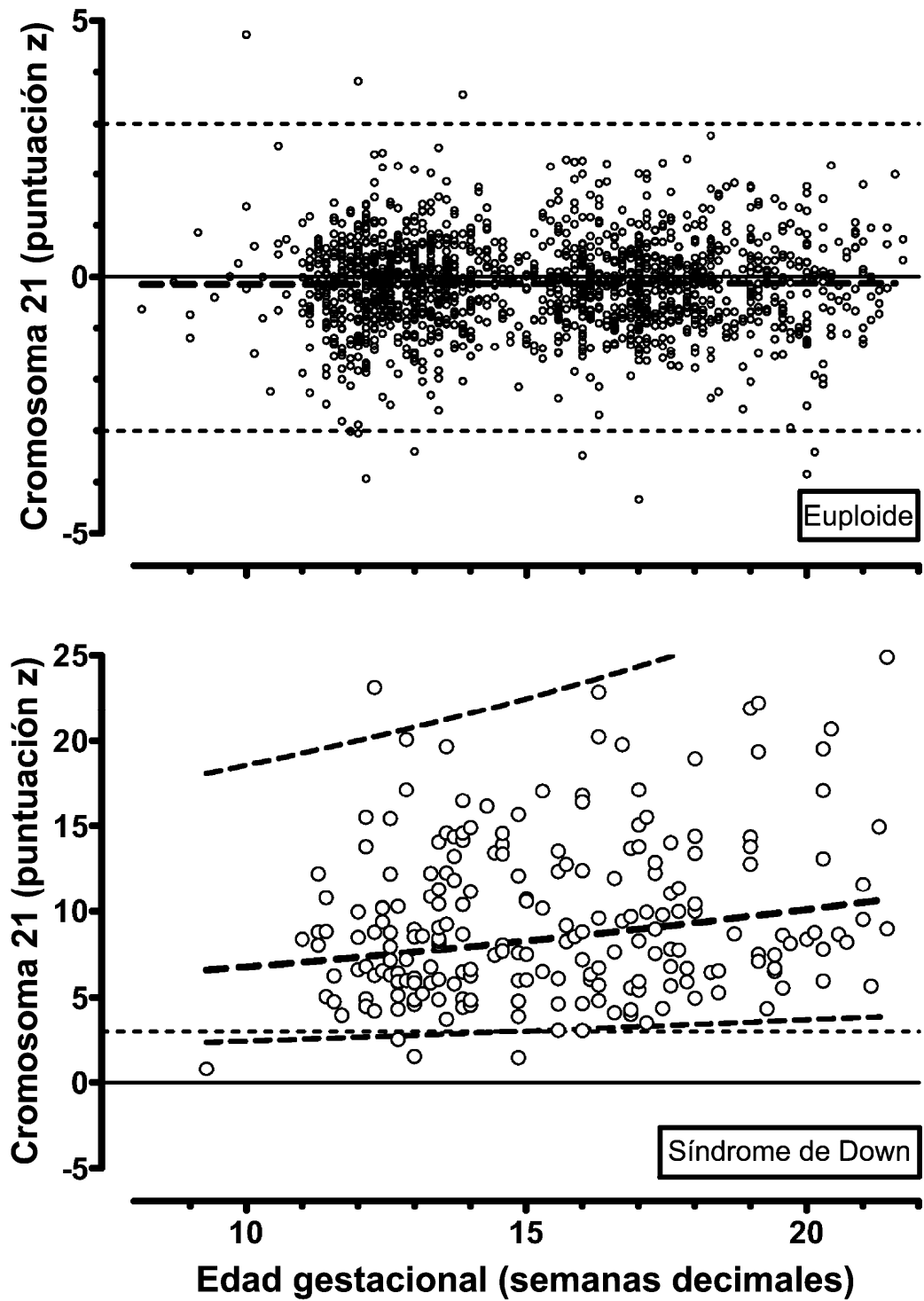


Fig. 7

