

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 968 644**

51 Int. Cl.:

G16B 20/00 (2009.01)

G16B 40/00 (2009.01)

G16B 20/10 (2009.01)

G16B 20/20 (2009.01)

G16B 30/10 (2009.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **02.10.2014** **PCT/US2014/058885**

87 Fecha y número de publicación internacional: **09.04.2015** **WO15051163**

96 Fecha de presentación y número de la solicitud europea: **02.10.2014** **E 14790884 (2)**

97 Fecha y número de publicación de la concesión europea: **18.10.2023** **EP 3053071**

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

30 Prioridad:

04.10.2013 US 201361887081 P

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

13.05.2024

73 Titular/es:

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

72 Inventor/es:

HANNUM, GREGORY

74 Agente/Representante:

DEL VALLE VALIENTE, Sonia

ES 2 968 644 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

Métodos y procedimientos para la evaluación no invasiva de variaciones genéticas

5 **Campo**

La presente invención se define mediante las reivindicaciones. La presente descripción de tecnología proporcionada en el presente documento se refiere en parte a métodos, procesos y máquinas 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.

20

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), trisomías 16 y 22, monosomía X (síndrome de Turner) y determinadas aneuploidías en cromosomas sexuales, tales como síndrome de Klinefelter (XXY), por ejemplo. Otra variación genética es el sexo del feto, que puede determinarse a menudo basándose en los cromosomas sexuales X e Y. Algunas variaciones genéticas pueden predisponer a un individuo a, o provocar, cualquiera de varias enfermedades tales como, por ejemplo, diabetes, arteriosclerosis, obesidad, diversas enfermedades autoinmunitarias y cáncer (por ejemplo, colorrectal, de mama, de ovario, de pulmón).

25

30

35

40

45

50

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 métodos, procesos y aparatos para determinar la presencia o ausencia de una aneuploidía cromosómica, microduplicación o microdelección en un feto por evaluación no invasiva de variaciones genéticas mediante análisis de decisión se describen en la solicitud de patente número WO 2014/190286 A2. Los métodos, procesos y aparatos para la evaluación no invasiva para variaciones genéticas para determinar el cariotipo del cromosoma sexual en un feto se describen en la solicitud de patente número US 2013/261983 A1. Un enfoque simple basado en un modelo aditivo generalizado para corregir simultáneamente diferentes fuentes de sesgos en datos de ARN-Seq fue desarrollado por Zheng y col. (Zheng W, Chung LM, Zhao H. Bias detection and correction in RNA-Sequencing data. BMC Bioinformatics 12, 290 (2011)). En determinadas implementaciones, la identificación de una o más varianzas o variaciones genéticas implica el análisis del ADN libre de células. El ADN libre de células (ADNlc) se compone de fragmentos de ADN que se originan a partir de la muerte celular y circulan en la sangre periférica. Las altas concentraciones de ADNlc pueden ser indicativas de determinadas afecciones clínicas, tales como cáncer, traumatismo, quemaduras, infarto de miocardio, accidente cerebrovascular, septicemia, infección y otras enfermedades. Adicionalmente, el ADN fetal libre de células (ADNflc) puede detectarse en el torrente sanguíneo materno y usarse para diversos diagnósticos prenatales no invasivos.

Resumen

La presente invención se define mediante las reivindicaciones. En consecuencia, la presente invención se refiere a un método para determinar la presencia o ausencia de una aneuploidía cromosómica, para una muestra que comprende:

55

(a) filtrar, según una distribución de densidad de lectura, porciones de un genoma de referencia, lo que proporciona de este modo un perfil de densidad de lectura para una muestra de prueba que comprende densidades de lectura de porciones filtradas, en donde:

60

(i) las densidades de lectura comprenden mediciones cuantitativas de recuentos de lecturas de secuencia mapeadas en porciones de un genoma de referencia, en donde las lecturas de secuencia son lecturas de ácido nucleico circulante libre de células de una muestra de prueba de una mujer embarazada, y

65

(ii) la distribución de densidad de lectura es una distribución de densidades de lectura promedio, media o en mediana, y se determina para densidades de lectura de porciones para múltiples muestras;

- (b) ajustar el perfil de densidad de lectura para la muestra de prueba al 1) restar un valor de mediana entrenado de las densidades de lectura de la muestra de prueba y 2) eliminar los componentes de las densidades de lectura de la muestra de prueba que se correlacionan con uno o más componentes principales del perfil, cuyos componentes principales (i) se obtienen a partir de un conjunto de entrenamiento de muestras euploides conocidas mediante un análisis de componentes principales, y (ii) representan uno o más sesgos en un perfil de densidad de lectura, lo que proporciona de este modo un perfil de densidad de lectura ajustado que comprende densidades de lectura ajustadas, en donde una pluralidad de sesgos se elimina del perfil de densidad de lectura ajustada;
- 5
- (c) comparar el perfil de densidad de lectura ajustado con un perfil de referencia que comprende densidades de lectura obtenidas de una o más muestras de referencia, lo que proporciona de este modo una comparación; y
- 10
- (d) determinar la presencia o ausencia de una aneuploidía cromosómica para la muestra de prueba según la comparación, en donde uno o más o todos los métodos de procesamiento se realizan por un procesador, un microprocesador, un ordenador, junto con memoria y/o por un aparato controlado por microprocesador.
- 15
- Según una realización preferida, el perfil de densidad de lectura se ajusta en (b) por 2 a 10 componentes principales.
- Según una realización preferida de la invención, la comparación en (c) comprende determinar un nivel de significancia, en donde el nivel de significancia indica una diferencia estadísticamente significativa entre el perfil de densidad de lectura ajustado y el perfil de referencia, y se determina la presencia de una aneuploidía cromosómica.
- 20
- Según una realización más preferida, determinar el nivel de significancia comprende determinar un valor de p.
- Según una realización preferida adicional, el perfil de referencia en (c) comprende:
- 25
- (i) densidades de lectura determinadas para una o más muestras euploides conocidas;
- (ii) densidades de lectura de porciones filtradas; y/o
- 30
- (iii) densidades de lectura ajustadas según el uno o más componentes principales.
- Según otra realización preferida:
- (i) las densidades de lectura de las porciones para las múltiples muestras en (a)(ii) son las densidades de lectura de la mediana;
- 35
- (ii) las densidades de lectura de las porciones filtradas para la muestra de prueba son las densidades de lectura de la mediana; y/o
- 40
- (iii) las densidades de lectura para el perfil de referencia en (c) son medianas de las densidades de lectura.
- Según una realización más preferida, las densidades de lectura para el perfil de densidad de lectura para la muestra de prueba, las densidades de lectura de porciones para las múltiples muestras en (a)(ii), y las densidades de lectura para el perfil de referencia en (c) se determinan según un proceso que comprende el uso de una estimación de densidad de núcleo.
- 45
- Según otra realización más preferida de la invención, el perfil de densidad de lectura para la muestra de prueba se determina según las densidades de lectura de la mediana para la muestra de prueba; y el perfil de referencia en (c) se determina según las densidades de lectura de la mediana para las una o más muestras de referencia.
- 50
- Según una realización preferida, las porciones de un genoma de referencia se filtran en (a) según una medida de incertidumbre para la distribución de densidad de lectura.
- Según una realización más preferida, la medida de incertidumbre es una DAM.
- 55
- Según una realización preferida, el perfil de densidad de lectura para la muestra de prueba y el perfil de referencia en (c) comprenden puntuaciones z de densidades de lectura.
- Según una realización preferida adicional, el perfil de densidad de lectura para la muestra de prueba es representativo de la dosificación cromosómica para la muestra de prueba, y el método comprende comparar la dosificación cromosómica para el perfil de densidad de lectura para la muestra de prueba con la dosificación cromosómica para el perfil de referencia en (c), lo que genera de este modo una comparación de dosificación cromosómica, en donde determinar la presencia o ausencia de una aneuploidía cromosómica para la muestra de prueba es según la comparación de dosificación cromosómica.
- 60
- 65

Según otra realización preferida, determinar la presencia o ausencia de una aneuploidía cromosómica para la muestra de prueba comprende identificar la presencia o ausencia de una copia de un cromosoma, dos copias de un cromosoma, tres copias de un cromosoma, cuatro copias de un cromosoma, cinco copias de un cromosoma, una delección de uno o más segmentos de un cromosoma o una inserción de uno o más segmentos de cromosoma.

Según otra realización preferida adicional, el método comprende antes de (a), obtener las lecturas de secuencia.

Finalmente, según una realización preferida, el método comprende antes de (b), ajustar el perfil de densidad de lectura para la muestra de prueba mediante una normalización de LOESS de GC.

Breve descripción de los dibujos

Las figuras ilustran implementaciones de la tecnología y no son limitativas. 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 implementaciones particulares.

La Figura 1 muestra una implementación de una densidad de GC proporcionada por un núcleo de Epanechnikov (banda ancha = 200 bp).

La Figura 2 muestra un gráfico de densidades de GC (eje y) para el gen HTRA1 donde las densidades de GC se normalizan en todo un genoma. Las posiciones genómicas se muestran en el eje x.

La Figura 3 muestra una distribución de estimaciones de sesgo de genoma local (por ejemplo, densidad de GC, eje x) para un genoma de referencia (línea continua) y para lecturas de secuencia obtenidas de una muestra (línea discontinua). Se muestran frecuencias de sesgo (por ejemplo, frecuencia de densidad) en el eje y. Las estimaciones de densidad de GC se normalizan en todo un genoma. En este ejemplo, la muestra tiene más lecturas con alto contenido de GC de lo que se esperaría de la referencia.

La Figura 4 muestra una comparación de una distribución de estimaciones de densidad de GC para un genoma de referencia y estimaciones de densidad de GC de lecturas de secuencia para una muestra usando una relación ajustada polinómica del 3.º orden ponderada. Las estimaciones de densidad de GC (eje x) se normalizaron en todo un genoma. Las frecuencias de densidad de GC se representan en el eje y como una relación log2 de frecuencias de densidad de la referencia dividida por aquellas de la muestra.

La Figura 5A muestra una distribución de las medianas de las densidades GC (eje x) para todas las porciones de un genoma. La Figura 5B muestra los valores de la mediana de desviación absoluta (DAM) (eje x) determinados según las distribuciones de densidad de GC para múltiples muestras. Las frecuencias de densidad de GC se muestran en el eje y. Las porciones se filtraron según la mediana de distribuciones de densidad de GC para múltiples muestras de referencia (por ejemplo, un conjunto de entrenamiento) y los valores de DAM se determinaron según distribuciones de densidad de GC de múltiples muestras. Las porciones que comprenden densidades de GC fuera de un umbral establecido (por ejemplo, cuatro veces el rango entre cuartiles de DAM) se eliminaron de la consideración según el proceso de filtrado.

La Figura 6A muestra un perfil de densidad de lectura de una muestra para un genoma que comprende la mediana de densidades de lectura (eje y, por ejemplo, densidad de lectura/porción) y posiciones relativas de cada porción genómica (eje x, índice de porción) dentro de un genoma. La Figura 6B muestra un primer componente principal (PC1) y la Figura 6C muestra un segundo componente principal (PC2) obtenido de un análisis de componentes principales de los perfiles de densidad de lectura obtenidos a partir de un conjunto de entrenamiento de 500 euploides.

La Figura 7A-C muestra un ejemplo de un perfil de densidad de lectura de una muestra para un genoma que comprende una trisomía del cromosoma 21 (por ejemplo, agrupado con dos líneas verticales). Las posiciones relativas de cada porción genómica se muestran en el eje x. Se proporcionan densidades de lectura en el eje y. La Figura 7A muestra un perfil de densidad de lectura sin procesar (por ejemplo, no ajustado). La Figura 7B muestra el perfil de 7A que comprende un primer ajuste que comprende una sustracción de la mediana de perfil. La Figura 7C muestra el perfil de 7B que comprende un segundo ajuste. El segundo ajuste comprende la resta de 8 veces perfiles de componentes principales, ponderados en base a su representación encontrada en esta muestra. (por ejemplo, se construye un modelo). Por ejemplo, un perfil de muestra = $A*PC1 + B*PC2 + C*PC3...$ y un perfil corregido, por ejemplo, como se muestra en 7C = perfil de muestra - $A*PC1 + B*PC2 + C*PC3...$

La Figura 8 muestra un gráfico QQ de valores de p de prueba a partir de muestras de entrenamiento de remuestreo tipo *bootstrapping* para una prueba T21. Una gráfica QQ generalmente compara dos distribuciones. La Figura 8 muestra una comparación de las puntuaciones de ChAI (eje y) de muestras de prueba a una distribución uniforme (es decir, la distribución esperada de valores de p, eje x). Cada punto representa puntuaciones de valor log-p de una muestra de prueba única. Las muestras se clasifican y asignan un valor "esperado" (eje x) en base a la distribución uniforme. La línea discontinua inferior representa la diagonal y la línea superior representa un umbral de Bonferroni. Se esperaría que las muestras que siguen a una distribución uniforme aterricen en la diagonal inferior (línea discontinua inferior).

Los valores de datos se encuentran bien fuera de las diagonales debido a las correlaciones en las porciones (por ejemplo, sesgo) que indican más muestras de puntuación alta (bajo valor de p) de lo esperado. Los métodos descritos en el presente documento (por ejemplo, ChAI, por ejemplo, véase el ejemplo 1) pueden corregir este sesgo observado.

5 La Figura 9A muestra un gráfico de densidad de lectura que muestra una diferencia en los coeficientes de PC2 para hombres y mujeres en un conjunto de entrenamiento. La Figura 9B muestra un gráfico de características operativas del receptor (ROC, por sus siglas en inglés) para las identificaciones de género con un coeficiente de PC2. Se usaron identificaciones de sexo realizadas por secuenciación para la referencia de la verdad.

10 Las Figuras 10A-10B muestran una implementación de un sistema.

La Figura 11 muestra una implementación de un sistema.

La Figura 12 muestra una implementación de un método proporcionado en el presente documento.

15 Descripción detallada

La secuenciación de nueva generación permite secuenciar ácidos nucleicos en una escala de todo el genoma mediante métodos que son más rápidos y más baratos que los métodos tradicionales de secuenciación. Los métodos, sistemas y productos proporcionados en el presente documento pueden utilizar tecnologías de secuenciación avanzadas para localizar e identificar variaciones genéticas y/o enfermedades y trastornos asociados. Los métodos, sistemas y productos proporcionados en el presente documento a menudo pueden proporcionar una evaluación no invasiva del genoma de un sujeto (por ejemplo, un genoma fetal) mediante el uso de una muestra de sangre, o parte de la misma, y a menudo son más seguros, más rápidos y/o menos costosos que técnicas más invasivas (por ejemplo, amniocentesis, biopsia). En algunas realizaciones, en el presente documento se proporcionan métodos que comprenden, en parte, obtener lecturas de secuencia de ácidos nucleicos presentes en una muestra, cuyas lecturas de secuencia a menudo se mapean a una secuencia de referencia, recuentos de procesamiento de lecturas de secuencia y determinación de la presencia o ausencia de una variación genética. Los sistemas, métodos y productos proporcionados en el presente documento son útiles para localizar y/o identificar variaciones genéticas y son útiles para diagnosticar y tratar enfermedades, trastornos y discapacidades asociadas con determinadas variaciones genéticas.

También se proporcionan en el presente documento, en algunas implementaciones, métodos de manipulación de datos para reducir y/o eliminar el sesgo de secuenciación introducido por diversos aspectos de una tecnología de secuenciación. El sesgo de secuenciación a menudo contribuye a una distribución no uniforme de lecturas a través de un genoma, o un segmento del mismo, y/o variaciones en la calidad de lectura. El sesgo de secuenciación puede corromper los datos de secuenciación genómica, perjudicar el análisis efectivo de datos, contaminar los resultados e impedir la interpretación precisa de los datos. A veces se puede reducir el sesgo de secuenciación al aumentar la cubierta de secuenciación; sin embargo, este enfoque a menudo infla costos de secuenciación y tiene una efectividad muy limitada. Los métodos de manipulación de datos descritos en el presente documento pueden reducir y/o eliminar el sesgo de secuenciación, lo que de este modo mejora la calidad de los datos de lectura de secuencia sin aumentar los costos de secuenciación. También se proporcionan en el presente documento sistemas, máquinas, aparatos, productos y módulos que, en algunas implementaciones, llevan a cabo métodos descritos en el presente documento.

45 Muestras

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

El ácido nucleico o una mezcla de ácido nucleico usados en los métodos, sistemas, máquinas y/o aparatos descritos en el presente documento se suele aislar de una muestra obtenida de un sujeto (por ejemplo, un sujeto de prueba). Un sujeto del cual se obtiene una muestra se denomina a veces en el presente documento un sujeto de prueba. Un sujeto puede ser cualquier organismo vivo o no vivo incluyendo, pero sin limitarse a, un ser humano, animal no humano, planta, bacteria, hongo, virus o protista. Puede seleccionarse cualquier animal humano o no humano incluyendo, pero sin limitarse a, mamífero, reptil, ave, anfibio, pez, ungulado, rumiante, bovino (por ejemplo, ganado vacuno), equino (por ejemplo, caballo), caprino y ovino (por ejemplo, oveja, cabra), porcino (por ejemplo, cerdo), camélido (por ejemplo, camello, llama, alpaca), mono, simio (por ejemplo, gorila, chimpancé), úrsido (por ejemplo, oso), ave de corral, perro, gato, ratón, rata, pescado, delfín, ballena y tiburón. Un sujeto puede ser un hombre o una mujer (por ejemplo, una mujer, una mujer embarazada). Un sujeto puede tener cualquier edad (por ejemplo, un embrión, un feto, un bebé, un niño, un adulto).

El ácido nucleico puede aislarse de cualquier tipo de espécimen o muestra biológica adecuada (por ejemplo, una muestra de prueba). Una muestra o muestra de prueba puede ser cualquier espécimen que se aisle u obtenga de un sujeto o parte del mismo (por ejemplo, un sujeto humano, una mujer embarazada, un feto). A menudo se obtiene una muestra de prueba a partir de un sujeto de prueba. A menudo se obtiene una muestra de prueba de un sujeto femenino gestante (por ejemplo, una mujer humana embarazada). Los ejemplos no limitativos de especímenes incluyen fluidos o tejidos de un sujeto, incluyendo, sin limitación, sangre o un producto sanguíneo (por ejemplo, suero, plasma o similares), sangre de cordón umbilical, vellosidades coriónicas, líquido amniótico, líquido cefalorraquídeo, líquido medular, líquido de lavado (por ejemplo, broncoalveolar, gástrico, peritoneal, ductal, ótico, artroscópico), muestra de biopsia (por ejemplo, de embrión preimplantado), muestra de celocentesis, células (células sanguíneas, células placentarias, células embrionarias o fetales, células nucleadas fetales o restos celulares fetales) o partes de las mismas (por ejemplo, mitocondriales, núcleos, extractos o similares), lavados del aparato genital femenino, orina, heces, esputo, saliva, mucosa nasal, fluido prostático, lavados, semen, fluido linfático, bilis, lágrimas, sudor, leche materna, fluido mamario, similares o combinaciones de los mismos. Una muestra de prueba puede comprender sangre o un producto sanguíneo (por ejemplo, plasma, suero, linfocitos, plaquetas, capas leucocitarias). Una muestra de prueba a veces comprende suero obtenido de una mujer embarazada. Una muestra de prueba a veces comprende plasma obtenido de una mujer embarazada. En algunas implementaciones, una muestra biológica es un hisopo cervicouterino de un sujeto. En algunas implementaciones, una muestra biológica puede ser sangre y, algunas veces, plasma o suero. El término "sangre", como se usa en el presente documento, se refiere a una muestra o preparación de sangre de un sujeto (por ejemplo, un sujeto de prueba, por ejemplo, una mujer embarazada o una mujer a la que se le está realizando una prueba de posible embarazo). El término abarca sangre completa, un hemoderivado o cualquier fracción de sangre, como suero, plasma, capa leucocitaria o similares, según se define convencionalmente. La sangre o fracciones de la misma comprenden a menudo nucleosomas (por ejemplo, nucleosomas maternos y/o fetales). Los nucleosomas comprenden ácidos nucleicos y, algunas veces, son sin células o intracelulares. La sangre comprende además capas leucocíticas. Algunas veces, las capas leucocíticas se aíslan utilizando un gradiente de Ficoll. Las capas leucocitarias pueden comprender glóbulos blancos (por ejemplo, leucocitos, linfocitos T, linfocitos B, plaquetas y similares). En determinadas implementaciones, las capas leucocitarias comprenden ácido nucleico materno y/o fetal. Plasma sanguíneo se refiere a la fracción de sangre completa que resulta de la centrifugación de sangre tratada con anticoagulantes. Suero sanguíneo se refiere a la porción acuosa del fluido que queda después de que se ha coagulado una muestra de sangre. A menudo, se recogen muestras de líquido o tejido según protocolos convencionales que siguen generalmente hospitales o clínicas. A menudo, para la sangre se recoge una cantidad adecuada de sangre periférica (por ejemplo, entre 3-40 mililitros) y puede almacenarse según procedimientos convencionales antes o después de la preparación. Una muestra de líquido o tejido de la que se extrae ácido nucleico puede ser acelular (por ejemplo, libre de células). En algunas implementaciones, una muestra de líquido o tejido puede contener elementos celulares o restos celulares. En algunas implementaciones, pueden incluirse células fetales o células cancerosas en la muestra.

Una muestra es a menudo heterogénea, lo que significa que más de un tipo de especie de ácido nucleico está presente en la muestra. Por ejemplo, el ácido nucleico heterogéneo puede incluir, pero no se limita a, (i) ácido nucleico derivado fetal y derivado materno, (ii) ácido nucleico de cáncer y distinto de cáncer, (iii) ácido nucleico patógeno y de huésped y, más generalmente, (iv) ácido nucleico mutado y de tipo natural. Una muestra puede ser heterogénea porque más de un tipo de célula está presente, tal como una célula fetal y una célula materna, una célula cancerosa y no cancerosa, o una célula patógena y de huésped. En algunas implementaciones, están 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 sometiéndose a prueba para un posible embarazo. La edad gestacional adecuada puede variar dependiendo de la prueba prenatal que se realiza. En determinadas implementaciones, un sujeto de sexo femenino gestante algunas veces está en el primer trimestre de embarazo, a veces en el segundo trimestre de embarazo o, algunas veces, en el tercer trimestre de embarazo. En determinadas implementaciones, se recoge un líquido o tejido de una mujer embarazada de aproximadamente 1 a aproximadamente 45 semanas de gestación fetal (por ejemplo, a las 1-4, 4-8, 8-12, 12-16, 16-20, 20-24, 24-28, 28-32, 32-36, 36-40 o 40-44 semanas de gestación fetal), y a veces de aproximadamente 5 a aproximadamente 28 semanas de gestación fetal (por ejemplo, a las 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26 o 27 semanas de gestación fetal). En determinadas implementaciones, se recoge una muestra de líquido o tejido de una mujer embarazada durante o justo después (por ejemplo, de 0 a 72 horas después) de dar a luz (por ejemplo, parto vaginal o no vaginal (por ejemplo, parto quirúrgico)).

Adquisición de muestras de sangre y extracción de ADN

Los métodos incluidos en el presente documento suelen incluir separar, enriquecer y analizar el ADN fetal que se encuentra en la sangre materna como medio no invasivo para detectar la presencia o ausencia de una variación genética materna y/o fetal y/o controlar la salud de un feto y/o una mujer embarazada durante la gestación y, a veces, tras ella. Por lo tanto, las primeras etapas para poner en práctica determinados métodos en el presente documento suelen incluir la obtención de una muestra de sangre de una mujer embarazada y la extracción de ADN de una muestra.

Adquisición de muestras de sangre

Puede obtenerse una muestra de sangre de una mujer embarazada a una edad gestacional adecuada para la prueba utilizando un método de la presente tecnología. Una edad gestacional adecuada puede variar en función del trastorno analizado, como se explica más adelante. La extracción de sangre de una mujer suele realizarse según el protocolo convencional que suelen seguir los hospitales o las clínicas. Se suele extraer una cantidad adecuada de sangre periférica, por ejemplo, normalmente entre 5-50 ml, que puede almacenarse según el procedimiento convencional antes de su posterior preparación. Las muestras de sangre se pueden extraer, almacenar o transportar de manera que se minimice la degradación o la calidad del ácido nucleico presente en la muestra.

Preparación de muestras de sangre

Puede realizarse un análisis del ADN fetal encontrado en la sangre materna utilizando, por ejemplo, sangre completa, suero o plasma. Se conocen métodos para preparar suero o plasma a partir de sangre materna. Por ejemplo, se puede colocar sangre de una mujer embarazada en un tubo que contiene EDTA o un producto comercial especializado como Vacutainer SST (Becton Dickinson, Franklin Lakes, N.J.) para evitar la coagulación de la sangre, y a continuación se puede obtener plasma de la sangre completa mediante centrifugación. El suero se puede obtener con o sin centrifugación, después de la coagulación de la sangre. Si se utiliza la centrifugación, normalmente, aunque no exclusivamente, se realiza a una velocidad adecuada, por ejemplo, 1500-3000 xg. El plasma o el suero se pueden someter a etapas de centrifugación adicionales antes de transferirlos a un tubo nuevo para la extracción del ADN.

Además de la porción acelular de la sangre completa, también se puede recuperar el ADN de la fracción celular, enriquecida en la porción de la capa leucocitaria, que se puede obtener tras la centrifugación de una muestra de sangre completa de la mujer y la extracción del plasma.

Extracción de ADN

Existen numerosos métodos conocidos para extraer el ADN de una muestra biológica que incluye sangre. Los métodos generales de preparación de ADN (por ejemplo, descritos por Sambrook y Russell, *Molecular Cloning: A Laboratory Manual* 3d ed., 2001) se pueden seguir; también se pueden utilizar diferentes reactivos o kits disponibles en el mercado, tales como el kit de ácido nucleico circulante QIAamp de Qiagen, el mini kit de ADN QiaAmp o el mini kit de sangre de ADN QiaAmp (Qiagen, Hilden, Alemania), kit de aislamiento de ADN de la sangre GenomicPrep™ (Promega, Madison, Wis.) y el kit de purificación de ADN de sangre genómica GFX™ (Amersham, Piscataway, NJ) para obtener ADN de una muestra de sangre de una mujer embarazada. También se pueden utilizar combinaciones de más de uno de estos métodos.

En algunas implementaciones, la muestra puede enriquecerse o enriquecerse relativamente en ácido nucleico fetal mediante uno o más métodos. Por ejemplo, se puede realizar la diferenciación del ADN fetal y del materno se mediante las composiciones y los procesos de la presente tecnología solos o junto con otros factores de diferenciación. Los ejemplos de estos factores incluyen, entre otros, diferencias de un solo nucleótido entre los cromosomas X e Y, secuencias específicas del cromosoma Y, polimorfismos ubicados en otras partes del genoma, diferencias de tamaño entre el ADN fetal y materno, y diferencias en el patrón de metilación entre los tejidos maternos y fetales.

Otros métodos para enriquecer una muestra para una especie particular de ácido nucleico se describen en la solicitud de patente PCT número PCT/US07/69991, presentada el 30 de mayo de 2007, solicitud de patente PCT número PCT/US2007/071232, presentada el 15 de junio de 2007, solicitudes provisionales estadounidenses números 60/968.876 y 60/968.878 (asignadas al solicitante), (solicitud de patente PCT número PCT/EP05/012707, presentada el 28 de noviembre de 2005). En determinadas implementaciones, el ácido nucleico materno se elimina selectivamente (bien parcial, sustancial, casi completa o completamente) de la muestra.

Las expresiones “ácido nucleico” y “molécula de ácido nucleico” pueden utilizarse indistintamente a lo largo de la divulgación. Los términos se refieren a ácidos nucleicos de cualquier composición de, tal como ADN (por ejemplo, ADN complementario (ADNc), ADN genómico (ADNg) y similares), ARN (por ejemplo, ARN mensajero (ARNm), ARN inhibidor corto (ARNic), ARN ribosómico (ARNr), ARNt, microARN, ARN altamente expresado por el feto o la placenta, y similares), y/o análogos de ADN o ARN (por ejemplo, que contengan análogos de bases, análogos de azúcares y/o una cadena principal no nativa y similares), moléculas híbridas de ARN/ARN y ácidos nucleicos poliamídicos (ANP), la totalidad de los cuales puede estar en forma monocatenaria o bicatenaria y, a menos que se limite de otro modo, pueden abarcar análogos conocidos de nucleótidos naturales que pueden funcionar de forma similar a los nucleótidos naturales. Un ácido nucleico puede ser, o puede proceder de, un plásmido, fago, una secuencia de replicación autónoma (ARS), un centrómero, cromosoma artificial, cromosoma u otro ácido nucleico capaz de replicar o replicarse *in vitro* o en una célula huésped, una célula, un núcleo celular o un citoplasma de una célula en determinadas implementaciones. Un molde de ácido nucleico, en algunas implementaciones, 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). A menos que se limite específicamente, la expresión abarca ácidos nucleicos que contienen análogos conocidos de nucleótidos naturales que tienen propiedades de unión similares a las del ácido nucleico de referencia y se metabolizan de manera similar a los nucleótidos naturales. A menos que se indique lo contrario, una secuencia de

ácido nucleico particular también incluye implícitamente variantes modificadas conservativamente de la misma (por ejemplo, sustituciones de codones degenerados), alelos, ortólogos, polimorfismos de un solo nucleótido (SNP) y secuencias complementarias, así como la secuencia explícitamente indicada. Específicamente, las sustituciones de codones degenerados pueden lograrse generando secuencias en las que la tercera posición de uno o más codones seleccionados (o todos) se sustituye con restos de base mixta y/o desoxiinosina. La expresión ácido nucleico se usa indistintamente con locus, gen, ADNc y ARNm codificados por un gen. La expresión también puede incluir, como equivalentes, derivados, variantes y análogos de ARN o ADN sintetizados a partir de análogos de nucleótidos, polinucleótidos monocatenarios ("sentido" o "antisentido", cadena "positiva" o cadena "negativa", marco de lectura "directo" o marco de lectura "inverso") y bicatenarios. El término "gen" significa el segmento de ADN que interviene en la producción de una cadena polipeptídica; incluye regiones que preceden y siguen a la región codificadora (líder y remolque) que interviene en la transcripción/traducción del producto génico y la regulación de la transcripción/traducción, así como secuencias intermedias (intrones) entre segmentos de codificación individuales (exones). Los desoxirribonucleótidos incluyen desoxiadenosina, desoxicitidina, desoxiguanosina y desoxitimidina. Para el ARN, la base citosina se reemplaza por uracilo. Un ácido nucleico molde puede prepararse con el uso de un ácido nucleico obtenido de un sujeto como molde.

Aislamiento y procesamiento de ácidos nucleicos

El ácido nucleico puede derivar de una o más fuentes (por ejemplo, células, suero, plasma, capa leucocítica, líquido linfático, piel, suelo, y similares) mediante métodos conocidos en la técnica. Los ácidos nucleicos a menudo se aíslan de una muestra de prueba. Se puede utilizar cualquier método adecuado para aislar, extraer y/o purificar el ADN de una muestra biológica (por ejemplo, de sangre o de un producto sanguíneo), cuyos ejemplos no limitativos incluyen métodos de preparación de ADN (por ejemplo, los descritos por Sambrook y Russell, "Molecular Cloning: A Laboratory Manual" 3.^a ed., 2001), diferentes reactivos o kits disponibles en el mercado, tal como el kit de ácido nucleico circulante QIAamp de Qiagen, el mini kit de ADN QIAamp o el mini kit de sangre de ADN QIAamp (Qiagen, Hilden, Alemania), kit de aislamiento de ADN de la sangre GenomicPrep™ (Promega, Madison, Wis.) y el kit de purificación de ADN de sangre genómica GFX™ (Amersham, Piscataway, NJ), similares o combinaciones de los mismos.

Los procedimientos y reactivos de lisis celular se conocen en la técnica y pueden realizarse generalmente mediante métodos químicos (por ejemplo, detergentes, disoluciones hipotónicas, procedimientos enzimáticos, y similares, o una combinación de los mismos), físicos (por ejemplo, prensa francesa, sonicación, y similares) o electrolíticos de lisis. Puede utilizarse cualquier procedimiento de lisis adecuado. Por ejemplo, los métodos químicos emplean generalmente agentes de lisis para perturbar las células y extraer los ácidos nucleicos de las células, seguido por el tratamiento con sales caotrópicas. Además, son útiles los métodos físicos, tales como congelación/descongelación seguida de trituración, el uso de prensas celulares y similares. Además, se usan habitualmente procedimientos de lisis con alto contenido de sal. Por ejemplo, puede utilizarse un procedimiento de lisis alcalina. Este último procedimiento incorpora tradicionalmente el uso de disoluciones de fenol-cloroformo, y puede usarse un procedimiento alternativo libre de fenol-cloroformo que involucra tres disoluciones. En los últimos procedimientos, una solución puede contener Tris 15 mM, pH 8,0; EDTA 10 mM y 100 µg/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).

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

El ácido nucleico extracelular puede incluir diferentes especies de ácido nucleico y, por tanto, se denomina en el presente documento "heterogéneo" en determinadas implementaciones. Por ejemplo, el suero o plasma sanguíneo

de una persona que tiene cáncer puede incluir ácido nucleico de células cancerosas y ácido nucleico de células no cancerosas. En otro ejemplo, el suero o plasma sanguíneo de una mujer embarazada puede incluir ácido nucleico materno y ácido nucleico fetal. En algunos casos, el ácido nucleico fetal algunas veces es de aproximadamente el 5 % a aproximadamente el 50 % del ácido nucleico global (por ejemplo, aproximadamente el 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48 o 49 % del ácido nucleico total es ácido nucleico fetal). En algunas implementaciones, la mayoría del ácido nucleico fetal del ácido nucleico tiene una longitud de aproximadamente 500 pares de bases o menor, aproximadamente 250 pares de bases o menor, aproximadamente 200 pares de bases o menor, aproximadamente 150 pares de bases o menor, aproximadamente 100 pares de bases o menor, aproximadamente 50 pares de bases o menor o aproximadamente 25 pares de bases o menor.

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

En algunas implementaciones, los ácidos nucleicos se fragmentan o escinden antes, durante o después de un método descrito en el presente documento. El ácido nucleico fragmentado o escindido puede tener una longitud nominal, media o promedio de aproximadamente 5 a aproximadamente 10.000 pares de bases, de aproximadamente 100 a aproximadamente 1000 pares de bases, de aproximadamente 100 a aproximadamente 500 pares de bases, o aproximadamente 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000 o 9000 pares de bases. Los fragmentos pueden generarse mediante un método adecuado conocido en la técnica, y la longitud promedio, media o nominal de los fragmentos de ácido nucleico puede controlarse mediante la selección de un procedimiento de generación de fragmentos adecuado.

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

En algunas implementaciones, el ácido nucleico se fragmenta o escinde mediante un método adecuado, cuyos ejemplos no limitativos incluyen métodos físicos (por ejemplo, cizalla, por ejemplo, sonicación, prensa francesa, calor, radiación UV, similares), procesos enzimáticos (por ejemplo, agentes de escisión (por ejemplo, una nucleasa adecuada, una enzima de restricción adecuada, una enzima de restricción sensible a la metilación adecuada)), métodos químicos (por ejemplo, alquilación, DMS, piperidina, hidrólisis ácida, hidrólisis básica, calor, similares o combinaciones de los mismos), procesos descritos en la publicación de solicitud de patente estadounidense n.º 20050112590, similares o combinaciones de los mismos.

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

Tal como se usa en el presente documento, “fragmentos”, “productos de escisión”, “productos escindidos” o variantes gramaticales de los mismos, se refieren a moléculas de ácido nucleico resultantes de una fragmentación o escisión de una molécula de gen molde de ácido nucleico o producto amplificado de la misma. Aunque tales fragmentos o productos escindidos pueden referirse a todas las moléculas de ácido nucleico resultantes de una reacción de escisión normalmente tales fragmentos o productos escindidos se refieren solamente a moléculas de ácido nucleico resultantes de una fragmentación o escisión de una molécula de gen molde de ácido nucleico o el segmento de un producto amplificado de la misma que contiene la secuencia de nucleótidos correspondiente de una molécula de gen molde de ácido nucleico. El término “amplificado”, tal como se usa en el presente documento, se refiere a someter un ácido nucleico diana en una muestra a un procedimiento que genera lineal o exponencialmente ácidos nucleicos de amplicón que tienen la misma secuencia de nucleótidos o prácticamente la misma secuencia de nucleótidos que el ácido nucleico diana o segmento del mismo. En determinadas implementaciones, el término “amplificado” se refiere a un método que comprende una reacción en cadena de la polimerasa (PCR). 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 segmentos o partes de moléculas de ácido nucleico amplificadas que contienen, al menos en parte, información de secuencia de nucleótidos de o basada en la molécula molde de ácido nucleico representativa.

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

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 se puede proporcionar en cualquier forma adecuada útil para realizar un análisis de secuencia adecuado.

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

Determinación del contenido de ácido nucleico fetal

La cantidad de ácido nucleico fetal (por ejemplo, concentración, cantidad relativa, cantidad absoluta, número de copias y similares) en ácido nucleico se determina en algunas implementaciones. En determinadas implantaciones, la cantidad de ácido nucleico fetal en una muestra se denomina “fracción fetal”. En algunas implantaciones, “fracción fetal” se refiere a la fracción de ácido nucleico fetal en ácido nucleico circulante libre de células en una muestra (por ejemplo, una muestra de sangre, una muestra de suero, una muestra de plasma) obtenida de una mujer embarazada. En determinadas implementaciones, la cantidad de ácido nucleico fetal se determina según marcadores específicos de un feto de sexo masculino (por ejemplo, marcadores STR del cromosoma Y (por ejemplo, marcadores DYS 19, DYS 385, DYS 392); marcador RhD en mujeres sin expresión de RhD), proporciones alélicas de secuencias polimórficas, o según uno o más marcadores específicos del ácido nucleico fetal y no del ácido nucleico materno (por ejemplo, biomarcadores epigenéticos diferenciales (por ejemplo, metilación; se describe con más detalle a

continuación) entre la madre y el feto, o marcadores de ARN fetal en el plasma sanguíneo materno (véase, por ejemplo, Lo, 2005, "Journal of Histochemistry and Cytochemistry" 53 (3): 293-296)).

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

Un ensayo cuantificador fetal (FQA) puede realizarse junto con cualquiera de los métodos descritos en el presente documento. Tal ensayo puede realizarse mediante cualquier método conocido en la técnica y/o descrito en la publicación de solicitud de patente estadounidense n.º 2010/0105049, tal como, por ejemplo, mediante un método que puede distinguir entre ADN materno y fetal basándose en el estado de metilación diferencial, y cuantificar (por ejemplo, determinar la cantidad de) el ADN fetal. Los métodos para diferenciar ácido nucleico basados en el estado de metilación incluyen, pero no se limitan a, captura sensible a la metilación, por ejemplo, usando un fragmento Fc de MBD2 en el cual el dominio de unión a metilo de MBD2 se fusiona al fragmento Fc de un anticuerpo (MBD-FC) (Gebhard y col. (2006) *Cáncer Res.* 66(12):6118-28); anticuerpos específicos de la metilación; métodos de conversión de bisulfito, por ejemplo, MSP (PCR sensible a la metilación), COBRA, extensión de cebador de un solo nucleótido sensible a la metilación (Ms-SNuPE) o tecnología MassCLEAVE™ de Sequenom; y el uso de enzimas de restricción sensibles a la metilación (por ejemplo, digestión de ADN materno en una muestra materna usando una o más enzimas de restricción sensibles a la metilación, enriqueciendo de este modo el ADN fetal). Las enzimas sensibles a metilo pueden usarse además para diferenciar ácido nucleico basándose en el estado de metilación que, por ejemplo, puede escindir o digerir preferente o sustancialmente en su secuencia de reconocimiento de ADN si esta última no está metilada. Por tanto, se corta una muestra de ADN no metilado en fragmentos más pequeños que una muestra de ADN metilado y no se escinde una muestra de ADN hipermetilado. Excepto cuando se indique explícitamente, puede usarse cualquier método para diferenciar ácido nucleico basado en el estado de metilación con las composiciones y métodos de la tecnología en el presente documento. La cantidad de ADN fetal puede determinarse, por ejemplo, introduciendo uno o más competidores a concentraciones conocidas durante una reacción de amplificación. Además, la determinación de la cantidad de ADN fetal puede realizarse, por ejemplo, mediante RT-PCR, extensión de cebadores, secuenciación y/o recuento. En determinados casos, la cantidad de ácido nucleico puede determinarse usando la tecnología BEAMing tal como se describe en la publicación de solicitud de patente estadounidense n.º 2007/0065823. En determinadas implantaciones, puede determinarse la eficiencia de restricción y el índice de eficiencia se usa para determinar además la cantidad de ADN fetal.

En determinadas implementaciones, se puede usar un ensayo cuantificador fetal (FQA) para determinar la concentración de ADN fetal en una muestra materna, por ejemplo, mediante el siguiente método: a) determinar la cantidad total de ADN presente en una muestra materna; b) digerir selectivamente el ADN materno de una muestra materna utilizando una o más enzimas de restricción sensibles a la metilación, enriqueciendo de este modo el ADN fetal; c) determinar la cantidad de ADN fetal de la etapa b); y d) comparar la cantidad de ADN fetal de la etapa c) con la cantidad total de ADN de la etapa a), determinando de este modo la concentración de ADN fetal de la muestra materna. En determinadas implantaciones, el número absoluto de copias de ácido nucleico fetal en una muestra materna puede determinarse, por ejemplo, mediante espectrometría de masas y/o un sistema que usa un enfoque competitivo de PCR para las mediciones absolutas del número de copias. Véase, por ejemplo, Ding y Cantor (2003) PNAS, EE. UU. 100:3059-3064, y la publicación de solicitud de patente estadounidense n.º 2004/0081993.

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

En determinadas implementaciones, la fracción fetal se puede determinar en función de uno o más niveles. La determinación de la fracción fetal según un nivel se describe, por ejemplo, en la Publicación de Solicitud Internacional

n°. WO 2014/055774. En algunas implementaciones, se determina una fracción fetal según un nivel categorizado como representativo de una variación del número de copias materno y/o fetal. Por ejemplo, determinar la fracción fetal puede comprender evaluar un nivel esperado para una variación del número de copias materno y/o fetal utilizada para la determinación de fracción fetal. En algunas implementaciones, se determina una fracción fetal para un nivel (por ejemplo, un primer nivel) categorizado como representativo de una variación del número de copias según un rango de nivel esperado determinado para el mismo tipo de variación del número de copias. Se puede determinar una fracción fetal según un nivel observado que se encuentra dentro de un rango de nivel esperado y, de este modo, se categoriza como una variación del número de copias materno y/o fetal. En algunas implementaciones, se determina una fracción fetal cuando un nivel observado (por ejemplo, un primer nivel) categorizado como una variación del número de copias materno y/o fetal es diferente del nivel esperado determinado para el mismo tipo de variación del número de copias materno y/o fetal. La fracción fetal puede proporcionarse como un porcentaje. Por ejemplo, una fracción fetal puede dividirse entre 100 proporcionando de ese modo un valor porcentual. Por ejemplo, para un primer nivel representativo de una duplicación homocigota materna y que tiene un nivel de 155 y un nivel esperado para una duplicación homocigota materna que tiene un nivel de 150, puede determinarse una fracción fetal como del 10 % (por ejemplo, (fracción fetal = $2 \times (155 - 150)$)).

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

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

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

Ácidos nucleicos enriquecedores

En algunas implementaciones, el ácido nucleico (por ejemplo, ácido nucleico extracelular) está enriquecido o relativamente enriquecido para una subpoblación o especie de ácido nucleico. Las subpoblaciones de ácidos nucleicos pueden incluir, por ejemplo, ácido nucleico fetal, ácido nucleico materno, ácido nucleico que comprende fragmentos de una longitud o rango de longitudes particular, o ácido nucleico de una región particular del genoma (por ejemplo, cromosoma único, conjunto de cromosomas, y/o determinadas regiones cromosómicas). Tales muestras enriquecidas pueden usarse junto con un método proporcionado en el presente documento. Por tanto, en determinadas implementaciones, los métodos de la tecnología comprenden una etapa adicional de enriquecimiento en una subpoblación de ácido nucleico en una muestra, tal como, por ejemplo, ácido nucleico fetal. En determinadas implementaciones, un método para determinar la fracción fetal descrita anteriormente puede usarse además para

enriquecer el ácido nucleico fetal. En determinadas implementaciones, el ácido nucleico materno se elimina selectivamente (parcial, sustancialmente, casi completamente o completamente) de la muestra. En determinadas implementaciones, el enriquecimiento de un ácido nucleico de especie de bajo número de copias particular (por ejemplo, ácido nucleico fetal) puede mejorar la sensibilidad cuantitativa. Los métodos para enriquecer una muestra para una especie particular de ácido nucleico se describen, por ejemplo, en la patente estadounidense n.º 6.927.028, la publicación de solicitud de patente internacional n.º WO2007/140417, la publicación de solicitud de patente internacional n.º WO2007/147063, la publicación de solicitud de patente internacional n.º WO2009/032779, la publicación de solicitud de patente internacional n.º WO2009/032781, la publicación de solicitud de patente internacional n.º WO2010/033639, la publicación de solicitud de patente internacional n.º WO2011/034631, la publicación de solicitud de patente internacional n.º WO2006/056480 y la publicación de solicitud de patente internacional n.º WO2011/143659.

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

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

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

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

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

identificarse una firma de secuencia a partir de cada perla. Se analiza cada secuencia de firma (etiqueta de MPSS) en un conjunto de datos MPSS, en comparación con todas las demás firmas, y se cuentan todas las firmas idénticas.

- En determinadas implementaciones, determinados métodos de enriquecimiento (por ejemplo, determinados métodos de enriquecimiento basados en MPS y/o MPSS) pueden incluir enfoques basados en la amplificación (por ejemplo, PCR). En determinadas implementaciones, pueden usarse métodos de amplificación específicos de locus (por ejemplo, usando cebadores de amplificación específicos de locus). En determinadas implementaciones, puede usarse un enfoque de PCR múltiplex de alelos con SNP. En determinadas implementaciones, puede usarse un enfoque de PCR múltiplex de alelos con SNP en combinación con secuenciación uniplex. Por ejemplo, tal método puede incluir el uso de PCR múltiplex (por ejemplo, el sistema MASSARRAY) y la incorporación de secuencias de sondas de captura en los amplicones seguido por secuenciación usando, por ejemplo, el sistema Illumina de MPSS. En determinadas implementaciones, puede usarse un método de PCR múltiplex de alelos con SNP junto con un sistema de tres cebadores y secuenciación indexada. Por ejemplo, dicho enfoque puede incluir el uso de PCR múltiplex (por ejemplo, sistema MASSARRAY) con cebadores que tienen una primera sonda de captura incorporada en determinados cebadores de PCR directos específicos de locus y secuencias adaptadoras incorporadas en los cebadores de PCR inversos específicos de locus, para generar de ese modo amplicones, seguido por una PCR secundaria para incorporar las secuencias de captura inversa y los códigos de barras de índice molecular para la secuenciación usando, por ejemplo, el sistema Illumina de MPSS. En determinadas implementaciones, puede usarse un método de PCR múltiplex de alelos con SNP junto con un sistema de cuatro cebadores y secuenciación indexada. Por ejemplo, tal enfoque que puede involucrar el uso de PCR múltiplex (por ejemplo, sistema MASSARRAY) con cebadores que tienen secuencias adaptadoras incorporadas en los cebadores de PCR directos específicos de locus e inversos específicos de locus, seguido de una PCR secundaria para incorporar las secuencias de captura directa e inversa y los códigos de barra de índice molecular para la secuenciación usando, por ejemplo, el sistema Illumina de MPSS. En determinadas implementaciones, puede usarse un enfoque microfluídico. En determinadas implementaciones, puede usarse un enfoque microfluídico basado en alineamientos. Por ejemplo, tal método puede incluir el uso de un alineamiento microfluídico (por ejemplo, Fluidigm) para la amplificación a una multiplicidad baja y la incorporación de sondas de índice y captura, seguido de secuenciación. En determinadas implementaciones, puede usarse un método microfluídico en emulsión, tal como, por ejemplo, PCR digital en gotas.
- En determinadas implementaciones, pueden usarse métodos de amplificación universales (por ejemplo, usando cebadores de amplificación universales o inespecíficos de locus). En determinadas implementaciones, los métodos de amplificación universales pueden usarse en combinación con enfoques de detección de interacciones ("pull-down"). En determinadas implementaciones, un método puede incluir la detección de interacciones de Ultramer biotinilado (por ejemplo, ensayos de detección de interacciones biotiniladas de Agilent o IDT) a partir de una biblioteca de secuenciación amplificada de manera universal. Por ejemplo, tal método puede implicar la preparación de una biblioteca convencional, el enriquecimiento de regiones seleccionadas mediante un ensayo de detección de interacciones y una etapa secundaria de amplificación universal. En determinadas implementaciones, los enfoques de detección de interacciones pueden usarse en combinación con métodos basados en ligamiento. En determinadas implementaciones, un método puede incluir la detección de interacciones de Ultramer biotinilado con ligamiento de adaptador específico de secuencia (por ejemplo, PCR HALOPLEX, Halo Genomics). Por ejemplo, tal método puede incluir el uso de sondas selectoras para capturar fragmentos digeridos por enzimas de restricción, seguido del ligamiento de productos capturados a un adaptador, y la amplificación universal seguida de secuenciación. En determinadas implementaciones, los enfoques de detección de interacciones pueden usarse en combinación con métodos basados en extensión y ligamiento. En determinadas implementaciones, un método puede incluir la extensión y ligamiento de la sonda de inversión molecular (MIP). Por ejemplo, tal enfoque puede incluir el uso de sondas de inversión molecular en combinación con adaptadores de secuencia seguido de amplificación y secuenciación universal. En determinadas implementaciones, el ADN complementario puede sintetizarse y secuenciarse sin amplificación.
- En determinadas implementaciones, los enfoques de extensión y ligamiento pueden realizarse sin un componente de detección de interacciones. En determinadas implementaciones, un método puede incluir hibridación de cebadores directos e inversos específicos de locus, extensión y ligamiento. Tales métodos pueden incluir además amplificación universal o síntesis de ADN complementario sin amplificación, seguida de secuenciación. Tales métodos pueden reducir o excluir secuencias de fondo durante el análisis, en determinadas implementaciones.
- En determinadas implementaciones, los enfoques de detección de interacciones pueden usarse con un componente de amplificación opcional o sin componente de amplificación. En determinadas implementaciones, un método puede incluir un ensayo de detección de interacciones modificado y ligamiento con incorporación completa de sondas de captura sin amplificación universal. Por ejemplo, tal método puede incluir el uso de sondas selectoras modificadas para capturar fragmentos digeridos por enzimas de restricción, seguido por el ligamiento de productos capturados a un adaptador, amplificación opcional y secuenciación. En determinadas implementaciones, un método puede incluir un ensayo de detección de interacciones biotiniladas con extensión y ligamiento de la secuencia adaptadora en combinación con ligamiento monocatenario circular. Por ejemplo, tal enfoque puede incluir el uso de sondas selectoras para capturar regiones de interés (por ejemplo, secuencias diana), extensión de las sondas, ligamiento de adaptador, ligamiento monocatenario circular, amplificación opcional y secuenciación. En determinadas implementaciones, el análisis del resultado de la secuenciación puede separar las secuencias diana del fondo.