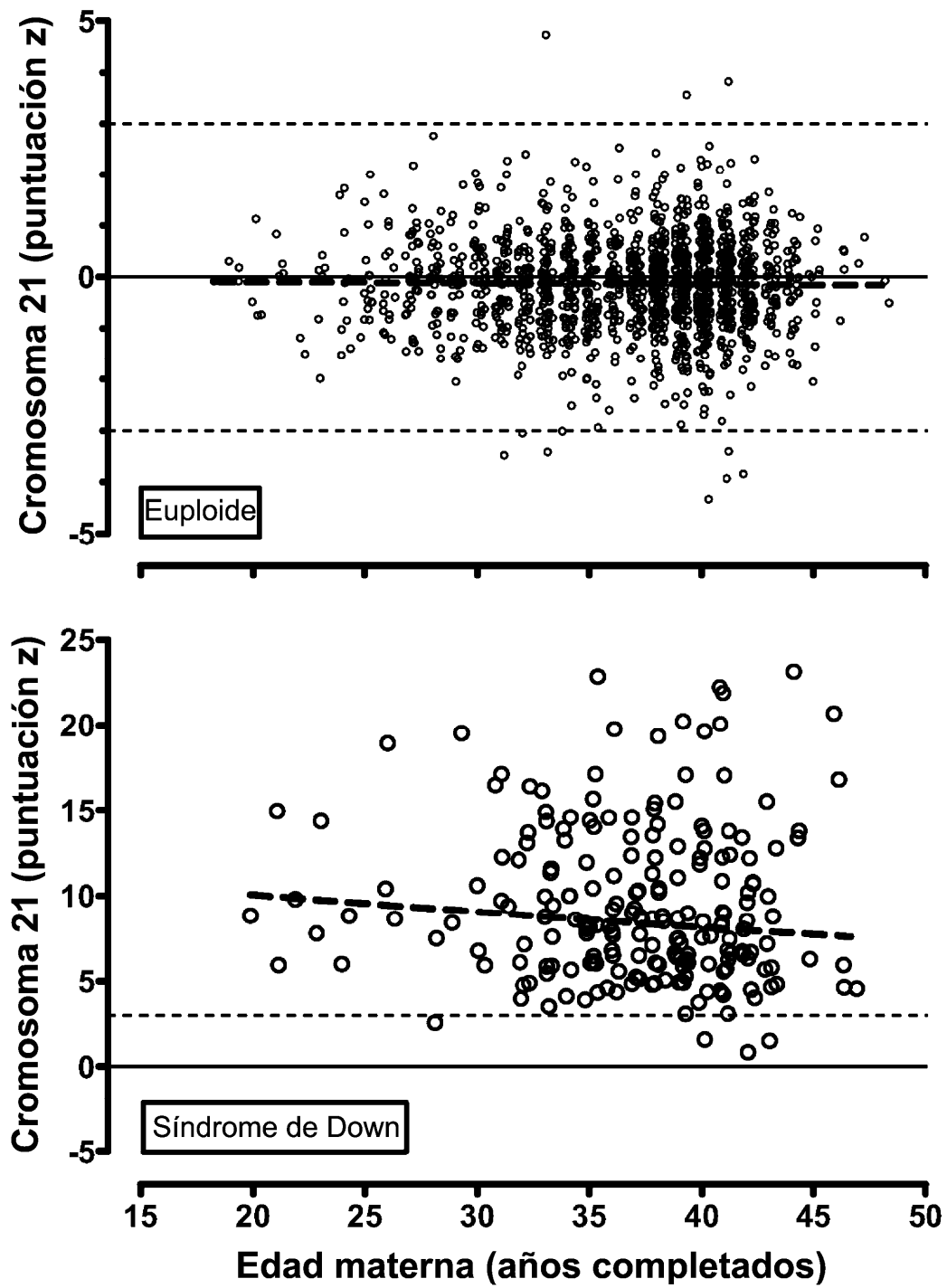


Fig. 8