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

En algunas implementaciones, se usa un procedimiento de captura de ácido nucleico selectivo para separar los fragmentos diana y/o de referencia de la muestra de ácido nucleico. Los sistemas de captura de ácidos nucleicos disponibles en el mercado incluyen, por ejemplo, el sistema de captura de secuencias Nimblegen (Roche NimbleGen, Madison, WI); la plataforma BEADARRAY de Illumina (Illumina, San Diego, CA); la plataforma GENECHIP de Affymetrix (Affymetrix, Santa Clara, CA); el sistema de enriquecimiento de diaans SureSelect de Agilent (Agilent Technologies, Santa Clara, CA); y plataformas relacionadas. Tales métodos implican normalmente la hibridación de un oligonucleótido de captura a un segmento o toda la secuencia de nucleótidos de un fragmento diana o de referencia y pueden incluir el uso de una fase sólida (por ejemplo, alineamiento de fase sólida) y/o una plataforma a base de disolución. Los oligonucleótidos de captura (a veces denominados “cebo”) pueden seleccionarse o diseñarse de tal manera que se hibriden preferiblemente con fragmentos de ácido nucleico de regiones o locus genómicos seleccionados (por ejemplo, uno de los cromosomas 21, 18, 13, X o Y, o un cromosoma de referencia). En determinadas implementaciones, puede utilizarse un método basado en la hibridación (por ejemplo, utilizando matrices de oligonucleótidos) para enriquecer en las secuencias de ácidos nucleicos de determinados cromosomas (por ejemplo, un cromosoma potencialmente aneuploide, un cromosoma de referencia u otro cromosoma de interés) o segmentos de interés de los mismos.

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

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

nucleico monocatenario y amplificación de los fragmentos marcados con etiqueta usando cebadores de amplificación específicos para al menos una etiqueta.

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

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

Biblioteca de ácidos nucleicos

En algunas implementaciones, una biblioteca de ácidos nucleicos es una pluralidad de moléculas de polinucleótidos (por ejemplo, una muestra de ácidos nucleicos) que se preparan, ensamblan y/o modifican para un proceso específico, cuyos ejemplos no limitativos incluyen la inmovilización en una fase sólida (por ejemplo, un soporte sólido, por ejemplo, una celda de flujo, una perla), enriquecimiento, amplificación, clonación, detección y/o secuenciación de ácidos nucleicos. En determinadas implementaciones, se prepara una biblioteca de ácidos nucleicos antes o durante un proceso de secuenciación. Una biblioteca de ácidos nucleicos (por ejemplo, una biblioteca de secuenciación) se puede preparar mediante un método adecuado como se conoce en la técnica. Una biblioteca de ácidos nucleicos se puede preparar mediante un proceso de preparación dirigido o no dirigido.

En algunas implementaciones, una biblioteca de ácidos nucleicos se modifica para que comprenda un resto químico (por ejemplo, un grupo funcional) configurado para la inmovilización de ácidos nucleicos en un soporte sólido. En algunas implementaciones, una biblioteca de ácidos nucleicos se modifica para que comprenda una biomolécula (por ejemplo, un grupo funcional) y/o un miembro de un par de unión configurado para la inmovilización de la biblioteca en un soporte sólido, cuyos ejemplos no limitativos incluyen globulina de unión a tiroxina, proteínas de unión a esteroides, anticuerpos, antígenos, haptenos, enzimas, lectinas, ácidos nucleicos, represores, proteína A, proteína G, avidina, estreptavidina, biotina, componente del complemento C1q, proteínas de unión a ácidos nucleicos, receptores, carbohidratos, oligonucleótidos, polinucleótidos, secuencias de ácidos nucleicos complementarias, similares y combinaciones de los mismos. Algunos ejemplos de pares de unión específicos incluyen, sin limitación: un resto de avidina y un resto de biotina; un epítopo antigénico y un anticuerpo o fragmento inmunológicamente reactivo del mismo; un anticuerpo y un hapteno; un resto de digoxigeno y un anticuerpo anti-digoxigeno; un resto de fluoresceína y un anticuerpo anti-fluoresceína; un operador y un represor; una nucleasa y un nucleótido; una lectina y un polisacárido; un esteroide y una proteína de unión a esteroides; un compuesto activo y un receptor de compuesto activo; una hormona y un receptor de hormonas; una enzima y un sustrato; una inmunoglobulina y proteína A; un oligonucleótido o polinucleótido y su complemento correspondiente; similares o combinaciones de los mismos.

En algunas implementaciones, una biblioteca de ácidos nucleicos se modifica para que comprenda uno o más polinucleótidos de composición conocida, cuyos ejemplos no limitativos incluyen un identificador (por ejemplo, una etiqueta, una etiqueta de indexación), una secuencia de captura, un marcador, un adaptador, un sitio de enzima de restricción, un promotor, un potenciador, un origen de replicación, una estructura de horquilla, una secuencia complementaria (por ejemplo, un sitio de unión del cebador, un sitio de hibridación), un sitio de integración adecuado (por ejemplo, un transposón, un sitio de integración vírica), un nucleótido modificado, similares o combinaciones de los mismos. Se pueden añadir polinucleótidos de secuencia conocida en una posición adecuada, por ejemplo, en el extremo 5', el extremo 3' o dentro de una secuencia de ácido nucleico. Los polinucleótidos de secuencia conocida pueden ser secuencias iguales o diferentes. En algunas implementaciones, un polinucleótido de secuencia conocida se configura para hibridarse con uno o más oligonucleótidos inmovilizados en una superficie (por ejemplo, una superficie en una celda de flujo). Por ejemplo, una molécula de ácido nucleico que comprende una secuencia conocida en 5' puede hibridarse con una primera pluralidad de oligonucleótidos mientras que la secuencia conocida en 3' puede hibridarse con una segunda pluralidad de oligonucleótidos. En algunas implementaciones, una biblioteca de ácido nucleico puede comprender etiquetas específicas de cromosomas, secuencias de captura, marcadores y/o adaptadores. En algunas implementaciones, una biblioteca de ácidos nucleicos comprende uno o más marcadores detectables. En algunas implementaciones, se pueden incorporar uno o más marcadores detectables a una biblioteca de ácidos nucleicos en un extremo 5', en un extremo 3' y/o en cualquier posición de nucleótido dentro de un ácido nucleico en la biblioteca. En algunas implementaciones, una biblioteca de ácidos nucleicos comprende oligonucleótidos hibridados. En determinadas implementaciones, los oligonucleótidos hibridados son sondas marcadas. En algunas implementaciones, una biblioteca de ácidos nucleicos comprende sondas de oligonucleótidos hibridadas antes de la inmovilización en una fase sólida.

En algunas implementaciones, un polinucleótido de secuencia conocida comprende una secuencia universal. Una secuencia universal es una secuencia de ácido nucleico específica que se integra en dos o más moléculas de ácido nucleico, o dos o más subconjuntos de moléculas de ácido nucleico, donde la secuencia universal es la misma para todas las moléculas o subconjuntos de moléculas en las que se integra. Una secuencia universal se suele diseñar para hibridar y/o amplificar una pluralidad de secuencias diferentes utilizando un único cebador universal que sea complementario a una secuencia universal. En algunas implementaciones, se utilizan dos (por ejemplo, un par) o más secuencias universales y/o cebadores universales. Un cebador universal suele comprender una secuencia universal. En algunas implementaciones, los adaptadores (por ejemplo, adaptadores universales) comprenden secuencias universales. En algunas implementaciones se utilizan una o más secuencias universales para capturar, identificar y/o detectar múltiples especies o subconjuntos de ácidos nucleicos.

En determinadas implementaciones de la preparación de una biblioteca de ácidos nucleicos (por ejemplo, en determinados procedimientos de secuenciación mediante síntesis), los ácidos nucleicos se seleccionan según el tamaño y/o se fragmentan en longitudes de varios cientos de pares de bases o menores (por ejemplo, en la preparación para la generación de bibliotecas). En algunas implementaciones, la preparación de la biblioteca se realiza sin fragmentación (por ejemplo, cuando se usa ADNfcc).

En determinadas implementaciones, se usa un método de preparación de bibliotecas basado en ligamiento (por ejemplo, ILLUMINA TRUSEQ, Illumina, San Diego CA). Los métodos de preparación de bibliotecas basados en ligamiento suelen utilizar un diseño (por ejemplo, un adaptador metilado) que puede incorporar una secuencia índice en la etapa de ligamiento inicial y a menudo pueden usarse para preparar muestras para secuenciación de una sola lectura, secuenciación de extremos apareados y secuenciación multiplexada. Por ejemplo, a veces los ácidos nucleicos (por ejemplo, ácidos nucleicos fragmentados o ADNfcc) se reparan en los extremos mediante una reacción de relleno, una reacción de exonucleasa o una combinación de las mismas. En algunas implementaciones, el ácido nucleico reparado de extremo romo resultante puede extenderse luego por un solo nucleótido que sea complementario a un solo nucleótido saliente en el extremo 3' de un adaptador/cebador. Puede usarse cualquier nucleótido para los nucleótidos de extensión/saliente. En algunas implementaciones, la preparación de la biblioteca de ácidos nucleicos comprende ligar un oligonucleótido adaptador. Los oligonucleótidos adaptadores suelen ser complementarios a los anclajes de las celdas de flujo y, a veces, se utilizan para inmovilizar una biblioteca de ácidos nucleicos en un soporte sólido, como la superficie interior de una celda de flujo, por ejemplo. En algunas implementaciones, un oligonucleótido adaptador comprende un identificador, uno o más sitios de hibridación de cebadores de secuenciación (por ejemplo, secuencias complementarias a los cebadores de secuenciación universales, cebadores de secuenciación de un solo extremo, cebadores de secuenciación de extremos apareados, cebadores de secuenciación multiplexados, y similares), o combinaciones de los mismos (por ejemplo, adaptador/secuenciación, adaptador/identificador, adaptador/identificador/secuenciación).

Un identificador puede ser un marcador detectable adecuado incorporado o unido a un ácido nucleico (por ejemplo, un polinucleótido) que permita la detección y/o identificación de ácidos nucleicos que comprendan el identificador. En algunas implementaciones, un identificador se incorpora o se une a un ácido nucleico durante un método de secuenciación (por ejemplo, mediante una polimerasa). Los ejemplos no limitativos de identificadores incluyen etiquetas de ácido nucleico, índices de ácido nucleico o códigos de barras, un marcador radioactivo (por ejemplo, un isótopo), un marcador metálico, un marcador fluorescente, un marcador quimioluminiscente, un marcador fosforescente, un inactivador de fluoróforo, un colorante, una proteína (por ejemplo, una enzima, un anticuerpo o parte del mismo, un conector, un miembro de un par de unión), similares o combinaciones de los mismos. En algunas implementaciones, un identificador (por ejemplo, un índice de ácido nucleico o un código de barras) es una secuencia única, conocida y/o identificable de nucleótidos o análogos de nucleótidos. En algunas implementaciones, los identificadores son seis o más nucleótidos contiguos. Existe una multitud de fluoróforos con una variedad de diferentes espectros de excitación y emisión. Se puede usar cualquier tipo y/o número adecuado de fluoróforos como identificador. En algunas implementaciones, se usan 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, 20 o más, 30 o más, o 50 o más identificadores diferentes en un método descrito en el presente documento (por ejemplo, un método de detección y/o secuenciación de ácidos nucleicos). En algunas implementaciones, uno o dos tipos de identificadores (por ejemplo, marcadores fluorescentes) están vinculados a cada ácido nucleico de una biblioteca. La detección y/o cuantificación de un identificador puede realizarse mediante un método, una máquina o un aparato adecuado, cuyos ejemplos no limitativos incluyen citometría de flujo, reacción en cadena de la polimerasa cuantitativa (qPCR), electroforesis en gel, un luminómetro, un fluorómetro, un espectrofotómetro, un análisis de genochips o micromatrices adecuado, transferencia Western, espectrometría de masas, cromatografía, análisis citofluorimétrico, microscopía de fluorescencia, un método de generación de imágenes digitales o de fluorescencia adecuado, microscopía de barrido láser confocal, citometría de barrido láser, cromatografía de afinidad, separación manual por lotes, suspensión de campo eléctrico, un método adecuado de secuenciación de ácidos nucleicos y/o un aparato de secuenciación de ácidos nucleicos, similares y combinaciones de los mismos.

En algunas implementaciones, se usa un método de preparación de bibliotecas basado en transposones (por ejemplo, EPICENTRE NEXTERA, Epicentre, Madison WI). Los métodos basados en transposones usan normalmente la transposición *in vitro* para fragmentar y marcar con etiqueta simultáneamente ADN en una reacción de un solo tubo (que a menudo permite la incorporación de etiquetas específicas de plataforma y códigos de barras opcionales), y preparar bibliotecas listas para el secuenciador.

En algunas implementaciones, se amplifica una biblioteca de ácidos nucleicos o partes de la misma (por ejemplo, se amplifica mediante un método basado en PCR). En algunas implementaciones, un método de secuenciación comprende la amplificación de una biblioteca de ácidos nucleicos. Una biblioteca de ácidos nucleicos se puede amplificar antes o después de la inmovilización en un soporte sólido (por ejemplo, un soporte sólido en una celda de flujo). La amplificación de ácidos nucleicos incluye el proceso de amplificación o aumento del número de un molde de ácido nucleico y/o de un complemento del mismo que están presentes (por ejemplo, en una biblioteca de ácidos nucleicos), mediante la producción de una o más copias del molde y/o su complemento. La amplificación puede llevarse a cabo mediante un método adecuado. Una biblioteca de ácidos nucleicos puede amplificarse mediante un método de termociclado o mediante un método de amplificación isotérmico. En algunas implementaciones, se utiliza un método de amplificación de círculo rodante. En algunas implementaciones, la amplificación tiene lugar en un soporte sólido (por ejemplo, dentro de una celda de flujo) donde se inmoviliza una biblioteca de ácidos nucleicos o una porción de la misma. En determinados métodos de secuenciación, se añade una biblioteca de ácidos nucleicos a una celda de flujo y se inmoviliza mediante hibridación con anclajes en condiciones adecuadas. Este tipo de amplificación de ácidos nucleicos se suele denominar amplificación en fase sólida. En algunas implementaciones de la amplificación en fase sólida, la totalidad o una porción de los productos amplificados se sintetizan mediante una extensión que se inicia a partir de un cebador inmovilizado. Las reacciones de amplificación en fase sólida son análogas a las amplificaciones en fase de solución convencionales excepto que al menos uno de los oligonucleótidos de amplificación (por ejemplo, cebadores) se inmoviliza en un soporte sólido.

En algunas implementaciones, la amplificación en fase sólida comprende una reacción de amplificación de ácido nucleico que comprende solo una especie de cebador oligonucleotídico inmovilizado en una superficie. En determinadas implementaciones, la amplificación en fase sólida comprende una pluralidad de diferentes especies de cebadores oligonucleotídicos inmovilizados. En algunas implementaciones, la amplificación en fase sólida puede comprender una reacción de amplificación de ácido nucleico que comprende una especie de cebador oligonucleotídico inmovilizado sobre una superficie sólida y una segunda especie de cebador oligonucleotídico diferente en solución. Se pueden usar múltiples especies diferentes de cebadores inmovilizados o basados en solución. Los ejemplos no limitativos de reacciones de amplificación de ácidos nucleicos en fase sólida incluyen amplificación interfacial, amplificación puente, PCR en emulsión, amplificación WildFire (por ejemplo, publicación de patente estadounidense US20130012399), similares o combinaciones de las mismas.

Secuenciación

En algunas implementaciones, se secuencian ácidos nucleicos (por ejemplo, fragmentos de ácido nucleico, ácido nucleico de muestra, ácido nucleico libre de células). En determinadas implementaciones, se obtiene una secuencia completa o sustancialmente completa y, algunas veces, se obtiene una secuencia parcial.

En algunas implementaciones, algunos o todos los ácidos nucleicos de una muestra se enriquecen y/o amplifican (por ejemplo, de forma no específica, por ejemplo, mediante un método basado en PCR) antes o durante la secuenciación. En determinadas implementaciones, las porciones o subconjuntos de ácidos nucleicos específicos de una muestra se enriquecen y/o amplifican antes o durante la secuenciación. En algunas implementaciones, una porción o un subconjunto de una combinación preseleccionada de ácidos nucleicos se secuencia aleatoriamente. En algunas implementaciones, los ácidos nucleicos de una muestra no se enriquecen y/o amplifican antes o durante la secuenciación.

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

La longitud de una lectura de secuencia se asocia a menudo con la tecnología de secuenciación particular. Los métodos de alto rendimiento, por ejemplo, proporcionan lecturas de secuencia que pueden variar en tamaño desde decenas hasta cientos de pares de bases (pb). La secuenciación por nanoporos, por ejemplo, puede proporcionar lecturas de secuencia que pueden variar en tamaño desde decenas hasta cientos y hasta miles de pares de bases. En algunas implementaciones, las lecturas de secuencia tienen una longitud media, promedio o absoluta, o mediana de la longitud, de aproximadamente 15 pb a aproximadamente 900 pb de longitud. En determinadas implementaciones, las lecturas de secuencia tienen una longitud media, promedio o absoluta, o mediana de la longitud, de aproximadamente 1000 pb o más.

En determinadas implementaciones, la longitud nominal, promedio, media o absoluta de las lecturas de un solo extremo es a veces de aproximadamente 15 nucleótidos contiguos a aproximadamente 50 nucleótidos contiguos o más, de aproximadamente 15 nucleótidos contiguos a aproximadamente 40 nucleótidos contiguos o más, y a veces de aproximadamente 15 nucleótidos contiguos a aproximadamente 36 nucleótidos contiguos o más. En determinadas implementaciones, la longitud nominal, promedio, media o absoluta de las lecturas de un solo extremo es de aproximadamente 20 a aproximadamente 30 bases, o de aproximadamente 24 a aproximadamente 28 bases de

longitud. En determinadas implementaciones, la longitud nominal, promedio, media o absoluta de las lecturas de un solo extremo es de aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 28 o aproximadamente 29 bases o más de longitud.

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

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

En algunas implementaciones, se secuencia una fracción representativa de un genoma y, a veces, se denomina "cobertura" o "cobertura de pliegues". Por ejemplo, una cobertura de 1 vez indica que aproximadamente el 100 % de las secuencias de nucleótidos del genoma están representadas por lecturas. En algunas implementaciones, "las veces de cobertura" es una expresión relativa que se refiere a una ejecución de secuenciación anterior como referencia. Por ejemplo, una segunda ejecución de secuenciación puede tener la mitad de cobertura que una primera ejecución de secuenciación. En algunas implementaciones, se secuencia un genoma con redundancia, donde una región dada del genoma puede estar cubierta por dos o más lecturas o lecturas superpuestas (por ejemplo, "veces de cobertura" mayor de 1, por ejemplo, 2 veces de cobertura).

En algunas implementaciones, se secuencia una muestra de ácido nucleico de un individuo. En determinadas implementaciones, se secuencian los ácidos nucleicos de cada una de dos o más muestras, donde las muestras son de un individuo o de diferentes individuos. En determinadas implementaciones, las muestras de ácido nucleico de dos o más muestras biológicas se combinan, donde cada muestra biológica es de un individuo o de dos o más individuos, y se secuencia la combinación. En estas últimas implementaciones, una muestra de ácido nucleico de cada muestra biológica suele identificarse mediante uno o más identificadores únicos.

En algunas implementaciones, un método de secuenciación utiliza identificadores que permiten la multiplexación de reacciones de secuencia en un proceso de secuenciación. Cuanto mayor sea el número de identificadores únicos, mayor será el número de muestras y/o cromosomas para la detección, por ejemplo, que puedan multiplexarse en un proceso de secuenciación. Un proceso de secuenciación puede realizarse utilizando cualquier número adecuado de identificadores únicos (por ejemplo, 4, 8, 12, 24, 48, 96 o más).

Un proceso de secuenciación a veces hace uso de una fase sólida y, a veces, la fase sólida comprende una celda de flujo en la que se puede unir el ácido nucleico de una biblioteca y los reactivos pueden fluir y ponerse en contacto con el ácido nucleico anexo. Una celda de flujo a veces incluye carriles de celdas de flujo, y el uso de identificadores puede facilitar el análisis de varias muestras en cada carril. 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 algunas implementaciones, el número de muestras analizadas en un carril de celda de flujo determinado depende del número de identificadores únicos utilizados durante la preparación de la biblioteca y/o el carril de celda de flujo único del diseño de la sonda. La multiplexación usando 12 identificadores, por ejemplo, permite el análisis simultáneo de 96 muestras (por ejemplo, igual al número de pocillos en una placa de micropocillos de 96 pocillos) en una celda de flujo de 8 carriles. De manera similar, la multiplexación usando 48 identificadores, por ejemplo, permite el análisis simultáneo de 384 muestras (por ejemplo, igual al número de pocillos en una placa de micropocillos de 384 pocillos) en una celda de flujo de 8 carriles. Los ejemplos no limitativos de kits de secuenciación múltiplex disponibles comercialmente incluyen el kit de oligonucleótidos para la preparación de muestras de multiplexación de Illumina y los cebadores de secuenciación de multiplexación y el kit de control PhiX (por ejemplo, números de catálogo de Illumina PE-400-1001 y PE-400-1002, respectivamente).

Puede usarse cualquier método adecuado de secuenciación de ácidos nucleicos, cuyos ejemplos no limitativos incluyen Maxim & Gilbert, métodos de terminación de cadena, secuenciación por síntesis, secuenciación por ligación, secuenciación por espectrometría de masas, técnicas basadas en microscopía, similares o combinaciones de los mismos. En algunas implementaciones, se puede utilizar una tecnología de primera generación tal como, por ejemplo, los métodos de secuenciación de Sanger, incluidos los métodos automatizados de secuenciación de Sanger, incluida la secuenciación de Sanger microfluídica, en un método proporcionado en el presente documento. En algunas implementaciones, se pueden utilizar tecnologías de secuenciación que incluyan el uso de tecnologías de obtención de imágenes de ácidos nucleicos (por ejemplo, microscopía electrónica de transmisión (TEM) y microscopía de fuerza atómica (AFM)). En algunas implementaciones, se usa un método de secuenciación de alto rendimiento. Los métodos de secuenciación de alto rendimiento generalmente incluyen moldes de ADN amplificados clonalmente o moléculas únicas de ADN que se secuencian de forma masiva en paralelo, a veces dentro de una celda de flujo. Las técnicas de secuenciación de próxima generación (por ejemplo, 2.^a y 3.^a generación) capaces de secuenciar el ADN de forma masiva en paralelo se pueden utilizar para los métodos descritos en el presente documento y se denominan colectivamente en el presente documento "secuenciación masiva en paralelo" (MPS). En algunas implementaciones, los métodos de secuenciación MPS utilizan un enfoque dirigido, donde se secuencian cromosomas, genes o regiones de interés específicos. En determinadas implementaciones, se utiliza un enfoque no dirigido donde la mayoría o todos los ácidos nucleicos de una muestra se secuencian, amplifican y/o capturan aleatoriamente.

En algunas implementaciones se utiliza un enfoque de enriquecimiento, amplificación y/o secuenciación dirigido. Un enfoque dirigido suele aislar, seleccionar y/o enriquecer un subconjunto de ácidos nucleicos de una muestra para su posterior procesamiento mediante el uso de oligonucleótidos específicos de la secuencia. En algunas implementaciones, se utiliza una biblioteca de oligonucleótidos específicos de la secuencia para dirigirse a (por ejemplo, hibridar con) uno o más conjuntos de ácidos nucleicos de una muestra. Los oligonucleótidos y/o cebadores específicos de la secuencia suelen ser selectivos para secuencias particulares (por ejemplo, secuencias de ácido nucleico únicas) presentes en uno o más cromosomas, genes, exones, intrones y/o regiones reguladoras de interés. Puede usarse cualquier método adecuado o combinación de métodos para el enriquecimiento, amplificación y/o secuenciación de uno o más subconjuntos de ácidos nucleicos diana. En algunas implementaciones, las secuencias diana se aíslan y/o enriquecen mediante la captura en una fase sólida (por ejemplo, una celda de flujo, una perla) utilizando uno o más anclajes específicos de la secuencia. En algunas implementaciones, las secuencias diana se enriquecen y/o amplifican mediante un método basado en polimerasa (por ejemplo, un método basado en PCR, mediante cualquier extensión adecuada basada en polimerasa) utilizando cebadores y/o conjuntos de cebadores específicos de la secuencia. Los anclajes específicos de la secuencia se suelen poder utilizar como cebadores específicos de la secuencia.

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

La secuenciación por síntesis, en algunas implementaciones, comprende la adición iterativa (por ejemplo, por adición covalente) de un nucleótido a un cebador o cadena de ácido nucleico preexistente de manera dirigida por molde. Cada adición iterativa de un nucleótido se detecta y el proceso se repite varias veces hasta que se obtiene una secuencia de una cadena de ácido nucleico. La longitud de una secuencia obtenida depende, en parte, del número de etapas de adición y detección que se realicen. En algunas implementaciones de secuenciación por síntesis, se añaden y detectan uno, dos, tres o más nucleótidos del mismo tipo (por ejemplo, A, G, C o T) en una serie de adición de nucleótidos. Los nucleótidos se pueden añadir mediante cualquier método adecuado (por ejemplo, enzimática o químicamente). Por ejemplo, en algunas implementaciones, una polimerasa o una ligasa añade un nucleótido a un cebador o a una cadena de ácido nucleico preexistente de manera dirigida por el molde. En algunas implementaciones de secuenciación por síntesis, se utilizan diferentes tipos de nucleótidos, análogos de nucleótidos y/o identificadores. En algunas implementaciones se utilizan terminadores reversibles y/o identificadores extraíbles (por ejemplo, escindibles). En algunas implementaciones se utilizan nucleótidos y/o análogos de nucleótidos marcados con fluorescencia. En determinadas implementaciones, la secuenciación por síntesis comprende una escisión (por ejemplo, escisión y extracción de un identificador) y/o una etapa de lavado. En algunas implementaciones, la adición de uno o más nucleótidos se detecta mediante un método adecuado descrito en el presente documento o conocido en la técnica, cuyos ejemplos no limitativos incluyen cualquier aparato de formación de imágenes adecuado, una cámara adecuada, una cámara digital, un CCD (Dispositivo de par de carga) (por ejemplo, una cámara CCD), un aparato de formación de imágenes basado en CMOS (silicio de óxido de metal complementario) (por ejemplo, una cámara CMOS), un fotodiodo (por ejemplo, un tubo fotomultiplicador), microscopía electrónica, un transistor de efecto de campo (por

ejemplo, un transistor de efecto de campo de ADN), un sensor de iones ISFET (por ejemplo, un sensor CHEMFET), similares o combinaciones de los mismos. Otros métodos de secuenciación que pueden usarse para llevar a cabo los métodos en el presente documento incluyen PCR digital y secuenciación por hibridación.

Otros métodos de secuenciación que pueden usarse para llevar a cabo los métodos en el presente documento incluyen PCR digital y secuenciación por hibridación. La reacción en cadena de la polimerasa digital (PCR digital o dPCR) puede usarse para identificar y cuantificar directamente los ácidos nucleicos en una muestra. La PCR digital puede llevarse a cabo en una emulsión, en algunas implementaciones. Por ejemplo, los ácidos nucleicos individuales se separan, por ejemplo, en un dispositivo de cámara microfluídico, y cada ácido nucleico se amplifica individualmente mediante PCR. Los ácidos nucleicos pueden separarse de tal manera que no haya más de un ácido nucleico por pocillo. En algunas implementaciones, pueden usarse sondas diferentes para distinguir diversos alelos (por ejemplo, alelos fetales y alelos maternos). Pueden enumerarse alelos para determinar el número de copias.

En determinadas implementaciones, se puede utilizar 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 algunas implementaciones. El patrón de hibridación al alineamiento puede usarse para determinar las secuencias de polinucleótidos presentes en la muestra. En algunas implementaciones, cada sonda se conecta a una perla, por ejemplo, una perla magnética o similar. La hibridación a las perlas puede identificarse y usarse para identificar la pluralidad de secuencias de polinucleótidos dentro de la muestra.

En algunas implementaciones, la secuenciación por nanoporos puede usarse en un método descrito en el presente documento. La secuenciación por nanoporos es una tecnología de secuenciación de una sola molécula mediante la cual una sola molécula de ácido nucleico (por ejemplo, ADN) se secuencia directamente a medida que pasa a través de un nanoporo.

Se puede utilizar un método, sistema o plataforma tecnológica de MPS adecuado para llevar a cabo los métodos descritos en el presente documento para obtener lecturas de secuencias de ácidos nucleicos. Los ejemplos no limitativos de plataformas MPS incluyen Illumina/Solex/HiSeq (por ejemplo, el analizador genómico de Illumina; analizador genómico II; HISEQ 2000; HISEQ), SOLiD, Roche/454, PACBIO y/o SMRT, Helicos True Single Molecule Sequencing, Ion Torrent y la secuenciación basada en el semiconductor Ion (por ejemplo, desarrollado por Life Technologies), Wild Fire, 5500, 5500xl W y/o tecnologías basadas en el analizador genético 5500xl W (por ejemplo, desarrolladas y comercializadas por Life Technologies, publicación de patente de EE. UU. n.º US20130012399); Secuenciación de Colonia, pirosecuenciación, secuenciación masiva en paralelo (MPSS), secuenciación de ARN polimerasa (RNAP), sistemas y métodos LaserGen, plataformas basadas en nanoporos, matriz de transistores de efecto de campo sensible a productos químicos (CHEMFET), secuenciación basada en microscopía electrónica (por ejemplo, como la desarrollada por ZS Genetics, Halcyon Molecular), secuenciación de nanobolas, similares o combinaciones de los mismos.

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

En algunas implementaciones, se generan, obtienen, reúnen, ensamblan, manipulan, transforman, procesan y/o proporcionan lecturas de secuencia por un módulo de secuencia. Una máquina que comprende un módulo de secuencias puede ser una máquina y/o un aparato adecuado que determine la secuencia de un ácido nucleico utilizando una tecnología de secuenciación conocida en la técnica. En algunas implementaciones, un módulo de secuencias puede alinear, ensamblar, fragmentar, complementar, complementar de manera inversa, y/o comprobar errores (por ejemplo, corregir errores en las lecturas de secuencias).

En algunas implementaciones, las lecturas de secuencia de nucleótidos obtenidas de una muestra son lecturas de secuencia de nucleótidos parciales. Como se usa en el presente documento, "lecturas de secuencia de nucleótidos parciales" se refiere a lecturas de secuencia de cualquier longitud con información de secuencia incompleta, también denominada ambigüedad de secuencia. Las lecturas de secuencia de nucleótidos parciales pueden carecer de información sobre identidad de nucleobase y/o posición u orden de nucleobase. Las lecturas de secuencia de nucleótidos parciales generalmente no incluyen lecturas de secuencia en las que la única información de secuencia incompleta (o en la que menos de todas las bases se secuencian o determinan) es de errores de secuenciación inadvertidos o no intencionales. Dichos errores de secuenciación pueden ser inherentes a determinados procesos de

secuenciación e incluyen, por ejemplo, identificaciones incorrectas para identidad de nucleobase y nucleobases faltantes o adicionales. Por lo tanto, para las lecturas de secuencia de nucleótidos parciales en el presente documento, determinada información sobre la secuencia a menudo se excluye deliberadamente. Es decir, se obtiene deliberadamente información de secuencia con respecto a menos de todas las nucleobases o que de cualquier otra manera podría caracterizarse como o ser un error de secuenciación. En algunas implementaciones, una lectura de secuencia de nucleótidos parcial puede abarcar una porción de un fragmento de ácido nucleico. En algunas implementaciones, una lectura de secuencia de nucleótidos parcial puede abarcar la longitud completa de un fragmento de ácido nucleico. Las lecturas de secuencia de nucleótidos parciales se describen, por ejemplo, en la publicación de solicitud de patente internacional n°. WO 2013/052907.

Lecturas de mapeo

Las lecturas de secuencia pueden mapearse. Se puede usar cualquier método de mapeo adecuado (por ejemplo, proceso, algoritmo, programa, software, módulo, similares o una combinación de los mismos) y determinados aspectos de los procesos de mapeo se describen a continuación.

El mapeo de las lecturas de secuencia de nucleótidos (por ejemplo, la información de la secuencia de un fragmento cuya posición genómica física se desconoce) puede realizarse de varias maneras, y suele comprender la alineación de las lecturas de secuencia obtenidas con una secuencia coincidente en un genoma de referencia. En dichas alineaciones, las lecturas de secuencia generalmente se alinean con una secuencia de referencia y las que se alinean se designan como “mapeadas”, “una lectura de secuencia mapeada” o “una lectura mapeada”.

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

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

En algunas implementaciones, las lecturas de secuencia mapeadas y/o la información asociada con una lectura de secuencia mapeada se almacenan en y/o se acceden desde un medio de almacenamiento legible por ordenador no transitorio en un formato legible por ordenador adecuado. Un “formato legible por ordenador” a veces se denomina generalmente en el presente documento un formato. En algunas implementaciones, las lecturas de secuencia mapeadas se almacenan y/o se acceden en un formato binario adecuado, un formato de texto, similares o una combinación de los mismos. Un formato binario a veces es un formato BAM. Un formato de texto a veces es un formato de alineación/mapa de secuencia (SAM). Los ejemplos no limitantes de formatos binarios y/o de texto incluyen BAM, SAM, SRF, FASTQ, Gzip, similares, o combinaciones de los mismos. En algunas implementaciones, las lecturas de secuencia mapeadas se almacenan en y/o se convierten en un formato que requiere menos espacio de almacenamiento (por ejemplo, menos bytes) que un formato tradicional (por ejemplo, un formato SAM o un formato BAM). En algunas implementaciones, las lecturas de secuencia mapeadas en un primer formato se comprimen en un segundo formato que requiere menos espacio de almacenamiento que el primer formato. El término “comprimido” como se usa en el presente documento se refiere a un proceso de compresión de datos, codificación de origen y/o reducción de velocidad de bits donde un archivo de datos legible por ordenador se reduce en tamaño. En algunas implementaciones, las lecturas de secuencia mapeadas se comprimen de un formato SAM en un formato binario. Algunos datos a veces se pierden después de que se comprime un archivo. A veces no se pierden datos en un proceso de compresión. En algunas implementaciones de compresión de archivos, algunos datos se reemplazan con un índice y/o una referencia a otro archivo de datos que comprende información sobre una lectura de secuencia mapeada. En algunas implementaciones, una lectura de secuencia mapeada se almacena en un formato binario que comprende o

consiste en un recuento de lectura, un identificador de cromosoma (por ejemplo, que identifica un cromosoma al que se mapea una lectura) y un identificador de posición cromosómica (por ejemplo, que identifica una posición en un cromosoma al que se mapea una lectura). En algunas implementaciones, un formato binario comprende una matriz de 20 bytes, una matriz de 16 bytes, una matriz de 8 bytes, una matriz de 4 bytes o una matriz de 2 bytes. En algunas implementaciones, la información de lectura mapeada se almacena en una matriz en un formato de 10 bytes, formato de 9 bytes, formato de 8 bytes, formato de 7 bytes, formato de 6 bytes, formato de 5 bytes, formato de 4 bytes, formato de 3 bytes o formato de 2 bytes. Algunas veces, los datos de lectura mapeados se almacenan en una matriz de 4 bytes que comprende un formato de 5 bytes. En algunas implementaciones, un formato binario comprende un formato de 5 bytes que comprende una ordinal variante de 1 byte y una posición cromosómica de 4 bytes. En algunas implementaciones, las lecturas mapeadas se almacenan en un formato binario comprimido que es aproximadamente 100 veces, aproximadamente 90 veces, aproximadamente 80 veces, aproximadamente 70 veces, aproximadamente 60 veces, aproximadamente 55 veces, aproximadamente 50 veces, aproximadamente 45 veces, aproximadamente 40 veces o aproximadamente 30 veces más pequeño que un formato de alineación de secuencia/mapa (SAM). En algunas implementaciones, las lecturas mapeadas se almacenan en un formato binario de compresión que es aproximadamente 2 veces menor a aproximadamente 50 veces menor que (por ejemplo, aproximadamente 30, 25, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6 o aproximadamente 5 veces menor que) el formato GZip.

En algunas implementaciones, un sistema comprende un módulo de compresión (por ejemplo, 4, Figura 10A). En algunas implementaciones, la información de lectura de secuencia mapeada almacenada en un medio de almacenamiento legible por ordenador no transitorio en un formato legible por ordenador es comprimida por un módulo de compresión. Un módulo de compresión a veces convierte lecturas de secuencia mapeadas en y desde un formato adecuado. Un módulo de compresión puede aceptar lecturas de secuencia mapeadas en un primer formato (por ejemplo, 1), convertirlas en un formato comprimido (por ejemplo, un formato binario, 5) y transferir las lecturas comprimidas a otro módulo (por ejemplo, un módulo de densidad de sesgo 6) en algunas implementaciones. Un módulo de compresión a menudo proporciona lecturas de secuencia en un formato binario 5 (por ejemplo, un formato BReads). Los ejemplos no limitantes de un módulo de compresión incluyen GZIP, BGZF y BAM, similares o modificaciones de los mismos).

A continuación se proporciona un ejemplo de conversión de un número entero en una matriz de 4 bytes usando java:

```
public static final byte[]  
  
convertToByteArray(int value)  
  
{  
    return new byte[] {  
        (byte)(value >>> 24),  
        (byte)(value >>> 16),  
        (byte)(value >>> 8),  
        (byte)value};  
}
```

En algunas implementaciones, una lectura puede mapearse de manera única o no única en un 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 un genoma de referencia. En algunas implementaciones, las lecturas mapeadas de manera no única se eliminan del análisis posterior (por ejemplo, cuantificación). Puede permitirse que un determinado grado pequeño de apareamiento erróneo (0-1) tenga en cuenta los polimorfismos de un solo nucleótido que pueden existir entre el genoma de referencia y las lecturas de las muestras individuales que se mapean, en determinadas implementaciones. En algunas implementaciones, no se permite ningún grado de apareamiento erróneo para que una lectura se mapee en una secuencia de referencia.

Tal como se usa en el presente documento, la expresión “genoma de referencia” puede referirse a cualquier genoma conocido, secuenciado o caracterizado, ya sea parcial o completo, de cualquier organismo o virus que pueda usarse para referenciar secuencias identificadas de un sujeto. Un genoma de referencia a veces se refiere a un segmento de un genoma de referencia (por ejemplo, un cromosoma o parte del mismo, por ejemplo, una o más porciones de un genoma de referencia). Los genomas humanos, los conjuntos de genoma humano y/o genomas de cualquier otro organismo pueden usarse como genoma de referencia. Se pueden encontrar uno o más genomas humanos, conjuntos de genoma humano así como genomas de otros organismos en el Centro Nacional de Información Biotecnológica en www.ncbi.nlm.nih.gov. Un “genoma” se refiere a la información genética completa de un organismo o virus, expresada

en secuencias de ácido nucleico. Tal como se usa en el presente documento, una secuencia de referencia o un genoma de referencia es a menudo una secuencia genómica ensamblada o parcialmente ensamblada de un individuo o múltiples individuos. En algunas implementaciones, un genoma de referencia es una secuencia genómica ensamblada o parcialmente ensamblada de uno o más individuos humanos. En algunas implementaciones, un genoma de referencia comprende secuencias asignadas a cromosomas.

El término “secuencia de referencia” como se usa en el presente documento se refiere a una o más secuencias de polinucleótidos de una o más muestras de referencia. En algunas implementaciones, las secuencias de referencia comprenden lecturas de secuencia obtenidas de una muestra de referencia. En algunas implementaciones, las secuencias de referencia comprenden lecturas de secuencia, un ensamblaje de lecturas, una secuencia de ADN consenso (por ejemplo, un cóntigo de secuencia), densidades de lectura y/o perfiles de densidad de lectura obtenidos de una o más muestras de referencia. Un perfil de densidad de lectura obtenido a partir de una muestra de referencia a veces se denomina en el presente documento un perfil de referencia. Un perfil de densidad de lectura obtenido a partir de una muestra de prueba y/o un sujeto de prueba a veces se denomina en el presente documento perfil de prueba. En algunas implementaciones, una muestra de referencia se obtiene a partir de un sujeto de referencia sustancialmente libre de una variación genética (por ejemplo, una variación genética en cuestión). En algunas implementaciones, se obtiene una muestra de referencia a partir de un sujeto de referencia que comprende una variación genética conocida. El término “referencia” como se usa en el presente documento puede referirse a un genoma de referencia, una secuencia de referencia, muestra de referencia y/o un sujeto de referencia.

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

En determinadas implementaciones, la mapeabilidad se evalúa para una región genómica (por ejemplo, porciones, porciones genómicas). La capacidad de mapeo es la capacidad de alinear de manera inequívoca una lectura secuencia de nucleótidos con una porción de un genoma de referencia normalmente hasta un número especificado de apareamientos erróneos, incluyendo, por ejemplo, 0, 1, 2 o más apareamientos erróneos. En algunas implementaciones, la capacidad de mapeo se proporciona como una puntuación o valor donde la puntuación o valor se genera mediante un algoritmo de mapeo adecuado o software de mapeo de ordenador. Para una región genómica dada, la capacidad de mapeo esperada puede estimarse usando un enfoque de ventana deslizante de una longitud de lectura predeterminada y promediando los valores de capacidad de mapeo a nivel de lectura resultantes. Las regiones genómicas que comprenden tramos de secuencia de nucleótidos única tienen, algunas veces, un alto valor de capacidad de mapeo.

Las lecturas de secuencia pueden mapearse por un módulo de mapeo o por una máquina 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 secuencias mediante un método adecuado conocido en la técnica. En algunas implementaciones, se requiere un módulo de mapeo o una máquina que comprende un módulo de mapeo para proporcionar lecturas de secuencia mapeadas.

Recuentos

Las lecturas de secuencia que se asignan se pueden cuantificar para determinar el número de lecturas que se asignan a una región o porción de un genoma de referencia. En determinadas implementaciones, una lectura que se correlaciona con un genoma de referencia, o una región, porción o segmento de la misma, se denomina recuento. En algunas implementaciones, un recuento comprende un valor. En determinadas implementaciones, un valor de recuento se determina mediante un proceso matemático. Un recuento puede determinarse mediante un método, una operación o un procedimiento matemático adecuado. En determinadas implementaciones, un recuento se pondera, se elimina, se filtra, se normaliza, se ajusta, se promedia, se agrega o se resta o se procesa mediante una combinación de los mismos. En determinadas implementaciones, un recuento se deriva de una lectura de secuencia que se procesa o manipula mediante un método, operación o proceso matemático adecuado descrito en el presente documento o conocido en la técnica. Por ejemplo, un recuento a menudo se normaliza y/o se pondera según uno o más sesgos asociados con una lectura de secuencia. En algunas implementaciones, un recuento se normaliza y/o se pondera según el sesgo de GC asociado con una lectura de secuencia. En algunas implementaciones, un recuento se deriva de lecturas de secuencia sin procesar y/o lecturas de secuencia filtradas. En algunas implementaciones, uno o más recuentos no se manipulan matemáticamente. El término “recuento sin procesar” y “recuentos sin procesar” como se usa en el presente documento se refiere a uno o más recuentos que no se han manipulado matemáticamente.

En algunas implementaciones, se determina un recuento para algunas o todas las lecturas de secuencia mapeadas en un genoma de referencia, o una región, porción o segmento del mismo. En determinadas implementaciones, se determinan recuentos a partir de un subconjunto predefinido de lecturas de secuencia mapeadas. Los subconjuntos predefinidos (por ejemplo, subconjuntos seleccionados) de lecturas de secuencia mapeadas pueden definirse o seleccionarse con el uso de cualquier característica o variable adecuada. En algunas implementaciones, los subconjuntos predefinidos de lecturas de secuencia mapeadas pueden incluir de 1 a n lecturas de secuencia, en el que n representa un número igual a la suma de todas las lecturas de secuencia generadas a partir de una muestra de sujeto de prueba o de sujeto de referencia.

Los recuentos a menudo se derivan de lecturas de secuencia obtenidas de un sujeto (por ejemplo, un sujeto de prueba). Algunas veces, los recuentos se derivan de lecturas de secuencias obtenidas a partir de una muestra de ácido nucleico de una mujer embarazada que porta un feto. Los recuentos de lecturas de secuencia de ácido nucleico son a menudo recuentos representativos tanto de un feto como de una madre de un feto (por ejemplo, para un sujeto femenino gestante). En determinadas implementaciones, donde un sujeto es una mujer embarazada, algunos recuentos se derivan de un genoma fetal y algunos recuentos se derivan de un genoma materno.

Densidad de lectura

Los recuentos de lecturas de secuencia (por ejemplo, recuentos ponderados) a menudo se representan como una densidad de lectura. Una densidad de lectura a menudo se determina y/o genera para una o más porciones de un genoma. En determinadas implementaciones, se determina una densidad de lectura y/o se genera para uno o más cromosomas. En algunas implementaciones, una densidad de lectura comprende una medida cuantitativa de recuentos de lecturas de secuencia mapeadas en una porción de un genoma de referencia. Una densidad de lectura puede determinarse mediante un proceso adecuado. En algunas implementaciones, una densidad de lectura se determina mediante una distribución adecuada y/o una función de distribución adecuada. Los ejemplos no limitantes de una función de distribución incluyen una función de probabilidad, función de distribución de probabilidad, función de densidad de probabilidad (PDF, por sus siglas en inglés), una función de densidad de núcleo (estimación de densidad de núcleo), una función de distribución acumulativa, función de masa de probabilidad, distribución de probabilidad discreta, una distribución univariada absolutamente continua, similares, cualquier distribución adecuada, o combinaciones de las mismas. En determinadas implementaciones, una PDF comprende una función de densidad de núcleo (estimación de densidad de núcleo). Los ejemplos no limitantes de una función de densidad de núcleo que se puede usar para generar una estimación de sesgo de genoma local incluyen una función de densidad de núcleo uniforme (núcleo uniforme), una función de densidad de núcleo gaussiano (núcleo gaussiano), una función de densidad de núcleo triangular (núcleo triangular), una función de densidad de núcleo biponderal (núcleo de biponderación), una función de densidad de núcleo de tricubo (núcleo de tricubo), una función de densidad de núcleo triponderal (núcleo triponderal), funciones de núcleo de coseno (núcleo de coseno), una función de densidad de núcleo de Epanechnikov (núcleo de Epanechnikov), una función de densidad de núcleo normal (núcleo normal), similares o una combinación de los mismos. Una densidad de lectura es a menudo una estimación de densidad derivada de una función de densidad de probabilidad adecuada. Una estimación de densidad es la construcción de una estimación, basada en datos observados, de una función de densidad de probabilidad subyacente. En algunas implementaciones, una densidad de lectura comprende una estimación de densidad (por ejemplo, una estimación de densidad de probabilidad, una estimación de densidad de núcleo). Una estimación de densidad a menudo comprende una estimación de densidad de núcleo. En algunas implementaciones, una densidad de lectura es una estimación de densidad de núcleo, determinada según una función de densidad de núcleo. A menudo se genera una densidad de lectura según un proceso que comprende generar una estimación de densidad para cada una de las una o más porciones de un genoma donde cada porción comprende recuentos de lecturas de secuencia. A menudo se genera una densidad de lectura para recuentos normalizados y/o ponderados mapeados en una porción. En algunas implementaciones, cada lectura mapeada a una porción a menudo contribuye a una densidad de lectura, un valor (por ejemplo, un recuento) igual a su peso obtenido a partir de un proceso de normalización descrito en el presente documento. En algunas implementaciones se ajustan las densidades de lectura para una o más porciones. Las densidades de lectura se pueden ajustar mediante un método adecuado. Por ejemplo, las densidades de lectura para una o más porciones pueden ponderarse y/o normalizarse.

En algunas implementaciones, un sistema comprende un módulo de distribución 12. Un módulo de distribución a menudo genera y/o proporciona densidades de lectura (por ejemplo, 22, 24) para porciones (por ejemplo, porciones filtradas) de un genoma. Un módulo de distribución puede proporcionar densidades de lectura, distribuciones de densidad de lectura 14 y/o una medida asociada de incertidumbre (por ejemplo, una DAM, un cuantil) para una o más muestras de referencia, un conjunto de entrenamiento (por ejemplo, 3) y/o una muestra de prueba. Un módulo de distribución puede aceptar, recuperar y/o almacenar lecturas de secuencia (por ejemplo, 1, 3, 5) y/o recuentos (por ejemplo, recuentos normalizados 11, recuentos ponderados). Un módulo de distribución a menudo acepta (por ejemplo, entradas de usuario y parámetros de usuario para porciones), recupera y/o almacena porciones (por ejemplo, porciones no filtradas o filtradas). A veces, un módulo de distribución acepta y/o recupera porciones (por ejemplo, porciones filtradas y/o porciones seleccionadas 20) de un módulo de filtrado 18. En algunas implementaciones, un módulo de distribución comprende instrucciones para un microprocesador (por ejemplo, un algoritmo, una secuencia de comandos) en forma de código y/o código fuente (por ejemplo, una colección de secuencias de comandos).

estándares o personalizadas) y/o uno o más paquetes de software (por ejemplo, paquetes de software estadísticos) que llevan a cabo las funciones de un módulo de distribución. En algunas implementaciones, un módulo de distribución comprende código (por ejemplo, secuencia de comandos) escrito en java, S o R que utiliza un paquete adecuado (por ejemplo, un paquete S, un paquete R). En el ejemplo 2 se proporciona un ejemplo no limitante de un módulo de distribución.

En algunas implementaciones se determina un perfil de densidad de lectura. En algunas implementaciones, un perfil de densidad de lectura comprende al menos una densidad de lectura, y a menudo comprende dos o más densidades de lectura (por ejemplo, un perfil de densidad de lectura comprende a menudo múltiples densidades de lectura). En algunas implementaciones, un perfil de densidad de lectura comprende un valor cuantitativo adecuado (por ejemplo, una media, una mediana, una puntuación Z o similares). Un perfil de densidad de lectura comprende a menudo valores resultantes de una o más densidades de lectura. Un perfil de densidad de lectura a veces comprende valores resultantes de una o más manipulaciones de densidades de lectura en base a uno o más ajustes (por ejemplo, normalizaciones). En algunas implementaciones, un perfil de densidad de lectura comprende densidades de lectura no manipuladas. En algunas implementaciones, uno o más perfiles de densidad de lectura se generan a partir de diversos aspectos de un conjunto de datos que comprende densidades de lectura, o una derivación de la misma (por ejemplo, producto de una o más etapas de procesamiento de datos matemáticos y/o estadísticas conocidas en la técnica y/o descritas en el presente documento). En determinadas implementaciones, un perfil de densidad de lectura comprende densidades de lectura normalizadas. En algunas implementaciones, un perfil de densidad de lectura comprende densidades de lectura ajustadas. En determinadas implementaciones, un perfil de densidad de lectura comprende densidades de lectura sin procesar (por ejemplo, no manipuladas, no ajustadas o normalizadas), densidades de lectura normalizadas, densidades de lectura ponderadas, densidades de lectura de porciones filtradas, puntuaciones z de densidades de lectura, valores de p de densidades de lectura, valores integrales de densidades de lectura (por ejemplo, área bajo la curva), densidades de lectura promedio, medias o en mediana, componentes principales, similares, o combinaciones de los mismos. A menudo, las densidades de lectura de un perfil de densidad de lectura y/o un perfil de densidad de lectura se asocian con una medida de incertidumbre (por ejemplo, una DAM). En determinadas implementaciones, un perfil de densidad de lectura comprende una distribución de mediana de densidades de lectura. En algunas implementaciones, un perfil de densidad de lectura comprende una relación (por ejemplo, una relación ajustada, una regresión o similar) de una pluralidad de densidades de lectura. Por ejemplo, a veces un perfil de densidad de lectura comprende una relación entre densidades de lectura (por ejemplo, valor de densidades de lectura) y ubicaciones genómicas (por ejemplo, porciones, ubicaciones de porciones). En algunas implementaciones, se genera un perfil de densidad de lectura usando un procedimiento de ventana estática y, en determinadas implementaciones, se genera un perfil de densidad de lectura usando un procedimiento de ventana deslizante. El término “perfil de densidad de lectura”, tal como se usa en el presente documento, se refiere a un producto de una manipulación matemática y/o estadística de densidades de lecturas que puede facilitar la identificación de patrones y/o correlaciones en grandes cantidades de datos de lectura de secuencia.