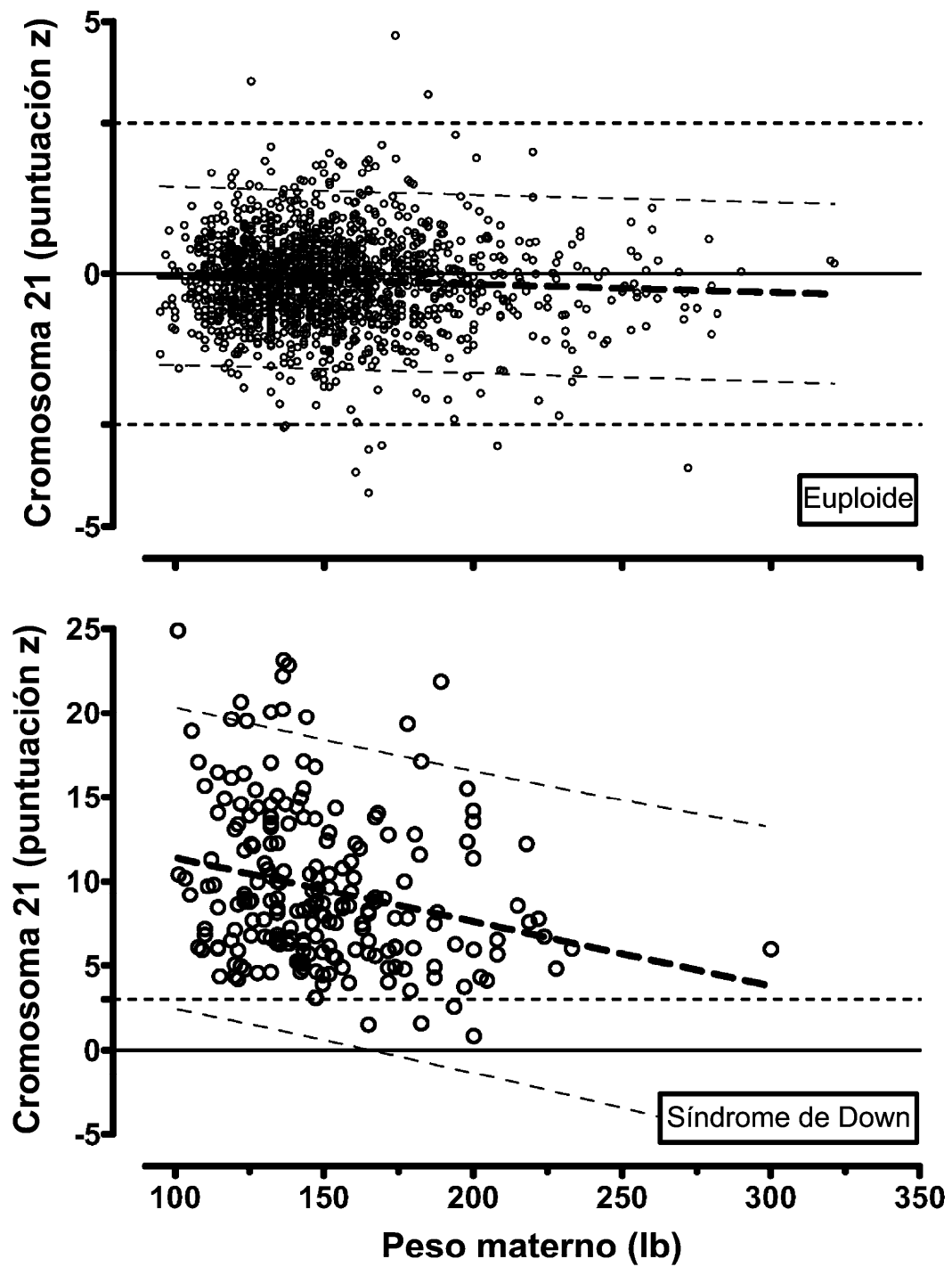


Fig. 9