En algunas implementaciones, un perfil de densidad de lectura a veces se imprime y/o visualiza (por ejemplo, se muestra como una representación visual, por ejemplo, una representación o un gráfico).

Un perfil de densidad de lectura comprende a menudo múltiples puntos de datos, donde cada punto de datos representa un valor cuantitativo de una o más densidades de lectura. Cualquier número adecuado de puntos de datos puede incluirse en un perfil de densidad de lectura dependiendo de la naturaleza y/o complejidad de un conjunto de datos. En determinadas implementaciones, los perfiles de densidad de lectura 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, 100.000 o más puntos de datos o 1.000.000 o más puntos de datos. En algunas implementaciones, un punto de datos es un valor cuantitativo y/o estimación de recuentos de lecturas de secuencia mapeadas o asociadas con una o más porciones. En algunas implementaciones, un punto de datos en un perfil de densidad de lectura comprende los resultados de una manipulación de datos de recuentos mapeados en una o más porciones. En determinadas implementaciones, un punto de datos es a menudo un valor cuantitativo y/o estimación de una o más densidades de lectura (por ejemplo, una densidad de lectura media). Un perfil de densidad de lectura a menudo comprende múltiples densidades de lectura asociadas y/o mapeadas en múltiples porciones de un genoma de referencia. En algunas implementaciones, un perfil de densidad de lectura comprende densidades de lectura de 2 a aproximadamente 1.000.000 porciones. En algunas implementaciones, las densidades de lectura de 2 a aproximadamente 500.000, de 2 a aproximadamente 100.000, 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 porciones determinan un perfil de densidad de lectura. En algunas implementaciones, las densidades de lectura de aproximadamente 10 a aproximadamente 50 porciones determinan un perfil de densidad de lectura.

En algunas implementaciones, un perfil de densidad de lectura corresponde a un conjunto de porciones (por ejemplo, un conjunto de porciones de un genoma de referencia, un conjunto de porciones de un cromosoma o un subconjunto de porciones de un segmento de un cromosoma). En algunas implementaciones, un perfil de densidad de lectura

comprende densidades de lectura y/o recuentos asociados con una colección (por ejemplo, un conjunto, un subconjunto) de porciones. En algunas implementaciones, se determina un perfil de densidad de lectura para las densidades de lectura de porciones que son contiguas. En algunas implementaciones, las porciones contiguas comprenden huecos que comprenden segmentos de una secuencia de referencia y/o lecturas de secuencia que no están incluidas en un perfil de densidad (por ejemplo, porciones eliminadas por un filtrado). Algunas veces, las porciones (por ejemplo, un conjunto de porciones) que son contiguas representan segmentos vecinos de un genoma o segmentos vecinos de un cromosoma o gen. Por ejemplo, dos o más porciones contiguas, cuando se alinean uniendo las porciones extremo con extremo, pueden representar un conjunto de secuencias de una secuencia de ADN más larga que cada porción. Por ejemplo, dos o más porciones contiguas pueden representar un genoma intacto, cromosoma, gen, intrón, exón o segmento del mismo. Algunas veces, se determina un perfil de densidad de lectura a partir de una colección (por ejemplo, un conjunto, un subconjunto) de porciones contiguas y/o porciones no contiguas. En algunos casos, un perfil de densidad de lectura comprende una o más porciones, que pueden ponderarse, eliminarse, filtrarse, normalizarse, ajustarse, promediarse, derivarse como media, sumarse, restarse, procesarse o transformarse mediante cualquier combinación de las mismas.

En algunas implementaciones, un perfil de densidad de lectura comprende densidades de lectura para porciones de un genoma que comprende una variación genética. En algunas implementaciones, un perfil de densidad de lectura comprende densidades de lectura para porciones de un genoma que no comprenden una variación genética (por ejemplo, porciones de un genoma que están sustancialmente libres de una variación genética). En determinadas implementaciones, un perfil de densidad de lectura comprende densidades de lectura para porciones de un genoma que comprende una variación genética y densidades de lectura para porciones de un genoma que están sustancialmente libres de una variación genética.

A menudo se determina un perfil de densidad de lectura para una muestra y/o una referencia (por ejemplo, una muestra de referencia). A veces se genera un perfil de densidad de lectura para un genoma completo, uno o más cromosomas, o para una parte o segmento de un genoma o un cromosoma. En algunas implementaciones, uno o más perfiles de densidad de lectura se determinan para un genoma o segmentos del mismo. En algunas implementaciones, un perfil de densidad de lectura es representativo de la totalidad de un conjunto de densidades de lectura de una muestra, y en determinadas implementaciones, un perfil de densidad de lectura es representativo de una parte o subconjunto de densidades de lectura de una muestra. Es decir, algunas veces, un perfil de densidad de lectura comprende o se genera a partir de densidades de lectura representativas de datos que no se han filtrado para eliminar cualquier dato y, a veces, un perfil de densidad de lectura incluye o se genera a partir de puntos de datos representativos de datos que se han filtrado para eliminar datos no deseados.

En algunas implementaciones, se determina un perfil de densidad de lectura para una referencia (por ejemplo, una muestra de referencia, un conjunto de entrenamiento). Un perfil de densidad de lectura para una referencia a veces se denomina en el presente documento un perfil de referencia. En algunas implementaciones, un perfil de referencia comprende unas densidades de lectura obtenidas de una o más referencias (por ejemplo, secuencias de referencia, muestras de referencia). En algunas implementaciones, un perfil de referencia comprende densidades de lectura determinadas para una o más (por ejemplo, un conjunto de) muestras euploides conocidas. En algunas implementaciones, un perfil de referencia comprende densidades de lectura de porciones filtradas. En algunas implementaciones, un perfil de referencia comprende densidades de lectura ajustadas según el uno o más componentes principales.

En algunas implementaciones, un sistema comprende un módulo de generación de perfil (por ejemplo, 26). Un módulo de generación de perfiles a menudo acepta, recupera y/o almacena densidades de lectura (por ejemplo, 22, 24). Un módulo de generación de perfil puede aceptar y/o recuperar densidades de lectura (por ejemplo, densidades de lectura ajustadas, ponderadas, normalizadas, medias, promediadas, en medianas y/o integradas) de otro módulo adecuado (por ejemplo, un módulo de distribución). Un módulo de generación de perfil puede aceptar y/o recuperar densidades de lectura de una fuente adecuada (por ejemplo, uno o más sujetos de referencia, un conjunto de entrenamiento, uno o más sujetos de prueba y similares). Un módulo de generación de perfil a menudo genera y/o proporciona perfiles de densidad de lectura (por ejemplo, 32, 30, 28) a otro módulo adecuado (por ejemplo, un módulo de estadísticas de PCA 33, un módulo de ponderación de porción 42, un módulo de puntuación 46) y/o a un usuario (por ejemplo, mediante trazado, graficado y/o impresión). Un ejemplo de un módulo de generación de perfil, o parte del mismo, se proporciona en el ejemplo 2.

Porciones

En algunas implementaciones, las lecturas y/o recuentos de secuencia mapeados se agrupan conjuntamente según diversos parámetros y se asignan a segmentos y/o regiones particulares de un genoma de referencia denominado en el presente documento como "porción" o "una porción". En algunas implementaciones, una porción es un cromosoma entero, un segmento de un cromosoma, un segmento de un genoma de referencia, un segmento que abarca varios cromosomas, varios segmentos de cromosomas y/o combinaciones de los mismos. En algunas implementaciones, una porción está predefinida basándose en parámetros específicos (por ejemplo, longitudes predeterminadas, espaciado predeterminado, un contenido de GC predeterminado o cualquier otro parámetro adecuado). En algunas implementaciones, una porción se define arbitrariamente basándose en la partición de un genoma (por ejemplo,

partición por tamaño, contenido de GC, regiones contiguas, regiones contiguas de un tamaño definido arbitrariamente, y similares). En algunas implementaciones, una porción se delinea basándose en uno o más parámetros que incluyen, por ejemplo, la longitud o una o varias características particulares de la secuencia. En algunas implementaciones, una porción se basa en una longitud concreta de secuencia genómica. Las porciones pueden tener aproximadamente la misma longitud o las porciones pueden tener longitudes diferentes. En algunas implementaciones, las porciones tienen aproximadamente la misma longitud. En algunas implementaciones, las porciones de longitudes diferentes se ajustan o ponderan. Una porción puede ser de cualquier longitud adecuada. En algunas implementaciones, una porción es de aproximadamente 10 kilobases (kb) a aproximadamente 100 kb, de aproximadamente 20 kb a aproximadamente 80 kb, de aproximadamente 30 kb a aproximadamente 70 kb, de aproximadamente 40 kb a aproximadamente 60 kb, y a veces de aproximadamente 50 kb. En algunas implementaciones, una porción es de aproximadamente 10 kb a aproximadamente 20 kb. Una porción no se limita a tramos contiguos de secuencia. Por tanto, las porciones pueden estar formadas por secuencias contiguas y/o no contiguas.

En algunas implementaciones, una porción comprende una ventana que comprende un número preseleccionado de bases. Una ventana puede comprender cualquier número adecuado de bases determinada por una longitud de porción. En algunas implementaciones, un genoma, o segmentos del mismo, se divide en una pluralidad de ventanas. Las regiones que abarcan las regiones de un genoma pueden o no solaparse. En algunas implementaciones, las ventanas se colocan a distancias iguales entre sí. En algunas implementaciones, las ventanas se colocan a diferentes distancias entre sí. En determinada implementación, un genoma o segmento del mismo se divide en una pluralidad de ventanas deslizantes, donde una ventana se desliza gradualmente a través de un genoma, o segmento del mismo, donde cada ventana en cada incremento representa una porción. Una ventana se puede deslizar a través de un genoma en cualquier incremento adecuado o según cualquier patrón numérico o secuencia definida atemática. En algunas implementaciones, ventanas se deslizan a través de un genoma, o un segmento del mismo, en un incremento de aproximadamente 100.000 pb o menos, aproximadamente 50.000 pb o menos, aproximadamente 25.000 pb o menos, aproximadamente 10.000 pb o menos, aproximadamente 5.000 pb o menos, aproximadamente 1.000 pb o menos, aproximadamente 500 pb o menos, o aproximadamente 100 pb o menos. Por ejemplo, una ventana puede comprender aproximadamente 100.000 pb y puede deslizarse a través de un genoma en incrementos de 50.000 pb.

En algunas implementaciones, las porciones pueden ser segmentos cromosómicos particulares de un cromosoma de interés, tales como, por ejemplo, un cromosoma donde se evalúe una variación genética (por ejemplo, una aneuploidía de los cromosomas 13, 18 y/o 21 o un cromosoma sexual). Una porción no se limita a un único cromosoma. En algunas implementaciones, una o más porciones incluyen la totalidad o parte de un cromosoma o la totalidad o parte de dos o más cromosomas. En algunas implementaciones, una o más porciones pueden abarcar uno, dos o más cromosomas enteros. Además, las porciones pueden abarcar regiones unidas o desunidas de múltiples cromosomas. Las porciones pueden ser genes, fragmentos génicos, secuencias reguladoras, intrones, exones y similares.

En algunas implementaciones, determinadas regiones de un genoma se filtran antes de dividir un genoma, o segmento del mismo, en porciones. Las regiones de un genoma pueden seleccionarse para exclusión de un proceso de partición mediante el uso de cualquier método adecuado. A menudo, las regiones que comprenden regiones similares (por ejemplo, regiones o secuencias idénticas u homólogas, por ejemplo, regiones repetitivas) se eliminan y/o filtran. Algunas veces se excluyen las regiones no mapeables. En algunas implementaciones, solo se retienen regiones únicas. Las regiones eliminadas durante la partición pueden estar dentro de un único cromosoma o abarcar múltiples cromosomas. En algunas implementaciones, un genoma dividido se recorta y optimiza para una alineación más rápida, lo que suele permitir centrarse en secuencias identificables de forma única. En algunas implementaciones, la partición de un genoma en regiones (por ejemplo, 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 un perfil de valor de p que mide la significación de ubicaciones genómicas particulares para distinguir entre grupos de sujetos normales y anómalos confirmados (por ejemplo, sujetos euploides y con trisomía, respectivamente). En algunas implementaciones, la partición de un genoma en regiones (por ejemplo, regiones que trascienden a los cromosomas) puede basarse en cualquier otro criterio, tal como, por ejemplo, la velocidad/conveniencia al alinear las lecturas, el contenido de GC (por ejemplo, alto o bajo contenido de GC), la uniformidad del contenido de GC, otras medidas del contenido de la secuencia (por ejemplo, la fracción de nucleótidos individuales, la fracción de pirimidinas o purinas, la fracción de ácidos nucleicos naturales frente a los no naturales, la fracción de nucleótidos metilados y el contenido de CpG), el estado de metilación, la temperatura de fusión de la estructura doble, la susceptibilidad a la secuenciación o la PCR, la medida de la incertidumbre asignada a porciones individuales de un genoma de referencia y/o una búsqueda dirigida para características concretas.

Un “segmento” de un genoma es en ocasiones una región que comprende uno o más cromosomas, o una parte de un cromosoma. Típicamente, un “segmento” es una parte diferente de un genoma que una porción. Un “segmento” de un genoma y/o cromosoma a veces se encuentra en una región diferente de un genoma o cromosoma que una porción, a veces no comparte un polinucleótido con una porción y, a veces, incluye un polinucleótido que se encuentra en una porción. Un segmento de un genoma o cromosoma suele contener una cantidad mayor de nucleótidos que una porción (por ejemplo, un segmento a veces incluye una o más porciones) y, a veces, un segmento de un cromosoma contiene una cantidad menor de nucleótidos que una porción (por ejemplo, un segmento a veces está dentro de una porción).

Filtración de porciones

En determinadas implementaciones, una o más porciones (por ejemplo, porciones de un genoma) se eliminan de la consideración mediante un proceso de filtrado. En determinadas implementaciones, una o más porciones se filtran (por ejemplo, se someten a un proceso de filtrado) lo que proporciona de este modo porciones filtradas. En algunas implementaciones, un proceso de filtrado elimina determinadas porciones y retiene porciones (por ejemplo, un subconjunto de porciones). Después de un proceso de filtrado, las porciones retenidas a menudo se denominan en el presente documento porciones filtradas. En algunas implementaciones, se filtran porciones de un genoma de referencia. En algunas implementaciones, las porciones de un genoma de referencia que se eliminan mediante un proceso de filtrado no se incluyen en una determinación de la presencia o ausencia de una variación genética (por ejemplo, una aneuploidía cromosómica). En algunas implementaciones, se filtran porciones de un cromosoma en un genoma de referencia. En algunas implementaciones, las porciones asociadas con las densidades de lectura (por ejemplo, donde una densidad de lectura es para una porción) se eliminan mediante un proceso de filtrado y las densidades de lectura asociadas con las porciones eliminadas no se incluyen en una determinación de la presencia o ausencia de una variación genética (por ejemplo, una aneuploidía cromosómica). En algunas implementaciones, un perfil de densidad de lectura comprende y/o consiste en densidades de lectura de porciones filtradas. Las porciones pueden seleccionarse, filtrarse y/o eliminarse de la consideración utilizando cualquier criterio y/o método adecuado conocido en la técnica o descrito en el presente documento. Los ejemplos no limitantes de criterios utilizados para las porciones de filtrado incluyen datos redundantes (por ejemplo, lecturas mapeadas redundantes o superpuestas), datos no informativos (por ejemplo, porciones de un genoma de referencia con cero recuentos mapeados), porciones de un genoma de referencia con secuencias sobrerrepresentadas o subrepresentadas, contenido de GC, datos con ruido, capacidad de mapeo, recuentos, variabilidad de recuento, densidad de lectura, variabilidad de la densidad de lectura, una medida de incertidumbre, una medida de repetibilidad, similares o combinaciones de los anteriores. Las porciones a veces se filtran según una distribución de recuentos y/o una distribución de densidades de lectura. En algunas implementaciones, las porciones se filtran según una distribución de recuentos y/o densidades de lectura donde los recuentos y/o densidades de lectura se obtienen a partir de una o más muestras de referencia. Una o más muestras de referencia a veces se denominan en el presente documento un conjunto de entrenamiento. En algunas implementaciones, las porciones se filtran según una distribución de recuentos y/o densidades de lectura donde los recuentos y/o densidades de lectura se obtienen de una o más muestras de prueba. En algunas implementaciones, las porciones se filtran según una medida de incertidumbre para una distribución de densidad de lectura. En determinadas implementaciones, las porciones que demuestran una gran desviación en las densidades de lectura se eliminan mediante un proceso de filtrado. Por ejemplo, se puede determinar una distribución de densidades de lectura (por ejemplo, una distribución de densidades de lectura promedio o en mediana, por ejemplo, la Figura 5A), donde cada densidad de lectura en los mapas de distribución a la misma porción. Una medida de incertidumbre (por ejemplo, una DAM) puede determinarse comparando una distribución de densidades de lectura para múltiples muestras donde cada porción de un genoma está asociada con la medida de incertidumbre. Según el ejemplo anterior, las porciones pueden filtrarse según una medida de incertidumbre (por ejemplo, una desviación estándar (SD), una DAM) asociada con cada porción y un umbral predeterminado. La Figura 5B muestra una distribución de valores de DAM para porciones, determinadas según distribuciones de densidad de lectura para múltiples muestras. Un umbral predeterminado se indica mediante las líneas verticales discontinuas que encierran un rango de valores de DAM aceptables. En el ejemplo de la Figura 5B, se retienen porciones que comprenden valores DAM dentro del rango aceptable y se eliminan las porciones que comprenden valores DAM fuera del rango aceptable de la consideración mediante un proceso de filtrado. En algunas implementaciones, según el ejemplo anterior, las porciones que comprenden valores de densidades de lectura (por ejemplo, densidades de lectura en mediana, medias o promedio) fuera de una medida predeterminada de incertidumbre a menudo se eliminan de la consideración mediante un proceso de filtrado. En algunas implementaciones, las porciones que comprenden valores de densidades de lectura (por ejemplo, densidades de lectura en mediana, medias o promedio) fuera de un rango entre cuartiles de una distribución se eliminan de la consideración mediante un proceso de filtrado. En algunas implementaciones, las porciones que comprenden densidades de lectura de valores fuera por más de 2 veces, 3 veces, 4 veces o 5 veces un rango entre cuartiles de una distribución se eliminan de la consideración mediante un proceso de filtrado. En algunas realizaciones, las porciones que comprenden valores de densidades de lectura fuera por más de 2 sigma, 3 sigma, 4 sigma, 5 sigma, 6 sigma, 7 sigma u 8 sigma (por ejemplo, donde sigma es un rango definido por una desviación estándar) se eliminan de la consideración mediante un proceso de filtrado.

En algunas implementaciones, un sistema comprende un módulo de filtrado 18. Un módulo de filtrado a menudo acepta, recupera y/o almacena porciones (por ejemplo, porciones de tamaños predeterminados y/o superposición, ubicaciones de porciones dentro de un genoma de referencia) y densidades de lectura asociadas con porciones, a menudo de otro módulo adecuado (por ejemplo, un módulo de distribución 12). En algunas implementaciones, las porciones seleccionadas (por ejemplo, 20, por ejemplo, porciones filtradas) son proporcionadas por un módulo de filtrado. En algunas implementaciones, se requiere un módulo de filtrado para proporcionar porciones filtradas y/o para eliminar porciones de la consideración. En determinadas implementaciones, un módulo de filtrado elimina las densidades de lectura de la consideración donde las densidades de lectura están asociadas con las porciones eliminadas. Un módulo de filtrado a menudo proporciona porciones seleccionadas (por ejemplo, porciones filtradas) a otro módulo adecuado (por ejemplo, un módulo de distribución 21). En el ejemplo 3 se proporciona un ejemplo no limitante de un módulo de filtrado.

Estimaciones de sesgo

Las tecnologías de secuenciación pueden ser vulnerables a múltiples fuentes de sesgo. A veces, el sesgo de secuenciación es un sesgo local (por ejemplo, un sesgo del genoma local). El sesgo local a menudo se manifiesta a nivel de una lectura de secuencia. Un sesgo del genoma local puede ser cualquier sesgo local adecuado. Los ejemplos no limitantes de un sesgo local incluyen sesgo de secuencia (por ejemplo, sesgo de GC, sesgo de AT y similares), sesgo correlacionado con la sensibilidad de ADNasa I, entropía, sesgo de secuencia repetitiva, sesgo de estructura de cromatina, sesgo de tasa de error de polimerasa, sesgo de palíndromo, sesgo de repetición invertido, sesgo relacionado con PCR, similares o combinaciones de los mismos. En algunas implementaciones, la fuente de un sesgo local no se determina o se conoce.

En algunas implementaciones se determina una estimación local del sesgo del genoma. Una estimación local del sesgo del genoma a veces se denomina en el presente documento una estimación de sesgo local del genoma. Se puede determinar una estimación local del sesgo del genoma para un genoma de referencia, un segmento o una porción del mismo. En determinadas implementaciones, se determina una estimación local del sesgo del genoma para uno o más cromosomas en un genoma de referencia. En algunas implementaciones, se determina una estimación local del sesgo del genoma para una o más lecturas de secuencia (por ejemplo, algunas o todas las lecturas de secuencia de una muestra). A menudo se determina una estimación de sesgo de genoma local para una lectura de secuencia según una estimación de sesgo de genoma local para una ubicación y/o posición correspondiente de una referencia (por ejemplo, un genoma de referencia, un cromosoma en un genoma de referencia). En algunas implementaciones, una estimación local del sesgo del genoma comprende una medida cuantitativa de sesgo de una secuencia (por ejemplo, una lectura de secuencia, una secuencia de un genoma de referencia). Una estimación local del sesgo del genoma puede determinarse mediante un método adecuado o proceso matemático. En algunas implementaciones, una estimación local del sesgo del genoma se determina mediante una distribución adecuada y/o una función de distribución adecuada (por ejemplo, una PDF). En algunas implementaciones, una estimación local del sesgo del genoma comprende una representación cuantitativa de una PDF. En algunas implementaciones, una estimación de sesgo de genoma local (por ejemplo, una estimación de densidad de probabilidad (PDE, por sus siglas en inglés), una estimación de densidad de núcleo) se determina por una función de densidad de probabilidad (por ejemplo, una función de PDF, por ejemplo, una función de densidad de núcleo) de un contenido de sesgo local. En algunas implementaciones, una estimación de densidad comprende una estimación de densidad de núcleo. Una estimación local del sesgo del genoma a veces se expresa como un promedio, media o mediana de una distribución. A veces, una estimación local del sesgo del genoma se expresa como una suma o una integral (por ejemplo, un área bajo una curva (AUC) de una distribución adecuada.

Una PDF (por ejemplo, una función de densidad de núcleo, por ejemplo, una función de densidad de núcleo de Epanechnikov) comprende a menudo una variable de ancho de banda (por ejemplo, un ancho de banda). Una variable de ancho de banda a menudo define el tamaño y/o la longitud de una ventana a partir de la cual se deriva una estimación de densidad de probabilidad (PDE) cuando se usa una PDF. Una ventana a partir de la cual se deriva una PDE a menudo comprende una longitud definida de polinucleótidos. En algunas implementaciones, una ventana a partir de la cual se deriva una PDE es una porción. Una porción (por ejemplo, un tamaño de porción, una longitud de porción) a menudo se determina según una variable de ancho de banda. Una variable de ancho de banda determina la longitud o el tamaño de la ventana usada para determinar una estimación de sesgo de genoma local, una longitud de un segmento de polinucleótido (por ejemplo, un segmento contiguo de bases de nucleótidos) a partir de la cual se determina una estimación local de sesgo de genoma. Una PDE (por ejemplo, densidad de lectura, estimación de sesgo del genoma local (por ejemplo, densidad de GC)) puede determinarse usando cualquier ancho de banda adecuado, cuyos ejemplos no limitantes incluyen un ancho de banda de aproximadamente 5 bases a aproximadamente 100.000 bases, aproximadamente 5 bases a aproximadamente 50.000 bases, aproximadamente 5 bases a aproximadamente 25.000 bases, aproximadamente 5 bases a aproximadamente 10.000 bases, aproximadamente 5 bases a aproximadamente 5000 bases, aproximadamente 5 bases a aproximadamente 2500 bases, aproximadamente 5 bases a aproximadamente 1000 bases, aproximadamente 5 bases a aproximadamente 500 bases, aproximadamente 5 bases a aproximadamente 250 bases, aproximadamente 20 bases a aproximadamente 250 bases, o similares. En algunas implementaciones, se determina una estimación local del sesgo del genoma (por ejemplo, una densidad de GC) usando un ancho de banda de aproximadamente 400 bases o menos, aproximadamente 350 bases o menos, aproximadamente 300 bases o menos, aproximadamente 250 bases o menos, aproximadamente 225 bases o menos, aproximadamente 200 bases o menos, aproximadamente 175 bases o menos, aproximadamente 150 bases o menos, aproximadamente 125 bases o menos, aproximadamente 100 bases o menos, aproximadamente 75 bases o menos, aproximadamente 50 bases o menos o aproximadamente 25 bases o menos. En determinadas implementaciones, se determina una estimación de sesgo de genoma local (por ejemplo, una densidad de GC) usando un ancho de banda determinado según una longitud de lectura promedio, media, mediana o máxima de lecturas de secuencia obtenidas para un sujeto y/o muestra dados. Algunas veces, se determina una estimación de sesgo de genoma local (por ejemplo, una densidad de GC) usando un ancho de banda aproximadamente igual a una longitud de lectura promedio, media, mediana o máxima de lecturas de secuencia obtenidas para un sujeto y/o muestra dados. En algunas implementaciones, se determina una estimación local del sesgo del genoma (por ejemplo, una densidad de GC) usando un ancho de banda de aproximadamente 250, 240, 230, 220, 210, 200, 190, 180, 160, 150, 140, 130, 120, 110, 100, 90, 80, 70, 60, 50, 40, 30, 20 o aproximadamente 10 bases.

Puede determinarse una estimación de sesgo de genoma local a una única resolución de base, aunque las estimaciones locales de sesgo del genoma (por ejemplo, el contenido local de GC) pueden determinarse a una resolución más baja. En algunas implementaciones, se determina una estimación local del sesgo del genoma para un contenido de sesgo local. A menudo se determina una estimación local del sesgo del genoma (por ejemplo, según lo determinado mediante una PDF) usando una ventana. En algunas implementaciones, una estimación de sesgo local del genoma comprende el uso de una ventana que comprende un número de bases preseleccionado. Algunas veces, una ventana comprende un segmento de bases contiguas. Algunas veces, una ventana comprende una o más porciones de bases no contiguas. A veces, una ventana comprende una o más porciones (por ejemplo, porciones de un genoma). Un tamaño o longitud de la ventana a menudo se determina por un ancho de banda y según una PDF. En algunas implementaciones, una ventana es aproximadamente 10 o más, 8 o más, 7 o más, 6 o más, 5 o más, 4 o más, 3 o más, o aproximadamente 2 o más veces la longitud de un ancho de banda. Una ventana es a veces dos veces la longitud de un ancho de banda seleccionada cuando se usa una PDF (por ejemplo, una función de densidad de núcleo) para determinar una estimación de densidad. Una ventana puede comprender cualquier número adecuado de bases. En algunas implementaciones, una ventana comprende aproximadamente 5 bases a aproximadamente 100.000 bases, aproximadamente 5 bases a aproximadamente 50.000 bases, aproximadamente 5 bases a aproximadamente 25.000 bases, aproximadamente 5 bases a aproximadamente 10.000 bases, aproximadamente 5 bases a aproximadamente 5000 bases, aproximadamente 5 bases a aproximadamente 2500 bases, aproximadamente 5 bases a aproximadamente 1000 bases, aproximadamente 5 bases a aproximadamente 500 bases, aproximadamente 5 bases a aproximadamente 250 bases, o aproximadamente 20 bases a aproximadamente 250 bases. En algunas implementaciones, un genoma, o segmentos del mismo, se divide en una pluralidad de ventanas. Las regiones que abarcan las regiones de un genoma pueden o no solaparse. En algunas implementaciones, las ventanas se colocan a distancias iguales entre sí. En algunas implementaciones, las ventanas se colocan a diferentes distancias entre sí. En determinada implementación un genoma, o segmento del mismo, se divide en una pluralidad de ventanas deslizantes, donde una ventana se desliza gradualmente a través de un genoma, o segmento del mismo, donde cada ventana en cada incremento comprende una estimación de sesgo del genoma local (por ejemplo, una densidad de GC local). Una ventana se puede deslizar a través de un genoma en cualquier incremento adecuado, según cualquier patrón numérico o según cualquier secuencia definida atemática. En algunas implementaciones, para una determinación de estimación de sesgo de genoma local, una ventana se desliza a través de un genoma, o un segmento del mismo, en un incremento de aproximadamente 10.000 pb o más aproximadamente 5000 pb o más, aproximadamente 2500 pb o más, aproximadamente 1000 pb o más, aproximadamente 750 pb o más, aproximadamente 500 pb o más, aproximadamente 400 bases o más, aproximadamente 250 pb o más, aproximadamente 100 pb o más, aproximadamente 50 pb o más, o aproximadamente 25 pb o más. En algunas implementaciones, para una determinación de estimación de sesgo de genoma local, una ventana se desliza a través de un genoma, o un segmento del mismo, en un incremento de aproximadamente 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2 o aproximadamente 1 pb. Por ejemplo, para una determinación de estimación de sesgo de genoma local, una ventana puede comprender aproximadamente 400 pb (por ejemplo, un ancho de banda de 200 pb) y puede deslizarse a través de un genoma en incrementos de 1 pb. En algunas implementaciones, se determina una estimación de sesgo de genoma local para cada base en un genoma, o segmento del mismo, usando una función de densidad de núcleo y un ancho de banda de aproximadamente 200 pb.

En algunas implementaciones, una estimación local del sesgo del genoma es un contenido local de GC y/o una representación del contenido local de GC. El término "local" como se usa en el presente documento (por ejemplo, como se usa para describir un sesgo local, estimación de sesgo local, contenido de sesgo local, sesgo del genoma local, contenido de GC local y similares) se refiere a un segmento de polinucleótido de 10.000 pb o menos. En algunas implementaciones, el término "local" se refiere a un segmento de polinucleótido de 5000 pb o menos, 4000 pb o menos, 3000 pb o menos, 2000 pb o menos, 1000 pb o menos, 500 pb o menos, 250 pb o menos, 200 pb o menos, 175 pb o menos, 150 pb o menos, 100 pb o menos, 75 pb o menos, o 50 pb o menos. Un contenido de GC local es a menudo una representación (por ejemplo, una representación matemática, una representación cuantitativa) del contenido de GC para un segmento local de un genoma, un conjunto de lectura de secuencia, un conjunto de lectura de secuencia (por ejemplo, un cóntigo, un perfil y similares). Por ejemplo, un contenido de GC local puede ser una estimación local de sesgo de GC o una densidad de GC.

A menudo se determinan una o más densidades de GC para polinucleótidos de una referencia o muestra (por ejemplo, una muestra de prueba). En algunas implementaciones, una densidad de GC es una representación (por ejemplo, una representación matemática, una representación cuantitativa) del contenido local de GC (por ejemplo, para un segmento de polinucleótido de 5000 pb o menos). En algunas implementaciones, una densidad de GC es una estimación local del sesgo del genoma. Una densidad de GC puede determinarse usando un proceso adecuado descrito en el presente documento y/o conocido en la técnica. Se puede determinar una densidad de GC usando una PDF adecuada (por ejemplo, una función de densidad de núcleo (por ejemplo, una función de densidad de núcleo Epanechnikov, por ejemplo, véase la Figura 1)). En algunas implementaciones, una densidad de GC es una PDE (por ejemplo, una estimación de densidad de núcleo). En determinadas implementaciones, una densidad de GC se define por la presencia o ausencia de uno o más nucleótidos de guanina (G) y/o citosina (C). Inversamente, en algunas implementaciones, una densidad de GC puede definirse por la presencia o ausencia de uno o más nucleótidos adenina (A) y/o timidina (T). Las densidades de GC para el contenido local de GC, en algunas implementaciones, se normalizan según las densidades de GC determinadas para un genoma completo, o segmento del mismo (por ejemplo, autosomas, conjunto de cromosomas, un solo cromosoma, un gen, por ejemplo, véase la Figura 2). Se pueden

determinar una o más densidades de GC para polinucleótidos de una muestra (por ejemplo, una muestra de prueba) o una muestra de referencia. A menudo se determina una densidad de GC para un genoma de referencia. En algunas implementaciones, se determina una densidad de GC para una lectura de secuencia según un genoma de referencia. Una densidad de GC de una lectura se determina a menudo según una densidad de GC determinada para una ubicación y/o posición correspondiente de un genoma de referencia al que se mapea una lectura. En algunas implementaciones, se asigna una densidad de GC determinada para una ubicación en un genoma de referencia y/o se proporciona para una lectura, donde la lectura, o un segmento de la misma, se mapea a la misma ubicación en el genoma de referencia. Puede usarse cualquier método adecuado para determinar una ubicación de una lectura mapeada en un genoma de referencia con el fin de generar una densidad de GC para una lectura. En algunas implementaciones, una posición mediana de una lectura mapeada determina una ubicación en un genoma de referencia a partir del cual se determina una densidad de GC para la lectura. Por ejemplo, cuando la posición media de una lectura se mapea al cromosoma 12 en el número de base x de un genoma de referencia, la densidad de GC de la lectura a menudo se proporciona como la densidad de GC determinada por una estimación de densidad de núcleo para una posición ubicada en el cromosoma 12 en o cerca del número de base x del genoma de referencia. En algunas implementaciones, se determina una densidad de GC para algunas o todas las posiciones de base de una lectura según un genoma de referencia. A veces, una densidad de GC de una lectura comprende un promedio, suma, mediana o integral de dos o más densidades de GC determinadas para una pluralidad de posiciones de base en un genoma de referencia.

En algunas implementaciones, se cuantifica una estimación de sesgo local del genoma (por ejemplo, una densidad de GC) y/o se proporciona un valor. Una estimación de sesgo local del genoma (por ejemplo, una densidad de GC) a veces se expresa como un promedio, media y/o mediana. Una estimación de sesgo local del genoma (por ejemplo, una densidad de GC) a veces se expresa como una altura máxima de una PDE. A veces, una estimación de sesgo local del genoma (por ejemplo, una densidad de GC) se expresa como una suma o una integral (por ejemplo, un área bajo una curva (AUC)) de una PDE adecuada. En algunas implementaciones, una densidad de GC comprende un peso de núcleo. En determinadas implementaciones, una densidad de GC de una lectura comprende un valor aproximadamente igual a una altura de pico promedio, media, suma, mediana, altura máxima o integral de un peso de núcleo.

Frecuencias de sesgo

Las frecuencias de sesgo a veces se determinan según una o más estimaciones de sesgo del genoma local (por ejemplo, densidades de GC). Una frecuencia de sesgo es a veces un recuento o suma del número de apariciones de una estimación de sesgo de genoma local para una muestra, referencia (por ejemplo, un genoma de referencia, una secuencia de referencia, un cromosoma en un genoma de referencia) o parte del mismo. Una frecuencia de sesgo es a veces un recuento o suma del número de apariciones de una estimación de sesgo de genoma local (por ejemplo, cada estimación de sesgo de genoma local) para una muestra, referencia o parte de la misma. En algunas implementaciones, una frecuencia de sesgo es una frecuencia de densidad de GC. Una frecuencia de densidad de GC se determina a menudo según una o más densidades de GC. Por ejemplo, una frecuencia de densidad de GC puede representar el número de veces que una densidad de GC del valor x se representa en un genoma completo, o un segmento del mismo. Una frecuencia de sesgo es a menudo una distribución de estimaciones locales de sesgo del genoma, donde el número de apariciones de cada estimación de sesgo local del genoma se representa como una frecuencia de sesgo (por ejemplo, véase la Figura 3). Las frecuencias de sesgo a veces se manipulan matemáticamente y/o normalizan. Las frecuencias de sesgo pueden manipularse matemáticamente y/o normalizarse mediante un método adecuado. En algunas implementaciones, las frecuencias de sesgo se normalizan según una representación (por ejemplo, una fracción, un porcentaje) de cada estimación de sesgo local del genoma para una muestra, referencia o parte de la misma (por ejemplo, autosomas, un subconjunto de cromosomas, un solo cromosoma o lecturas del mismo). Las frecuencias de sesgo se pueden determinar para algunas o todas las estimaciones locales de sesgo del genoma de una muestra o referencia. En algunas implementaciones, las frecuencias de sesgo pueden determinarse para estimaciones de sesgo del genoma local para algunas o todas las lecturas de secuencia de una muestra de prueba.

En algunas implementaciones, un sistema comprende un módulo de densidad de sesgo 6. Un módulo de densidad de sesgo puede aceptar, recuperar y/o almacenar lecturas de secuencia mapeadas 5 y secuencias de referencia 2 en cualquier formato adecuado y generar estimaciones de sesgo del genoma local, distribuciones de sesgo del genoma local, frecuencias de sesgo, densidades de GC, distribuciones de densidad de GC y/o frecuencias de densidad de GC (representadas colectivamente por la caja 7). En algunas implementaciones, un módulo de densidad de sesgo transfiere datos y/o información (por ejemplo, 7) a otro módulo adecuado (por ejemplo, un módulo 8 de relación).

Relaciones

En algunas implementaciones, se generan una o más relaciones entre estimaciones de sesgo del genoma local y frecuencias de sesgo. El término "relación" como se usa en el presente documento se refiere a una relación matemática y/o gráfica entre dos o más variables o valores. Se puede generar una relación mediante un proceso matemático y/o gráfico adecuado. Los ejemplos no limitantes de una relación incluyen una representación matemática y/o gráfica de una función, una correlación, una distribución, una ecuación lineal o no lineal, una línea, una regresión,

una regresión ajustada, similares o una combinación de las mismas. A veces, una relación comprende una relación ajustada. En algunas implementaciones, una relación ajustada comprende una regresión ajustada. Algunas veces, una relación comprende dos o más variables o valores que se ponderan. En algunas implementaciones, una relación comprende una regresión ajustada donde una o más variables o valores de la relación se ponderan. Algunas veces, una regresión se ajusta de manera ponderada. Algunas veces, una regresión se ajusta sin ponderación. En determinadas implementaciones, generar una relación comprende realizar una representación o gráfico.

En algunas implementaciones, se determina una relación adecuada entre estimaciones de sesgo del genoma local y frecuencias de sesgo. En algunas implementaciones, generar una relación entre (i) estimaciones de sesgo del genoma local y (ii) frecuencias de sesgo para una muestra proporciona una relación de sesgo de muestra. En algunas implementaciones, generar una relación entre (i) estimaciones de sesgo del genoma local y (ii) frecuencias de sesgo para una referencia proporciona una relación de sesgo de referencia. En determinadas implementaciones, una relación se genera entre las densidades de GC y las frecuencias de densidad de GC. En algunas implementaciones, generar una relación entre (i) densidades de GC y (ii) frecuencias de densidad de GC para una muestra proporciona una relación de densidad de GC de muestra. En algunas implementaciones, generar una relación entre (i) densidades de GC y (ii) frecuencias de densidad de GC para una referencia proporciona una relación de densidad de GC de referencia. En algunas implementaciones, donde las estimaciones locales de sesgo del genoma son densidades de GC, una relación de sesgo de muestra es una relación de densidad de GC de muestra y una relación de sesgo de referencia es una relación de densidad de GC de referencia. Las densidades de GC de una relación de densidad de GC de referencia y/o una relación de densidad de GC de muestra son a menudo representaciones (por ejemplo, representación matemática o cuantitativa) del contenido local de GC. En algunas implementaciones, una relación entre estimaciones de sesgo del genoma local y frecuencias de sesgo comprende una distribución. En algunas implementaciones, una relación entre estimaciones de sesgo del genoma local y frecuencias de sesgo comprende una regresión lineal o no lineal ajustada (por ejemplo, una regresión polinómica). En determinadas implementaciones, una relación entre estimaciones de sesgo del genoma local y frecuencias de sesgo comprende una relación ponderada donde estimaciones de sesgo del genoma local y/o frecuencias de sesgo se ponderan mediante un proceso adecuado. En algunas implementaciones, se puede obtener una relación ajustada ponderada (por ejemplo, un ajuste ponderado) mediante un proceso que comprende una regresión cuantil, distribuciones parametrizadas o una distribución empírica con interpolación. En determinadas implementaciones, una relación entre estimaciones de sesgo del genoma local y frecuencias de sesgo para una muestra de prueba, una referencia o parte de la misma, comprende una regresión polinómica donde se ponderan estimaciones de sesgo del genoma local. En algunas implementaciones, un modelo ajustado ponderado comprende valores de ponderación de una distribución. Los valores de una distribución pueden ponderarse mediante un proceso adecuado. En algunas implementaciones, los valores ubicados cerca de las colas de una distribución se proporcionan menos peso que los valores más cercanos a la mediana de la distribución. Por ejemplo, para una distribución entre estimaciones de sesgo local del genoma (por ejemplo, densidades de GC) y frecuencias de sesgo (por ejemplo, frecuencias de densidad de GC), un peso se determina según la frecuencia de sesgo para una estimación de sesgo de genoma local dada, donde las estimaciones de sesgo del genoma local que comprenden frecuencias de sesgo más cercanas a la media de una distribución se proporcionan en mayor peso que las estimaciones locales de sesgo del genoma que comprenden frecuencias de sesgo más allá de la media.

En algunas implementaciones, un sistema comprende un módulo de relación 8. Un módulo de relación puede generar relaciones, así como funciones, coeficientes, constantes y variables que definen una relación. Un módulo de relación puede aceptar, almacenar y/o recuperar datos y/o información (por ejemplo, 7) de un módulo adecuado (por ejemplo, un módulo de densidad de sesgo 6) y generar una relación. Un módulo de relación a menudo genera y compara distribuciones de estimaciones locales de sesgo del genoma. Un módulo de relación puede comparar conjuntos de datos y, a veces, generar regresiones y/o relaciones ajustadas. En algunas implementaciones, un módulo de relación compara una o más distribuciones (por ejemplo, distribuciones de estimaciones de sesgo del genoma local de muestras y/o referencias) y proporciona factores de ponderación y/o asignaciones de ponderación 9 para recuentos de lecturas de secuencia en otro módulo adecuado (por ejemplo, un módulo de corrección de sesgo). A veces, un módulo de relación proporciona recuentos normalizados de lecturas de secuencia directamente a un módulo 21 de distribución donde los recuentos se normalizan según una relación y/o una comparación.

Generar una comparación y uso de la misma

En algunas implementaciones, un proceso para reducir el sesgo local en lecturas de secuencia comprende normalizar recuentos de lecturas de secuencia. Los recuentos de lecturas de secuencia a menudo se normalizan según una comparación de una muestra de prueba con una referencia. Por ejemplo, a veces, los recuentos de lecturas de secuencia se normalizan al comparar las estimaciones de sesgo del genoma local de las lecturas de secuencia de una muestra de prueba con estimaciones de sesgo del genoma local de una referencia (por ejemplo, un genoma de referencia, o parte del mismo). En algunas implementaciones, los recuentos de lecturas de secuencia se normalizan comparando las frecuencias de sesgo de las estimaciones de sesgo del genoma local de una muestra de prueba para desviar las frecuencias de estimaciones locales de sesgo del genoma de una referencia. En algunas implementaciones, los recuentos de lecturas de secuencia se normalizan al comparar una relación de sesgo de muestra y una relación de sesgo de referencia, lo que genera de este modo una comparación.

Los recuentos de lecturas de secuencia a menudo se normalizan según una comparación de dos o más relaciones. En determinadas implementaciones, se comparan dos o más relaciones, lo que proporciona de este modo una comparación que se usa para reducir el sesgo local en lecturas de secuencia (por ejemplo, recuentos de normalización). Se pueden comparar dos o más relaciones mediante un método adecuado. En algunas implementaciones, una comparación comprende añadir, restar, multiplicar y/o dividir una primera relación de una segunda relación. En determinadas implementaciones que comparan dos o más relaciones comprende el uso de una regresión lineal adecuada y/o una regresión no lineal. En determinadas implementaciones que comparan dos o más relaciones comprende una regresión polinomial adecuada (por ejemplo, una regresión polinómica del 3.º orden). En algunas implementaciones, una comparación comprende añadir, restar, multiplicar y/o dividir una primera regresión de una segunda regresión. En algunas implementaciones, dos o más relaciones se comparan mediante un proceso que comprende un marco inferencial de múltiples regresiones. En algunas implementaciones, dos o más relaciones se comparan mediante un proceso que comprende un análisis multivariado adecuado. En algunas implementaciones, dos o más relaciones se comparan mediante un proceso que comprende una función base (por ejemplo, una función de mezcla, por ejemplo, bases polinomiales, bases de Fourier o similares), esplines, una función base radial y/u ondículas.

En determinadas implementaciones, una distribución de estimaciones locales de sesgo del genoma que comprende frecuencias de sesgo para una muestra de prueba y una referencia se compara mediante un proceso que comprende una regresión polinómica donde se ponderan estimaciones locales del genoma. En algunas implementaciones se genera una regresión polinómica entre (i) relaciones, cada una de las cuales comprenden frecuencias de sesgo de estimaciones de sesgo local del genoma de una frecuencia de referencia y sesgo de estimaciones de sesgo local del genoma de una muestra y (ii) estimaciones locales de sesgo del genoma. En algunas implementaciones se genera una regresión polinómica entre (i) una relación de frecuencias de sesgo de estimaciones de sesgo del genoma local de una referencia a frecuencias de sesgo de estimaciones de sesgo del genoma local de una muestra y (ii) estimaciones de sesgo del genoma local. En algunas implementaciones, una comparación de una distribución de estimaciones de sesgo local del genoma para lecturas de una muestra de prueba y una referencia comprende determinar una relación logarítmica (por ejemplo, una relación $\log_2$) de frecuencias de sesgo de estimaciones de sesgo local del genoma para la referencia y la muestra. En algunas implementaciones, una comparación de una distribución de estimaciones locales de sesgo del genoma comprende dividir una relación logarítmica (por ejemplo, una relación $\log_2$) de frecuencias de sesgo de estimaciones de sesgo del genoma local para la referencia por una relación logarítmica (por ejemplo, una relación $\log_2$) de frecuencias de sesgo de estimaciones de sesgo del genoma local para la muestra (por ejemplo, véase el ejemplo 1 y la Figura 4).

Los recuentos de normalización según una comparación típicamente ajustan algunos recuentos y no otros. Los recuentos de normalización a veces ajustan todos los recuentos y, a veces, no ajustan ningún recuento de lecturas de secuencia. Un recuento para una lectura de secuencia a veces se normaliza mediante un proceso que comprende determinar un factor de ponderación y, a veces, el proceso no incluye generar y utilizar directamente un factor de ponderación. Normalizar recuentos según una comparación a veces comprende determinar un factor de ponderación para cada recuento de una lectura de secuencia. Un factor de ponderación a menudo es específico para una lectura de secuencia y se aplica a un recuento de una secuencia de lectura específica. A menudo se determina un factor de ponderación según una comparación de dos o más relaciones de sesgo (por ejemplo, una relación de sesgo de muestra en comparación con una relación de sesgo de referencia). Un recuento normalizado se determina a menudo ajustando un valor de recuento según un factor de ponderación. Ajustar un recuento según un factor de ponderación a veces incluye añadir, restar, multiplicar y/o dividir un recuento para una lectura de secuencia por un factor de ponderación. Un factor de ponderación y/o un recuento normalizado a veces se determina a partir de una regresión (por ejemplo, una línea de regresión). A veces se obtiene un recuento normalizado directamente desde una línea de regresión (por ejemplo, una línea de regresión ajustada) resultante de una comparación entre las frecuencias de sesgo de las estimaciones de sesgo del genoma local de una referencia (por ejemplo, un genoma de referencia, un cromosoma en un genoma de referencia) y una muestra de prueba. En algunas implementaciones, cada recuento de una lectura de una muestra se proporciona un valor de recuento normalizado según una comparación de (i) frecuencias de sesgo de una estimación de sesgo de genoma local de lecturas en comparación con (ii) frecuencias de sesgo de una estimación local de sesgo de genoma de una referencia. En determinadas implementaciones, los recuentos de lecturas de secuencia obtenidas para una muestra se normalizan y se reduce el sesgo en las lecturas de secuencia.

A veces, un sistema comprende un módulo de corrección de sesgo. En algunas implementaciones, las funciones de un módulo de corrección de sesgo se realizan mediante un módulo de modelado de relación. Un módulo de corrección de sesgo puede aceptar, recuperar y/o almacenar lecturas de secuencia mapeadas y factores de ponderación (por ejemplo, 9) de un módulo adecuado (por ejemplo, un módulo de relación 8, un módulo de compresión 4). En algunas implementaciones, un módulo de corrección de sesgo proporciona un recuento a lecturas mapeadas. En algunas implementaciones, un módulo de corrección de sesgo aplica asignaciones de ponderación y/o factores de corrección de sesgo a recuentos de lecturas de secuencia, lo que proporciona de este modo recuentos normalizados y/o ajustados. Un módulo de corrección de sesgo a menudo proporciona recuentos normalizados a otro módulo adecuado (por ejemplo, un módulo de distribución 21).

En determinadas implementaciones, la normalización de recuentos comprende factorizar una o más características además de la densidad de GC, y normalizar recuentos de las lecturas de secuencia. En determinadas implementaciones, la normalización de recuentos comprende factorizar una o más estimaciones de sesgo local diferentes, y normalizar recuentos de las lecturas de secuencia. En determinadas implementaciones, los recuentos de lecturas de secuencia se ponderan según una ponderación determinada según una o más características (por ejemplo, uno o más sesgos). En algunas implementaciones, los recuentos se normalizan según uno o más pesos combinados. A veces, la factorización de una o más características y/o recuentos de normalización según uno o más pesos combinados es mediante un proceso que comprende el uso de un modelo multivariado. Se puede usar cualquier modelo multivariado adecuado para normalizar los recuentos. Los ejemplos no limitantes de un modelo multivariado incluyen una regresión lineal multivariante, regresión cuantil multivariante, una interpolación multivariada de datos empíricos, un modelo multivariado no lineal, similar o una combinación de los mismos.