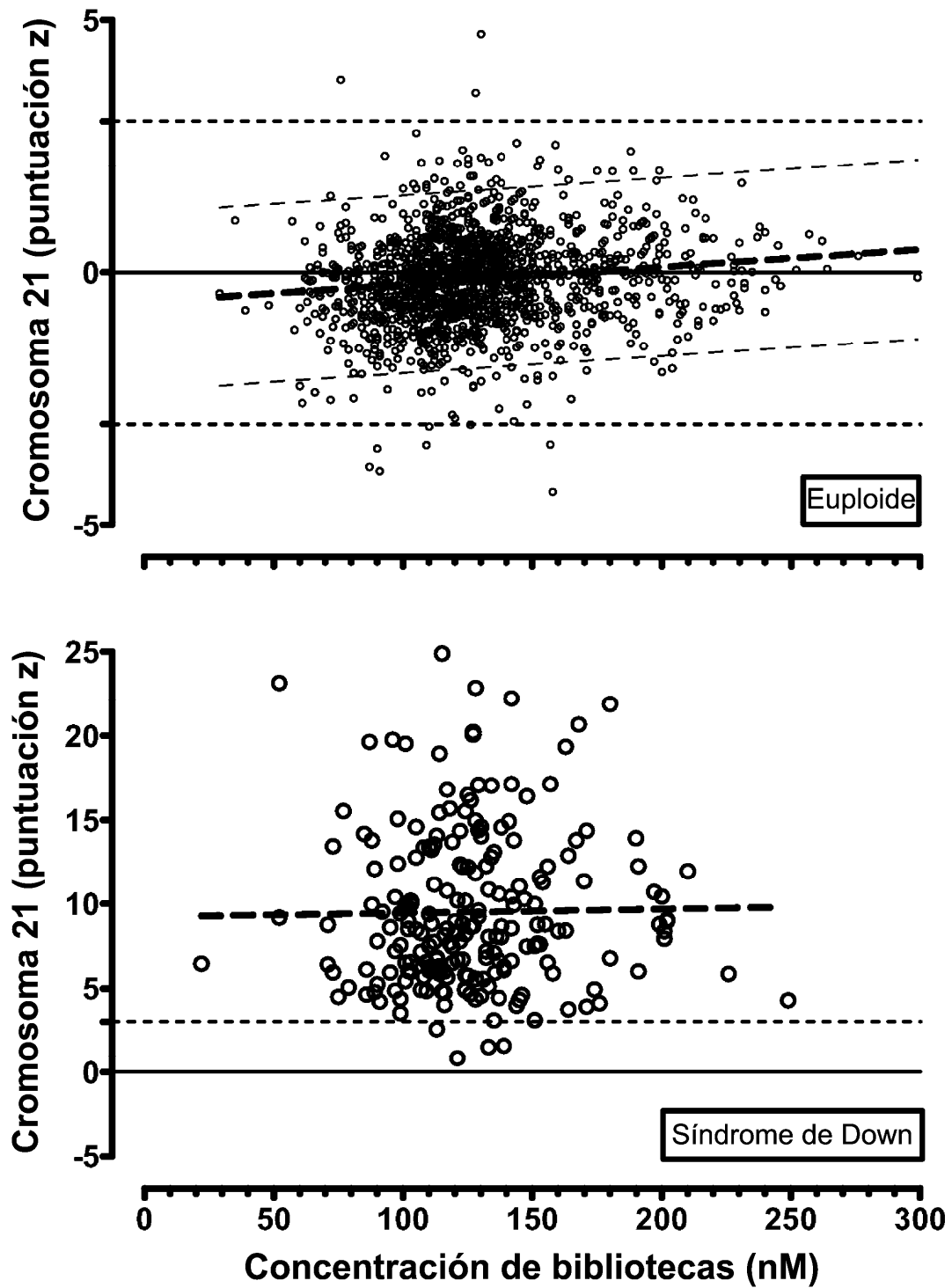


Fig. 10

Fig. 11