En algunas implementaciones, un sistema comprende un módulo de corrección multivariante 13. Un módulo de corrección multivariante puede realizar funciones de un módulo de densidad de sesgo 6, un módulo de relación 8 y/o un módulo de corrección de sesgo 10 varias veces lo que de este modo ajusta los recuentos para múltiples sesgos. En algunas implementaciones, un módulo de corrección multivariante comprende uno o más módulos de densidad de sesgo 6, módulos de relación 8 y/o módulos de corrección de sesgo 10. A veces, un módulo de corrección multivariante proporciona recuentos normalizados 11 a otro módulo adecuado (por ejemplo, un módulo de distribución 21).

Porciones ponderadas

En algunas implementaciones se ponderan porciones. En algunas implementaciones, se ponderan una o más porciones lo que proporciona de este modo porciones ponderadas. Las porciones que se ponderan a veces eliminan las dependencias de porciones. Las porciones pueden ponderarse mediante un proceso adecuado. En algunas implementaciones, una o más porciones se ponderan por una función eigen (por ejemplo, una función propia). En algunas implementaciones, una función eigen comprende sustituir porciones con porciones de eigen ortogonales. En algunas implementaciones, un sistema comprende un módulo de ponderación de porción 42. En algunas implementaciones, un módulo de ponderación acepta, recupera y/o almacena densidades de lectura, perfiles de densidad de lectura y/o perfiles de densidad de lectura ajustados. En algunas implementaciones, las porciones ponderadas son proporcionadas por un módulo de ponderación de porción. En algunas implementaciones, se requiere un módulo de ponderación para las porciones de peso. Un módulo de ponderación puede ponderar las porciones mediante uno o varios métodos de ponderación conocidos en la técnica o descritos en el presente documento. Un módulo de ponderación a menudo proporciona porciones ponderadas a otro módulo adecuado (por ejemplo, un módulo de puntuación 46, un módulo de estadísticas de PCA 33, un módulo de generación de perfil 26 y similares).

Análisis de componentes principales

En algunas implementaciones, un perfil de densidad de lectura (por ejemplo, un perfil de densidad de lectura de una muestra de prueba (por ejemplo, la Figura 7A) se ajusta según un análisis de componentes principales (PCA). Un perfil de densidad de lectura de una o más muestras de referencia y/o un perfil de densidad de lectura de un sujeto de prueba se puede ajustar según un PCA. Un perfil de densidad de lectura para un genoma, parte de un genoma, un cromosoma o un segmento de un cromosoma puede ajustarse según un PCA. La eliminación de sesgo de un perfil de densidad de lectura por un proceso relacionado con PCA se denomina en el presente documento como ajuste de un perfil. Un PCA puede realizarse mediante un método PCA adecuado, o una variación del mismo. Los ejemplos no limitantes de un método PCA incluyen un análisis de correlación canónico (CCA), una transformación de Karhunen-Loeve (KLT), una transformación de Hotelling, una descomposición ortogonal adecuada (POD), una descomposición de valores singulares (SVD) de X, una descomposición de valores eigen (EVD) de XTX , un análisis de factor, un teorema de Eckart-Young, un teorema Schmidt-Mirsky, funciones empíricas ortogonales (EOF), una descomposición de función propia empírica, un análisis de componentes empíricos, modos cuasiarmónicos, una descomposición espectral, un análisis empírico modal, similares, variaciones o combinaciones de los mismos. Un PCA a menudo identifica uno o más sesgos en un perfil de densidad de lectura. Un sesgo identificado por un PCA a veces se denomina en el presente documento como un componente principal. En algunas implementaciones, uno o más sesgos pueden eliminarse ajustando un perfil de densidad de lectura según uno o más componentes principales usando un método adecuado. Se puede ajustar un perfil de densidad de lectura al agregar, restar, multiplicar y/o dividir uno o más componentes principales de un perfil de densidad de lectura. En algunas implementaciones, uno o más sesgos se pueden eliminar de un perfil de densidad de lectura al restar uno o más componentes principales de un perfil de densidad de lectura. Aunque el sesgo en un perfil de densidad de lectura a menudo se identifica y/o cuantifica mediante un PCA de un perfil, los componentes principales a menudo se restan de un perfil al nivel de densidades de lectura. Los sesgos o características en un perfil de densidad de lectura que se identifican y/o cuantifican por un PCA de un perfil incluyen, pero no se limitan a, sexo fetal, sesgo de secuencia (por ejemplo, sesgo de guanina y citosina (GC)), fracción fetal, sesgo correlacionado con la sensibilidad de ADNasa I, entropía, sesgo de secuencia repetitiva, sesgo de estructura de cromatina, sesgo de tasa de error de polimerasa, sesgo de palíndromo, sesgo de repetición invertido, sesgo de amplificación de PCR y variación de número de copias oculta.

Un PCA a menudo identifica uno o más componentes principales. En algunas implementaciones, un PCA identifica un 1.°, 2.°, 3.°, 4.°, 5.°, 6.°, 7.°, 8.°, 9.°, y un 10.° o más componentes principales. En determinadas implementaciones 1,

2, 3, 4, 5, 6, 7, 8, 9, 10 o más componentes principales se usan para ajustar un perfil. En determinadas implementaciones, se utilizan 5 componentes principales para ajustar un perfil. A menudo, los componentes principales se usan para ajustar un perfil en el orden de su aparición en un PCA. Por ejemplo, cuando tres componentes principales se sustraen de un perfil de densidad de lectura, un 1.º, 2.º y 3.º componentes principales se usan. A veces, un sesgo identificado por un componente principal comprende una característica de un perfil que no se usa para ajustar un perfil. Por ejemplo, un PCA puede identificar una variación genética (por ejemplo, una aneuploidía, delección, translocación, inserción) y/o una diferencia de género (por ejemplo, como se ve en la Figura 6C) como componente principal. Por lo tanto, en algunas implementaciones, no se usan uno o más componentes principales para ajustar un perfil. Por ejemplo, a veces se usan un 1.º, 2.º y 4.º componente principal para ajustar un perfil donde un 3.º componente principal no se usa para ajustar un perfil. Un componente principal se puede obtener a partir de un PCA usando cualquier muestra o referencia adecuada. En algunas implementaciones, se obtienen componentes principales a partir de una muestra de prueba (por ejemplo, un sujeto de prueba). En algunas implementaciones, los componentes principales se obtienen a partir de una o más referencias (por ejemplo, muestras de referencia, secuencias de referencia, un conjunto de referencia). Como se muestra, por ejemplo, en la Figura 6, se realiza un PCA en un perfil de densidad de lectura medio obtenido de un conjunto de entrenamiento (Figura 6A) que comprende múltiples muestras que dan como resultado la identificación de un 1.º componente principal (Figura 6B) y un segundo componente principal (Figura 6C). En algunas implementaciones, los componentes principales se obtienen a partir de un conjunto de sujetos que se sabe que carecen de una variación genética en cuestión. En algunas implementaciones, los componentes principales se obtienen a partir de un conjunto de euploides conocidos. Los componentes principales a menudo se identifican según un PCA realizado mediante el uso de uno o más perfiles de densidad de lectura de una referencia (por ejemplo, un conjunto de entrenamiento). Uno o más componentes principales obtenidos de una referencia a menudo se resta de un perfil de densidad de lectura de un sujeto de prueba (por ejemplo, la Figura 7B) lo que proporciona de este modo un perfil ajustado (por ejemplo, la Figura 7C).

En algunas implementaciones, un sistema comprende un módulo de estadísticas de PCA 33. Un módulo de estadísticas de PCA puede aceptar y/o recuperar perfiles de densidad de lectura de otro módulo adecuado (por ejemplo, un módulo de generación de perfil 26). Un PCA a menudo se realiza mediante un módulo de estadísticas de PCA. Un módulo de estadísticas de PCA a menudo acepta, recupera y/o almacena perfiles de densidad de lectura y procesa perfiles de densidad de lectura de un conjunto de referencia 32, conjunto de entrenamiento 30 y/o de uno o más sujetos de prueba 28. Un módulo de estadísticas de PCA puede generar y/o proporcionar componentes principales y/o ajustar los perfiles de densidad de lectura según uno o más componentes principales. Los perfiles de densidad de lectura ajustados (por ejemplo, 40, 38) a menudo se proporcionan mediante un módulo de estadísticas de PCA. Un módulo de estadísticas de PCA puede proporcionar y/o transferir perfiles de densidad de lectura ajustados (por ejemplo, 38, 40) a otro módulo adecuado (por ejemplo, un módulo de ponderación de porción 42, un módulo de puntuación 46). En algunas implementaciones, un módulo de estadísticas de PCA puede proporcionar una identificación de género 36. Una identificación de género es a veces una determinación del sexo fetal determinado según un PCA y/o según uno o más componentes principales. En algunas implementaciones, un módulo de estadísticas de PCA comprende algunas, todas o una modificación del código R que se muestra a continuación. Un código R para calcular componentes principales generalmente comienza con la limpieza de los datos (por ejemplo, restar la mediana, filtrar las porciones y recortar los valores extremos):

```
#Clean the data outliers for PCA
dclean <- (dat - m)[mask,] for (j in 1 :ncol(dclean)) { q <- quantile(dclean[,j],c(.25,.75))
qmin <- q[1] - 4*(q[2]-q[1]) qmax <- q[2] + 4*(q[2]-q[1]) dclean[dclean[,j] < qmin,j] <- qmin dclean[dclean[,j] > qmax,j] <-
qmax }
```

A continuación, se ordenaron los componentes principales:

```
#Compute principal components pc <- prcomp(dclean)$x
```

Finalmente, el perfil ajustado por PCA de cada muestra puede calcularse con:

```
#Compute residuals mm <- model.matrix(~pc[,1:numpc]) for (j in 1 :ncol(dclean)) dclean[,j] <- dclean[,j] -
predict(lm(dclean[,j]~mm))
```

Comparación de perfiles

En algunas implementaciones, determinar un resultado comprende una comparación. En determinadas implementaciones, se utiliza un perfil de densidad de lectura, o una porción del mismo, para proporcionar un resultado. En determinadas implementaciones, se utiliza un perfil de densidad de lectura para un genoma, parte de un genoma, un cromosoma o un segmento de un cromosoma para proporcionar un resultado. En algunas implementaciones, determinar un resultado (por ejemplo, una determinación de la presencia o ausencia de una variación genética) comprende una comparación de dos o más perfiles de densidad de lectura. Comparar los perfiles de densidad de lectura a menudo comprende comparar los perfiles de densidad de lectura generados para un segmento seleccionado de un genoma. Por ejemplo, un perfil de prueba a menudo se compara con un perfil de referencia donde los perfiles de prueba y referencia se determinaron para un segmento de un genoma (por ejemplo, un genoma de referencia) que es sustancialmente el mismo segmento. La comparación de los perfiles de densidad de lectura a veces comprende

comparar dos o más subconjuntos de porciones de un perfil de densidad de lectura. Un subconjunto de porciones de un perfil de densidad de lectura puede representar un segmento de un genoma (por ejemplo, un cromosoma o segmento del mismo). Un perfil de densidad de lectura puede comprender cualquier cantidad de subconjuntos de porciones. A veces, un perfil de densidad de lectura comprende dos o más, tres o más, cuatro o más, o cinco o más subconjuntos. En determinadas implementaciones, un perfil de densidad de lectura comprende dos subconjuntos de porciones donde cada porción representa segmentos de un genoma de referencia que son adyacentes. En algunas implementaciones, un perfil de prueba puede compararse con un perfil de referencia donde el perfil de prueba y el perfil de referencia comprenden ambos un primer subconjunto de porciones y un segundo subconjunto de porciones donde el primer y segundo subconjuntos representan diferentes segmentos de un genoma. Algunos subconjuntos de porciones de un perfil de densidad de lectura pueden comprender variaciones genéticas y otros subconjuntos de porciones a veces están sustancialmente libres de variaciones genéticas. A veces, todos los subconjuntos de porciones de un perfil (por ejemplo, un perfil de prueba) están sustancialmente libres de una variación genética. A veces, todos los subconjuntos de porciones de un perfil (por ejemplo, un perfil de prueba) comprenden una variación genética. En algunas implementaciones, un perfil de prueba puede comprender un primer subconjunto de porciones que comprende una variación genética y un segundo subconjunto de porciones que están sustancialmente libres de una variación genética.

En algunas implementaciones, los métodos descritos en el presente documento comprenden preformar una comparación (por ejemplo, comparar un perfil de prueba con un perfil de referencia). Dos o más conjuntos de datos, dos o más relaciones y/o dos o más perfiles pueden compararse mediante un método adecuado. Ejemplos no limitantes de métodos estadísticos adecuados para comparar conjuntos de datos, relaciones y/o perfiles incluyen enfoque de Behrens-Fisher, remuestreo tipo *bootstrapping*, método de Fisher para combinar pruebas independientes de significación, pruebas de Neyman-Pearson, análisis de confirmación de datos, análisis de datos exploratorios, prueba exacta, prueba F, prueba Z, prueba T, calcular y/o comparar una medida de incertidumbre, una hipótesis nula, recuentos y similares, una prueba de chi-cuadrado, prueba de omnibus, calcular y/o comparar el nivel de significación (por ejemplo, significación estadística), un metaanálisis, un análisis multivariado, una regresión, una regresión lineal simple, una regresión lineal robusta, similares o combinaciones de los anteriores. En determinadas implementaciones que comparan dos o más conjuntos de datos, relaciones y/o perfiles comprende determinar y/o comparar una medida de incertidumbre. Una “medida de incertidumbre” como se usa en el presente documento se refiere a una medida de significación (por ejemplo, significación estadística), una medida de error, una medida de varianza, una medida de confianza, similares o una combinación de las mismas. Una medida de incertidumbre puede ser un valor (por ejemplo, un umbral) o un rango de valores (por ejemplo, un rango, un rango de confianza, un rango de confianza bayesiano, un rango de umbral). Los ejemplos no limitantes de una medida de incertidumbre incluyen valores de p, una medida adecuada de desviación (por ejemplo, desviación estándar, sigma, desviación absoluta, desviación absoluta media, similares), una medida adecuada de error (por ejemplo, error estándar, error cuadrático medio, error cuadrático medio de raíz, similares), una medida de varianza adecuada, una puntuación estándar adecuada (por ejemplo, desviaciones estándar, porcentajes acumulativos, equivalentes de percentil, puntuaciones Z, puntuaciones T, puntuaciones R, estándar nueve (stanine), porcentaje en stanine, similares), similares o combinaciones de los mismos. En algunas implementaciones, determinar el nivel de significancia comprende determinar una medida de incertidumbre (por ejemplo, un valor de p). En determinadas implementaciones, dos o más conjuntos de datos relaciones y/o perfiles pueden analizarse y/o compararse utilizando métodos 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 cualquier manipulación matemática y/o estadística (por ejemplo, a las que se hace referencia en el presente documento como manipulaciones) adecuada.

En determinadas implementaciones que comparan dos o más perfiles de densidad de lectura comprende determinar y/o comparar una medida de incertidumbre para dos o más perfiles de densidad de lectura. Los perfiles de densidad de lectura y/o las medidas asociadas de incertidumbre se comparan a veces para facilitar la interpretación de manipulaciones matemáticas y/o estadísticas de un conjunto de datos y/o para proporcionar un resultado. Un perfil de densidad de lectura generado para un sujeto de prueba a veces se compara con un perfil de densidad de lectura generado para una o más referencias (por ejemplo, muestras de referencia, sujetos de referencia y similares). En algunas implementaciones, se proporciona un resultado comparando un perfil de densidad de lectura de un ensayo sometido a un perfil de densidad de lectura de una referencia para un cromosoma, porciones o segmentos del mismo, donde se obtiene un perfil de densidad de lectura de referencia a partir de un conjunto de sujetos de referencia que no se sabe que poseen una variación genética (por ejemplo, una referencia). En algunas implementaciones, se proporciona un resultado comparando un perfil de densidad de lectura de un ensayo sometido a un perfil de densidad de lectura de una referencia para un cromosoma, porciones o segmentos del mismo, donde se obtiene un perfil de densidad de lectura de referencia a partir de un conjunto de sujetos de referencia que se sabe que poseen una variación genética específica (por ejemplo, una aneuploidía cromosómica, una trisomía).

En determinadas implementaciones, un perfil de densidad de lectura de un sujeto de prueba se compara con un valor predeterminado representativo de la ausencia de una variación genética, y a veces se desvía de un valor predeterminado en una o más ubicaciones genómicas (por ejemplo, porciones) correspondientes a una ubicación genómica en la que se ubica una variación genética. Por ejemplo, en sujetos de prueba (por ejemplo, sujetos con

riesgo de, o que padecen una afección médica asociada con una variación genética), se espera que los perfiles de densidad de lectura difieran significativamente de los perfiles de densidad de lectura de una referencia (por ejemplo, una secuencia de referencia, sujeto de referencia, conjunto de referencia) para porciones seleccionadas cuando un sujeto de prueba comprende una variación genética en cuestión. Los perfiles de densidad de lectura de un sujeto de prueba a menudo son sustancialmente iguales a los perfiles de densidad de lectura de una referencia (por ejemplo, una secuencia de referencia, un sujeto de referencia, un conjunto de referencia) para porciones seleccionadas cuando un sujeto de prueba no comprende una variación genética en cuestión. Los perfiles de densidad de lectura a menudo se comparan con un umbral y/o rango de umbral predeterminados (por ejemplo, véase la Figura 8). El término “umbral” como se utiliza en el presente documento se refiere a cualquier número que se calcula usando un conjunto de datos calificados y sirve como límite de diagnóstico de una variación genética (por ejemplo, una variación en el número de copias, una aneuploidía, una aberración cromosómica y similares). En determinadas implementaciones, se supera un umbral por los resultados obtenidos mediante los métodos descritos en el presente documento y a un sujeto se le diagnostica una variación genética (por ejemplo, una trisomía). En algunas implementaciones, un valor umbral o rango de valores a menudo se calcula mediante manipulación matemática y/o estadísticamente de los datos de lectura de secuencia (por ejemplo, de una referencia y/o sujeto). Un umbral predeterminado o rango umbral de valores indicativos de la presencia o ausencia de una variación genética puede variar mientras se proporciona un resultado útil para determinar la presencia o ausencia de una variación genética. En determinadas implementaciones, se genera un perfil de densidad de lectura que comprende densidades de lectura normalizadas y/o recuentos normalizados para facilitar la clasificación y/o proporcionar un resultado. Puede proporcionarse un resultado basado en una representación de un perfil de densidad de lectura que comprende recuentos normalizados (por ejemplo, usando una representación gráfica de tal perfil de densidad de lectura).

En algunas implementaciones, un sistema comprende un módulo de ranurado 46. Un módulo de puntuación puede aceptar, recuperar y/o almacenar perfiles de densidad de lectura (por ejemplo, perfiles de densidad de lectura normalizados ajustados) de otro módulo adecuado (por ejemplo, un módulo de generación de perfil 26, un módulo de estadísticas de PCA 33, un módulo de ponderación de porción 42 y similares). Un módulo de puntuación puede aceptar, recuperar, almacenar y/o comparar dos o más perfiles de densidad de lectura (por ejemplo, perfiles de prueba, perfiles de referencia, conjuntos de entrenamiento, sujetos de prueba). Un módulo de puntuación a menudo puede proporcionar una puntuación (por ejemplo, una gráfica, estadísticas de perfil, una comparación (por ejemplo, una diferencia entre dos o más perfiles), una puntuación Z, una medida de incertidumbre, una zona de identificación, una identificación de muestra 50 (por ejemplo, una determinación de la presencia o ausencia de una variación genética) y/o un resultado). Un módulo de puntuación puede proporcionar una puntuación a un usuario final y/o a otro módulo adecuado (por ejemplo, una pantalla, impresora, similar). En algunas implementaciones, un módulo de puntuación comprende alguna, todas o una modificación del código R mostrado a continuación, que comprende una función R para calcular las estadísticas Chi-cuadrado para una prueba específica (por ejemplo, recuentos de alto chr21).

Los tres parámetros son:

x = datos de lectura de muestra (porción x muestra)

m = valores medianos para porciones

y = vector de prueba (Ej. False for all portions except True for chr21) `getChisqP <- function(x,m,y)`

```
{
  ahigh <- apply(x[!y,],2,function(x) sum((x>m[!y])))
  alow <- sum(!y)-ahigh
  bhigh <- apply(x[y,],2,function(x) sum((x>m[y])))
  blow <- sum(y)-bhigh
  p <- sapply(1:length(ahigh), function(i) {
    p <- chisq.test(matrix(c(ahigh[i],alow[i],bhigh[i],blow[i]),2))$p.value/2 if (ahigh[i]/alow[i] > bhigh[i]/blow[i]) p <- max(p,1-
    p)
    else p <- min(p,1-p); p})
  return(p)
}
```

Condiciones experimentales

En determinadas implementaciones, un proceso de normalización de componentes principales puede ajustar los sesgos asociados con las condiciones experimentales. El procesamiento de datos en vista de las condiciones experimentales se describe, por ejemplo, en la publicación de solicitud de patente internacional n°. WO 2013/109981.

En determinados casos, las muestras pueden verse 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) (por ejemplo, sesgo) 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 determinadas implementaciones, 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. El procesamiento de datos en vista de las condiciones experimentales se describe, por ejemplo, en la publicación de solicitud de patente internacional n°. WO 2013/109981.

45 Detección de aneuploidía usando comparaciones

En algunas implementaciones, se usa un proceso de normalización de componentes principales junto con un método para determinar la presencia o ausencia de una aneuploidía según una comparación. La detección de aneuploidía usando comparaciones se describe, por ejemplo, en la publicación de solicitud de patente internacional n°. WO 2014/116598.

En esta sección, una comparación de relaciones o relaciones o valores de relación, evaluación de ploidía y valor de evaluación de ploidía se denominan colectivamente “comparación”. En algunas implementaciones, la presencia o ausencia de una aneuploidía cromosómica en un sujeto se determina según una o más comparaciones. En algunas implementaciones, la presencia o ausencia de una aneuploidía cromosómica en un sujeto se determina según una o más comparaciones para tres autosomas seleccionados (por ejemplo, donde uno o más de los tres autosomas seleccionados es un cromosoma de prueba). En algunas implementaciones, la presencia o ausencia de una aneuploidía cromosómica se determina según una o más comparaciones generadas para un conjunto de cromosomas distintos, una región euploide, una región aneuploide o una región euploide y una región aneuploide. En algunas implementaciones, la presencia o ausencia de una aneuploidía cromosómica (por ejemplo, una aneuploidía cromosómica en un feto) se determina según una comparación obtenida para un sujeto y una región euploide y/o una región aneuploide (por ejemplo, una región euploide y una región aneuploide determinada para un conjunto de referencia). En determinadas implementaciones, la presencia o ausencia de una aneuploidía cromosómica se determina según una relación entre una comparación obtenida para un sujeto y una región euploide y/o una región aneuploide. Por ejemplo, la presencia o ausencia de una aneuploidía cromosómica se determina según si una comparación está en una región euploide o región aneuploide, o cuán lejos está un valor de evaluación de ploidía de

una región euploide o región aneuploide, en algunas implementaciones. En algunas implementaciones, una relación es una proximidad o una distancia (por ejemplo, una diferencia matemática y/o una distancia gráfica, por ejemplo, una distancia entre un punto y una región). Una relación puede determinarse mediante un método adecuado conocido en la técnica o descrito en el presente documento, cuyos ejemplos no limitantes incluyen distribución de probabilidad, función de densidad de probabilidad, función de distribución acumulativa, función de probabilidad, comparación de modelos bayesianos, factor de Bayes, criterio de información de desviación, pruebas de chi cuadrado, distancia euclidiana, análisis espacial, distancia de Mahalanobis, distancia de Manhattan, distancia de Chebyshev, distancia de Minkowski, divergencia de Bregman, distancia de Bhattacharyya, distancia de Hellinger, espacio métrico, distancia de Canberra, casco convexo (por ejemplo, regla de devanado par-impar), similares o combinaciones de los mismos.

En algunas implementaciones, la ausencia de una aneuploidía cromosómica se determina según una comparación y una región euploide. En algunas implementaciones, la ausencia de una aneuploidía cromosómica se determina según una relación entre una comparación y una región euploide. En algunas implementaciones, una comparación que se encuentra dentro, en o cerca de una región euploide es una determinación de un cromosoma euploide (por ejemplo, una ausencia de un cromosoma aneuploide). En algunas implementaciones, una comparación que está en o cerca de una región euploide indica que cada cromosoma, del que se determinó la comparación, es euploide. Por ejemplo, a veces una comparación generada según recuentos mapeados en ChrA, ChrB y ChrC se encuentra dentro de una región euploide (por ejemplo, una región euploide determinada según recuentos mapeados en ChrA, ChrB y ChrC) y se determina una ausencia de una aneuploidía cromosómica. En algunas implementaciones, la ausencia de una aneuploidía cromosómica, según se determina según una comparación, indica que cada cromosoma (por ejemplo, cada cromosoma del que se derivó el valor de evaluación de ploidía) es euploide (por ejemplo, euploide en una madre y/o feto).

En algunas implementaciones, una comparación que cae fuera de una región aneuploide es una determinación de uno o más cromosomas euploides. En algunas implementaciones, una comparación que está fuera de una región euploide indica que uno o más cromosomas, a partir de los cuales se determinó la comparación, son euploides. Por ejemplo, a veces una comparación generada según recuentos mapeados en ChrA, ChrB y ChrC se encuentra fuera de una región euploide (por ejemplo, una región euploide determinada según recuentos mapeados en ChrA, ChrB y ChrC) y se determina una ausencia de una aneuploidía cromosómica. En algunas implementaciones, una comparación que está fuera de una región euploide indica que dos de tres cromosomas usados para la comparación o evaluación, y de los cuales se determinó la comparación, son euploides.