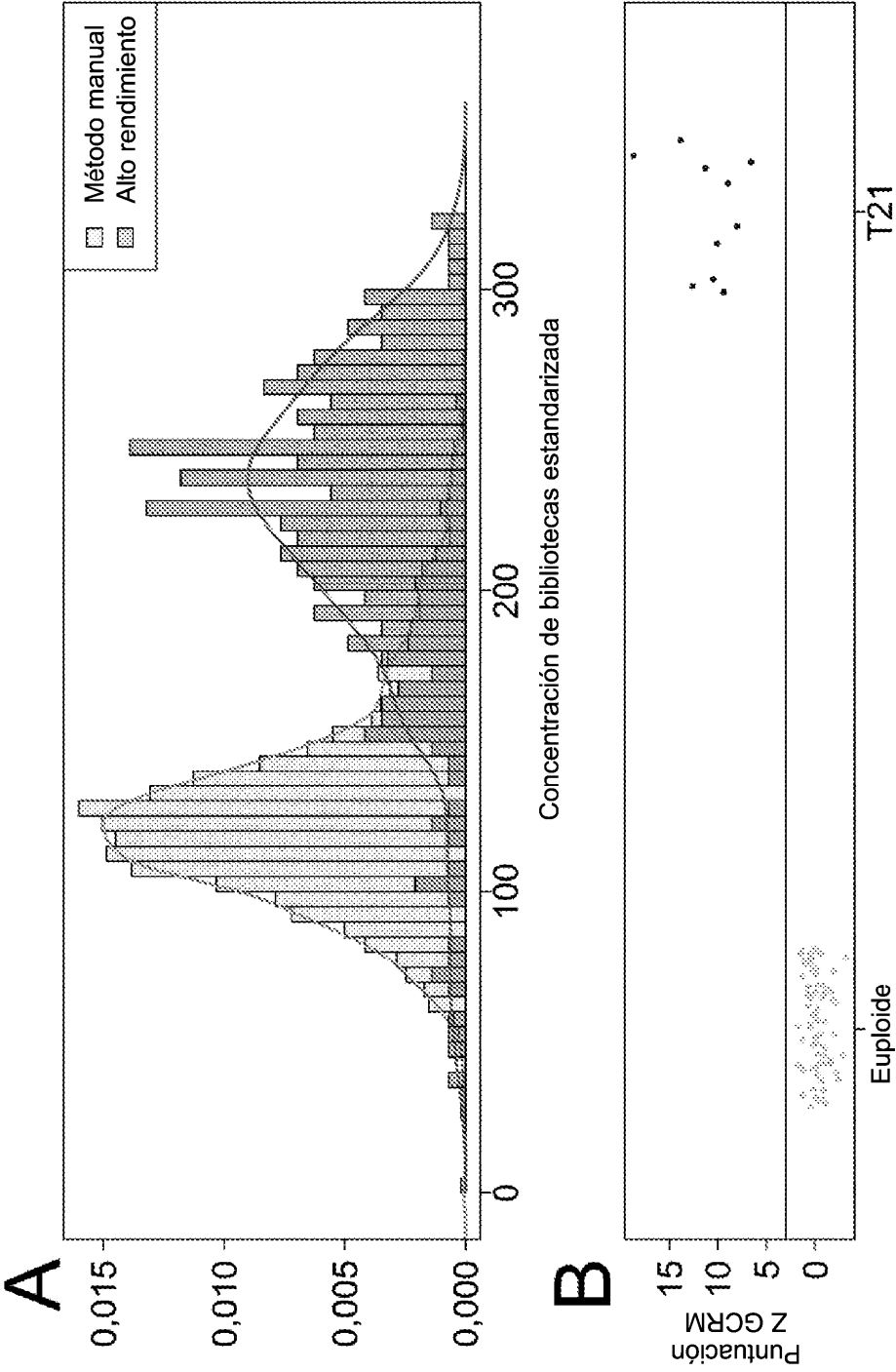


Fig. 12

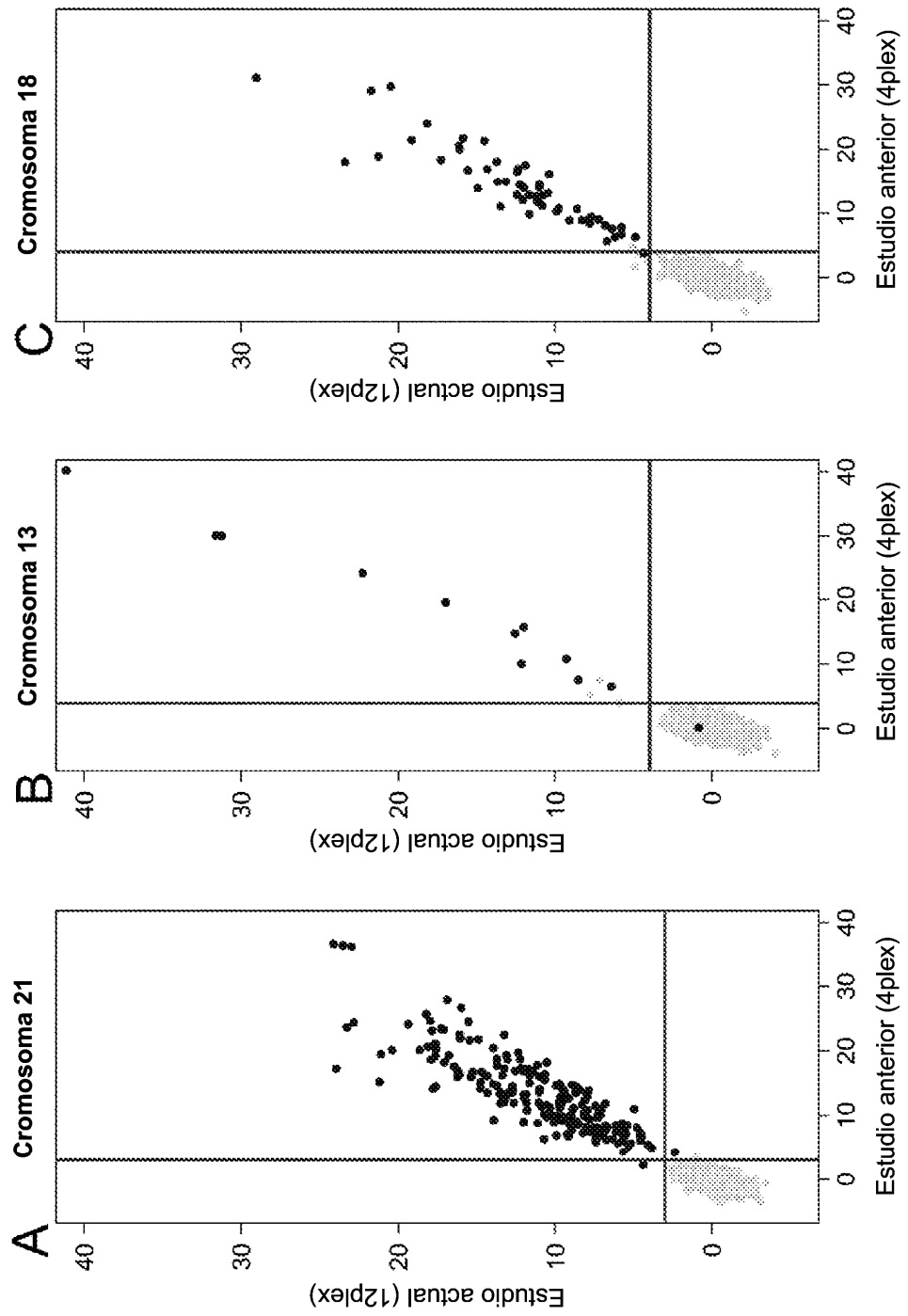


Fig. 13

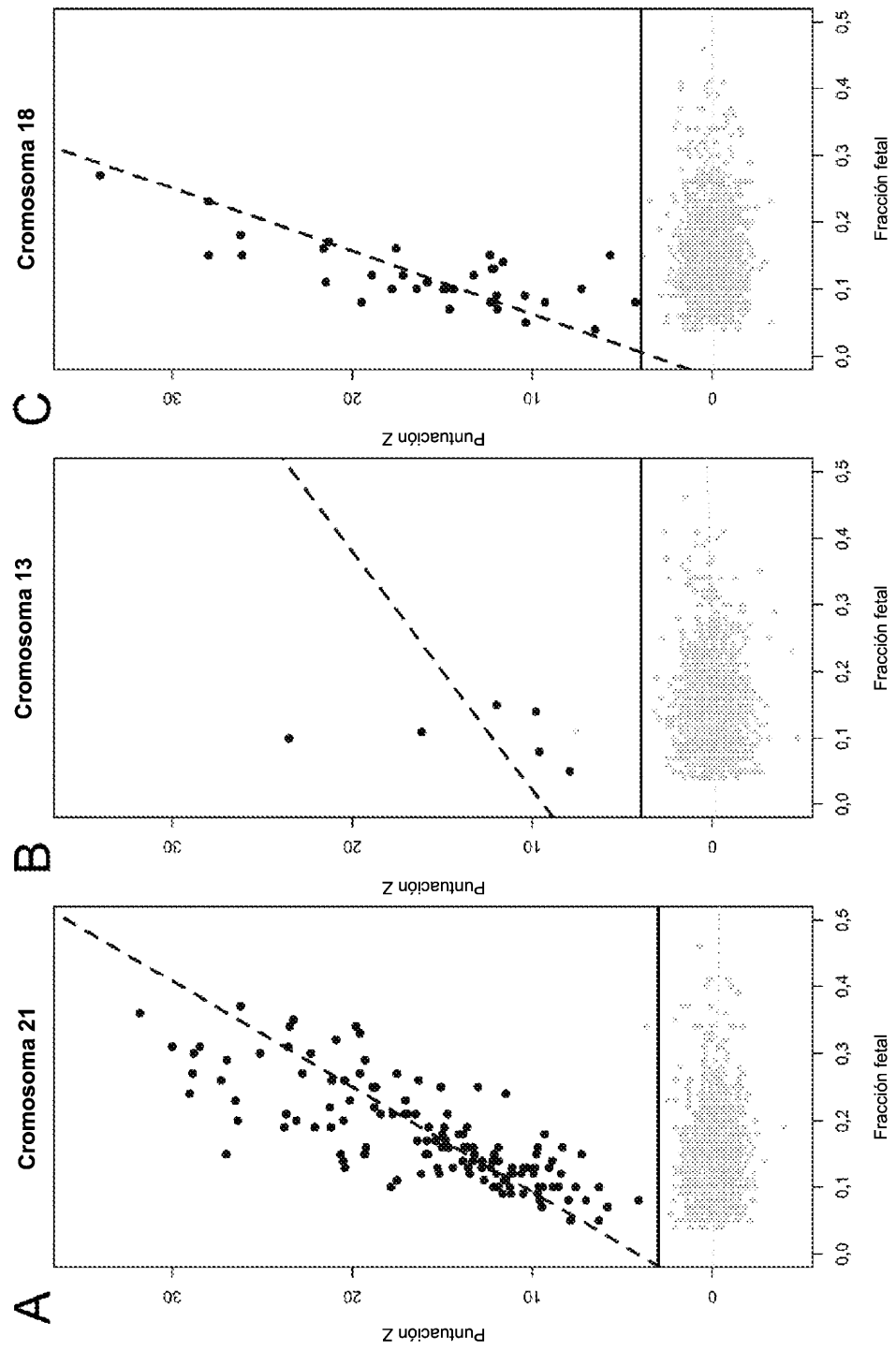


Fig. 14