En algunas implementaciones, una comparación cae dentro de una región aneuploide y uno o más cromosomas, a partir de los cuales se determinó la comparación, son euploides. Por ejemplo, a veces una comparación generada según recuentos mapeados en ChrA, ChrB y ChrC se encuentra dentro de una región aneuploide (por ejemplo, una región aneuploide determinada según recuentos mapeados en ChrA, ChrB y ChrC) y se determina una ausencia de una aneuploidía cromosómica para dos de los tres cromosomas.

En algunas implementaciones, la presencia de una aneuploidía cromosómica se determina según una comparación y una región euploide. En determinadas implementaciones, la presencia de una aneuploidía cromosómica se determina según una relación entre una comparación y una región euploide. En algunas implementaciones, una comparación que cae fuera de una región euploide es una determinación de un cromosoma aneuploide (por ejemplo, la presencia de un cromosoma aneuploide). En algunas implementaciones, una comparación que cae fuera de una región euploide indica que uno o más cromosomas, a partir de los cuales se determinó la comparación, es aneuploide. Por ejemplo, a veces una comparación generada según recuentos mapeados en ChrA, ChrB y ChrC se encuentra fuera de una región euploide (por ejemplo, una región euploide determinada según recuentos mapeados en ChrA, ChrB y ChrC) y se determina la presencia de una aneuploidía cromosómica.

En algunas implementaciones, una comparación que se encuentra dentro, en o cerca de una región aneuploide es una determinación de un cromosoma aneuploide (por ejemplo, una presencia de un cromosoma aneuploide). En algunas implementaciones, una comparación que está en o cerca de una región aneuploide indica que se determinó uno o más cromosomas, a partir de los cuales se determinó el valor de evaluación de ploidía, es aneuploide. En algunas implementaciones, una comparación que está en o cerca de una región aneuploide indica que 1, 2, 3, 4 y/o 5 cromosomas, a partir de los cuales se determinó la comparación, son aneuploides. En algunas implementaciones, una comparación que está en o cerca de una región aneuploide indica que uno de los tres cromosomas, a partir de los cuales se determinó la comparación, es aneuploide. Por ejemplo, a veces una comparación generada según recuentos mapeados en ChrA, ChrB y ChrC se encuentra dentro de una región aneuploide (por ejemplo, una región aneuploide determinada según recuentos mapeados en ChrA, ChrB y ChrC) y uno de los cromosomas es un cromosoma aneuploide.

En algunas implementaciones, una comparación que cae cerca de una región aneuploide es una determinación de un cromosoma aneuploide (por ejemplo, una presencia de un cromosoma aneuploide). En algunas implementaciones, una comparación que está cerca de una región aneuploide indica que uno o más cromosomas, a partir de los cuales se determinó la comparación, es aneuploide. En algunas implementaciones, un gráfico de referencia comprende una región euploide definida y tres regiones aneuploides definidas (por ejemplo, aneuploides para Chr13, Chr18 o Chr21) y una determinación de la presencia de una aneuploidía se realiza según una comparación que cae más cerca de una

de las regiones aneuploides. Por ejemplo, una comparación que está más cerca de una región aneuploide para Chr21 que a otra región (por ejemplo, una región aneuploide para Chr13 o Chr18, o una región euploide) puede indicar la presencia de una aneuploidía para Chr21.

5 En algunas implementaciones, una comparación generada según recuentos mapeados en Chr13, Chr18 y Chr21 se encuentra dentro de una región aneuploide (por ejemplo, una región aneuploide determinada según recuentos mapeados en Chr13, Chr18 y Chr21) y uno de los cromosomas es un cromosoma aneuploide. En algunas implementaciones, una comparación generada según recuentos mapeados en Chr13, Chr18 y Chr21 se encuentra dentro de una región aneuploide (por ejemplo, una región aneuploide determinada según recuentos mapeados en Chr13, Chr18 y Chr21), Chr18 y Chr21 se determina que son euploides y Chr13 se determina que es aneuploide. En algunas implementaciones, una comparación generada según recuentos mapeados en Chr13, Chr18 y Chr21 se encuentra dentro de una región aneuploide (por ejemplo, una región aneuploide determinada según recuentos mapeados en Chr13, Chr18 y Chr21), Chr13 y Chr21 se determina que son euploides y Chr18 se determina que es aneuploide. En algunas implementaciones, una comparación generada según recuentos mapeados en Chr13, Chr18 y Chr21 se encuentra dentro de una región aneuploide (por ejemplo, una región aneuploide determinada según recuentos mapeados en Chr13, Chr18 y Chr21), Chr18 y Chr13 se determina que son euploides y Chr21 se determina que es aneuploide.

20 En algunas implementaciones, la presencia o ausencia de una aneuploidía cromosómica se determina según una primera comparación y una segunda comparación donde ambas comparaciones se generan a partir de lecturas de secuencia mapeadas en el mismo conjunto de dos o más cromosomas. En algunas implementaciones, la presencia o ausencia de una aneuploidía cromosómica en un sujeto se determina según una relación (por ejemplo, una distancia) entre una primera comparación generada para un sujeto y una segunda comparación generada para un segundo sujeto. En algunas implementaciones, una segunda comparación es un conjunto de comparaciones (por ejemplo, una región) generada para uno o más sujetos. En algunas implementaciones, la presencia o ausencia de una aneuploidía cromosómica en un sujeto se determina según una relación (por ejemplo, una distancia) entre una primera comparación generada para el sujeto y un conjunto de referencia de comparaciones generadas para uno o más sujetos. En algunas implementaciones, una primera comparación es una comparación para un sujeto y una segunda comparación es una comparación o un conjunto de comparaciones que representan uno o más fetos euploides. En algunas implementaciones, una segunda comparación es un valor o conjunto de valores (por ejemplo, una región) esperada para un feto euploide. En algunas implementaciones, una segunda comparación es un valor o conjunto de valores generados para un sujeto (por ejemplo, un sujeto femenino gestante) donde se sabe que un feto es euploide para uno o más de los cromosomas a partir de los cuales se generó la comparación. En algunas implementaciones, la distancia se determina según un valor de incertidumbre (por ejemplo, una desviación estándar o DAM). En algunas implementaciones, la distancia entre una primera y una segunda comparación (por ejemplo, una segunda comparación representativa de uno o más sujetos euploides) es 1, 2, 3, 4, 5, 6 o más veces una incertidumbre asociada y la primera comparación se determina que es aneuploide. En algunas implementaciones, la distancia entre una primera y una segunda comparación (por ejemplo, una segunda comparación representativa de uno o más sujetos euploides) es 3 o más veces una incertidumbre asociada y la primera comparación se determina para representar un cromosoma aneuploide.

45 En algunas implementaciones, la presencia o ausencia de un cromosoma aneuploide se determina según una comparación generada según recuentos mapeados en uno o más cromosomas específicos y una región euploide, una región aneuploide, o una región euploide y una región aneuploide. En algunas implementaciones, la presencia o ausencia de una aneuploide cromosómica se determina según una comparación generada según las lecturas de secuencia mapeadas en uno o más cromosomas específicos y las lecturas de secuencia mapeadas en otros cromosomas no son necesarias para la determinación. En algunas implementaciones, la presencia o ausencia de una aneuploide cromosómica se determina según una comparación generada según las lecturas de secuencia mapeadas en 2, 3, 4, 5 o 6 cromosomas distintos y recuentos mapeados en otros cromosomas no se obtienen o requieren para la determinación. En algunas implementaciones, la presencia o ausencia de un cromosoma aneuploide se determina según una comparación generada según tres cromosomas distintos o segmentos de los mismos y la determinación no se basa en un cromosoma distinto de uno de los tres cromosomas distintos. Por ejemplo, cuando ChrA, ChrB y ChrC representan tres cromosomas distintos o segmentos de los mismos, la presencia o ausencia de un cromosoma aneuploide se determina a veces según una comparación generada según ChrA, ChrB y ChrC y la determinación no se basa en otro cromosoma distinto de ChrA, ChrB o ChrC. En algunas implementaciones, ChrA, ChrB y ChrC representan Chr13, Chr21 y Chr18 respectivamente.

Cariotipo de cromosoma sexual

60 En algunas implementaciones, se usa un proceso de normalización de componentes principales junto con un método para determinar un cariotipo de cromosoma sexual. Los métodos para determinar el cariotipo de cromosoma sexual se describen, por ejemplo, en la publicación de solicitud de patente internacional n°. WO 2013/192562.

65 En algunas implementaciones, los recuentos de lectura de secuencia que se correlacionan con uno o más cromosomas sexuales (es decir, cromosoma X, cromosoma Y) están normalizados. En algunas implementaciones, la normalización comprende una normalización de componentes principales. En algunas implementaciones, la

normalización implica determinar un sesgo experimental para porciones de un genoma de referencia. En algunas implementaciones, el sesgo experimental se puede determinar para múltiples muestras de una primera relación ajustada (por ejemplo, relación lineal ajustada, relación no lineal ajustada) para cada muestra entre recuentos de lecturas de secuencia mapeadas en cada una de las porciones de un genoma de referencia y una característica de mapeo (por ejemplo, contenido de GC) para cada una de las porciones. La pendiente de una relación ajustada (por ejemplo, relación lineal) generalmente se determina mediante regresión lineal. En algunas implementaciones, cada sesgo experimental está representado por un coeficiente de sesgo experimental. El coeficiente de sesgo experimental es la pendiente de una relación lineal entre, por ejemplo, (i) recuentos de lecturas de secuencia mapeadas en cada una de las porciones de un genoma de referencia, y (ii) una característica de mapeo para cada una de las porciones. En algunas implementaciones, el sesgo experimental puede comprender una estimación de curvatura de sesgo experimental.

En algunas implementaciones, un método comprende además calcular un nivel de sección genómica (por ejemplo, una elevación, un nivel) para cada una de las porciones genómicas de una segunda relación ajustada (por ejemplo, relación lineal ajustada, relación no lineal ajustada) entre el sesgo experimental y los recuentos de lecturas de secuencia mapeadas en cada una de las porciones y la pendiente de la relación puede determinarse mediante regresión lineal. Por ejemplo, si la primera relación ajustada es lineal y la segunda relación ajustada es lineal, el nivel de sección genómica L_i puede determinarse para cada una de las porciones del genoma de referencia según la Ecuación α :

$$L_i = (m_i - G_i S) I^{-1} \quad \text{Equation } \alpha$$

en donde G_i es el sesgo experimental, I es la ordenada en el origen de la segunda relación ajustada, S es la pendiente de la segunda relación, m_i son los recuentos medidos mapeados en cada porción del genoma de referencia e i es una muestra.

En algunas implementaciones, se aplica un proceso de normalización secundario a uno o más niveles de sección genómica calculados. En algunas implementaciones, la normalización secundaria comprende normalización de GC y a veces comprende el uso de la metodología PERUN. En algunas implementaciones, la normalización secundaria comprende una normalización de componentes principales.

Determinación de ploidía fetal

En algunas implementaciones, se usa un proceso de normalización de componentes principales junto con un método para determinar la ploidía fetal. Los métodos para determinar la ploidía fetal se describen, por ejemplo, en la publicación de solicitud de patente estadounidense n°. US 2013/0288244.

Se puede determinar una ploidía fetal, en parte, a partir de una medida de la fracción fetal y la determinación de ploidía fetal se usa para hacer una determinación de la presencia o ausencia de una variación genética (por ejemplo, una aneuploidía cromosómica, una trisomía). Una ploidía fetal puede determinarse, en parte, a partir de una medida de la fracción fetal determinada por cualquier método adecuado de determinación de la fracción fetal, incluidos los métodos descritos en el presente documento. En algunas implementaciones, el método requiere un recuento de referencia calculado F_i (a veces representado como f_i) determinado para una porción (es decir, un bin, i) de un genoma para múltiples muestras donde se sabe que la ploidía del feto para la porción i del genoma es euploide. En algunas implementaciones, se determina un valor de incertidumbre (por ejemplo, una desviación estándar, σ) para el recuento de referencia f_i . En algunas implementaciones, se utilizan un recuento de referencia f_i , un valor de incertidumbre, un recuento de muestra de prueba y/o una fracción fetal medida (F) para determinar la ploidía fetal. En algunas implementaciones, se normaliza un recuento de referencia (por ejemplo, un recuento de referencia promedio, medio o en mediana) mediante una normalización de componentes principales y/u otra normalización, tal como, por ejemplo, normalización en bins, normalización en porciones, normalización por contenido de GC, regresión lineal y no lineal de mínimos cuadrados, LOESS, LOESS de GC, LOWESS, PERUN, RM, GCRM y/o combinaciones de los mismos). En algunas implementaciones, un recuento de referencia de un segmento de un genoma que se sabe que es euploide es igual a 1 cuando el recuento de referencia se normaliza mediante normalización de componentes principales. En algunas implementaciones, tanto el recuento de referencia (por ejemplo, para un feto que se sabe que es euploide) como los recuentos de una muestra de prueba para una porción o segmento de un genoma están normalizados mediante normalización de componentes principales y el recuento de referencia es igual a 1. En algunas implementaciones, un recuento de referencia de un segmento de un genoma que se sabe que es euploide es igual a 1 cuando el recuento de referencia se normaliza mediante PERUN. En algunas implementaciones, tanto el recuento de referencia (por ejemplo, para un feto que se sabe que es euploide) como los recuentos de una muestra de prueba para una porción o segmento de un genoma están normalizados mediante PERUN y el recuento de referencia es igual a 1. Asimismo, en algunas implementaciones, un recuento de referencia de una porción o segmento de un genoma que se sabe que es euploide es igual a 1 cuando los recuentos se normalizan (es decir, se dividen entre) mediante una mediana del recuento de referencia. Por ejemplo, en algunas implementaciones, tanto el recuento de referencia (por ejemplo, para un feto que se sabe que es euploide) como los recuentos de una muestra de prueba para una porción o segmento de un genoma se normalizan mediante una mediana de un recuento de referencia, el recuento de referencia normalizado es igual a 1 y el recuento de la muestra de prueba se normaliza (por ejemplo, se divide entre)

mediante la mediana del recuento de referencia. En algunas implementaciones, tanto el recuento de referencia (por ejemplo, para un feto que se sabe que es euploide) como los recuentos de una muestra de prueba para una porción o segmento de un genoma se normalizan mediante normalización de componentes principales, GCRM, GC, RM o un método adecuado. En algunas implementaciones, un recuento de referencia es un recuento de referencia promedio, medio o en mediana. Un recuento de referencia suele ser un recuento normalizado para un bin (por ejemplo, un nivel de sección genómica normalizado). En algunas implementaciones, un recuento de referencia y los recuentos de una muestra de prueba son recuentos sin procesar. Un recuento de referencia, en algunas implementaciones, se determina a partir de un perfil de recuento promedio, medio o en mediana. En algunas implementaciones, un recuento de referencia es un nivel de sección genómica calculado. En algunas implementaciones, un recuento de referencia de una muestra de referencia y un recuento de una muestra de prueba (por ejemplo, una muestra de paciente, por ejemplo, y) se normalizan mediante el mismo método o proceso.

Procesamiento y normalización de datos adicionales

Las lecturas de secuencia mapeadas que se han contado se denominan en el presente documento datos sin procesar, ya que los datos representan recuentos sin manipular (por ejemplo, recuentos sin procesar). En algunas implementaciones, los datos de lectura de secuencia en un conjunto de datos pueden procesarse adicionalmente (por ejemplo, manipularse matemática y/o estadísticamente) y/o visualizarse para facilitar la provisión de un resultado. En determinadas implementaciones, los conjuntos de datos, incluyendo conjuntos de datos más grandes, pueden beneficiarse del preprocesamiento para facilitar un análisis adicional. El preprocesamiento de conjuntos de datos a veces implica la eliminación de porciones o porciones redundantes y/o poco informativas de un genoma de referencia (por ejemplo, porciones de un genoma de referencia con datos poco informativos, lecturas mapeadas redundantes, porciones con cero recuentos de mediana, secuencias representadas en exceso o en defecto). Sin desear limitarse por la teoría, el procesamiento y/o preprocesamiento de datos pueden (i) eliminar datos con ruido, (ii) eliminar datos no informativos, (iii) eliminar datos redundantes, (iv) reducir la complejidad de conjuntos de datos más grandes y/o (v) facilitar la transformación de los datos de una forma en una o más formas diferentes. Los términos “preprocesamiento” y “procesamiento”, cuando se usan con respecto a los datos o conjuntos de datos, se denominan colectivamente en el presente documento “procesamiento”. El procesamiento puede hacer que los datos sean más susceptibles de análisis adicional, y puede generar un resultado en algunas implementaciones. En algunas implementaciones, uno o más o todos los métodos de procesamiento (por ejemplo, métodos de normalización, filtrado de porciones, mapeo, validación, similares o combinaciones de los mismos) son realizados por un procesador, un microprocesador, un ordenador, junto con la memoria y/o por un aparato controlado por microprocesador.

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

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

Puede usarse cualquier procedimiento adecuado para procesar conjuntos de datos descritos en el presente documento. Los ejemplos no limitativos de procedimientos adecuados para su uso para procesar conjuntos de datos incluyen filtrado, normalización, ponderación, monitorización de alturas de pico, monitorización de áreas de pico, monitorización de bordes de pico, determinación de razones de área, procesamiento matemático de datos,

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

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

En algunas implementaciones, una o más etapas de procesamiento pueden comprender una o más etapas de filtrado. El término “filtrado”, tal y como se utiliza en el presente documento, se refiere a la eliminación de porciones o porciones de un genoma de referencia de la consideración. Las porciones de un genoma de referencia 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, porciones de un genoma de referencia con cero recuentos de mediana), porciones de un genoma de referencia con secuencias sobrerrepresentadas o subrepresentadas, datos con ruido, similares o combinaciones de los anteriores. Un proceso de filtrado a menudo implica eliminar una o más porciones de un genoma de referencia de la consideración y restar los recuentos en la una o más porciones de un genoma de referencia seleccionadas para su eliminación de los recuentos contados o sumados para las porciones de un genoma de referencia, cromosoma o cromosomas, o genoma en consideración. En algunas implementaciones, las porciones de un genoma de referencia pueden eliminarse sucesivamente (por ejemplo, de una en una para permitir la evaluación del efecto de la eliminación de cada porción individual) y, en determinadas implementaciones, todas las porciones de un genoma de referencia marcadas para su eliminación pueden eliminarse al mismo tiempo. En algunas implementaciones, se eliminan porciones de un genoma de referencia caracterizadas por una varianza por encima o por debajo de un determinado nivel, lo que a veces se denomina en el presente documento filtrado de porciones “con ruido” de un genoma de referencia. En determinadas implementaciones, un proceso de filtrado comprende la obtención de puntos de datos de un conjunto de datos que se desvían del nivel medio del perfil de una porción, un cromosoma o un segmento de un cromosoma en un múltiplo predeterminado de la varianza del perfil, y en determinadas implementaciones, un proceso de filtrado comprende la eliminación de puntos de datos de un conjunto de datos que no se desvían del nivel medio del perfil de una porción, un cromosoma o segmento de un cromosoma en un múltiplo predeterminado de la varianza del perfil. En algunas implementaciones, se utiliza un proceso de filtrado para reducir el número de porciones candidatas de un genoma de referencia analizado para detectar la presencia o ausencia de una variación genética. Reducir el número de porciones candidatas de un genoma de referencia analizado 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 algunas implementaciones, una o más etapas de procesamiento pueden comprender una o más etapas de normalización. La normalización puede realizarse mediante un método descrito en el presente documento adecuado o conocido en la técnica. En determinadas implementaciones, la normalización comprende ajustar los valores medidos en diferentes escalas a una escala teóricamente común. En determinadas implementaciones, la normalización comprende un ajuste matemático sofisticado para llevar a alineación distribuciones de probabilidad de valores

ajustados. En algunas implementaciones, la normalización comprende alinear distribuciones a una distribución normal. En determinadas implementaciones, 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). En determinadas implementaciones, la normalización comprende el escalado. La normalización comprende a veces la división de uno o más conjuntos de datos por una variable o fórmula predeterminada. La normalización comprende a veces la sustracción 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 porciones, normalización por contenido de GC, normalización por recuento de mediana (recuento de mediana de bins, mediana de recuento de porciones), regresión lineal y no lineal por mínimos cuadrados, LOESS, LOESS de GC, LOWESS (suavizado de diagrama de dispersión ponderado localmente), PERUN, ChAI, normalización de componentes principales, enmascaramiento de repetición (RM), normalización de GC y enmascaramiento de repetición (GCRM), cQn y/o combinaciones de los mismos. En algunas implementaciones, la determinación de una presencia o ausencia de una variación genética (por ejemplo, una aneuploidía, una microduplicación, una microdelección) utiliza un método de normalización (por ejemplo, normalización basada en porciones, normalización por contenido de GC, normalización por recuento de mediana (recuento de mediana de bins, mediana de recuento de porciones), regresión lineal y no lineal por mínimos cuadrados, LOESS, LOESS de GC, LOWESS (suavizado de diagrama de dispersión ponderado localmente), PERUN, ChAI, normalización de componentes principales, 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). En algunas implementaciones, la determinación de una presencia o ausencia de una variación genética (por ejemplo, una aneuploidía, una microduplicación, una microdelección) utiliza uno o más de LOESS, recuento medio (mediana del recuento de intervalos, recuento medio de porciones) y normalización de componentes principales. En algunas implementaciones, la determinación de una presencia o ausencia de una variación genética utiliza LOESS seguido de la normalización de la mediana (mediana del recuento de intervalos, recuento medio de porciones). En algunas implementaciones, la determinación de una presencia o ausencia de una variación genética utiliza LOESS seguido de una mediana de recuento (mediana del recuento de intervalos, recuento de porciones medianas) seguida de normalización de componentes principales.

Puede usarse cualquier número adecuado de normalizaciones. En algunas implementaciones, los conjuntos de datos pueden normalizarse 1 o más, 5 o más, 10 o más o incluso 20 o más veces. Los conjuntos de datos pueden normalizarse a valores (por ejemplo, valor de normalización) representativos de cualquier característica o variable adecuada (por ejemplo, datos de muestra, datos de referencia, o ambos). Los ejemplos no limitativos de tipos de normalizaciones de datos que se pueden usar incluyen normalizar los datos de recuento sin procesar para una o más pruebas o porciones 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 porción o secciones genómicas seleccionadas; normalizar los datos de recuento sin procesar de una o varias porciones seleccionadas con respecto a la mediana de un recuento de referencia de una o varias porciones o del cromosoma en el que se mapea una o varias porciones seleccionadas; 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 algunas implementaciones, pueden usarse una o más normalizaciones con respecto a un valor derivado estadísticamente para minimizar las diferencias de los datos y disminuir la importancia de los datos atípicos. La normalización de porciones, o porciones de un genoma de referencia, con respecto a un valor de normalización, a veces se denomina “normalización basada en porciones”.

En determinadas implementaciones, una etapa de procesamiento que comprende la normalización incluye la normalización con respecto a una ventana estática y, en algunas implementaciones, una etapa de procesamiento que comprende normalización incluye la normalización con respecto a una ventana móvil o deslizante. El término “ventana”, tal como se usa en el presente documento, se refiere a una o más porciones elegidas para el análisis y, algunas veces, usadas como referencia para la comparación (por ejemplo, usadas para la normalización y/u otra manipulación matemática o estadística). La expresión “normalizar con respecto a una ventana estática”, tal como se usa en el presente documento, se refiere a un procedimiento de normalización que usa una o más porciones seleccionadas para la comparación entre un sujeto de prueba y el conjunto de datos del sujeto de referencia. En algunas implementaciones, las porciones seleccionadas se utilizan para generar un perfil. Una ventana estática incluye generalmente un conjunto predeterminado de porciones que no cambian durante las manipulaciones y/o el análisis. Las expresiones “normalizar con respecto a una ventana móvil” y “normalizar con respecto a una ventana deslizante”, tal y como se utilizan en el presente documento, se refieren a normalizaciones realizadas con respecto a porciones ubicadas en la región genómica (por ejemplo, el entorno genético inmediato, la porción o secciones adyacentes, y similares) de una porción de prueba seleccionada, donde una o más porciones de prueba seleccionadas se normalizan con respecto a porciones que rodean inmediatamente a la porción de prueba seleccionada. En determinadas implementaciones, las porciones seleccionadas se utilizan para generar un perfil. Una normalización de ventana móvil o deslizante incluye a menudo desplazarse o deslizarse repetidamente a una porción de prueba adyacente, y normalizar la porción de prueba recién seleccionada a porciones inmediatamente circundantes o adyacentes a la

porción de prueba recién seleccionada, donde las ventanas adyacentes tienen una o más porciones en común. En determinadas implementaciones, una pluralidad de porciones y/o cromosomas de prueba seleccionados pueden analizarse mediante un procedimiento de ventana deslizante.

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

20 Se describen con mayor detalle a continuación en el presente documento determinados ejemplos de procesos de normalización que pueden utilizarse, tales como métodos de normalización LOESS, PERUN, ChAI y componentes principales, por ejemplo.

25 En algunas implementaciones, una etapa de procesamiento comprende una ponderación. Los términos “ponderado” y “ponderación” o la expresión “función de ponderación” o sus derivados gramaticales o equivalentes, tal y como se utilizan en el presente documento, se refieren a una manipulación matemática de una porción o de la totalidad de un conjunto de datos utilizada en ocasiones 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 importancia y/o la contribución de los datos contenidos en una o varias porciones o porciones de un genoma de referencia, en función de la calidad o la utilidad de los datos de la porción o porciones seleccionadas de un genoma de referencia). Una función de ponderación puede usarse para aumentar la influencia de los datos con una varianza de medición relativamente pequeña, y/o para disminuir la influencia de los datos con una varianza de medición relativamente grande, en algunas implementaciones. Por ejemplo, las porciones de un genoma de referencia con datos de secuencia poco representados o de baja calidad pueden “ponderarse a la baja” para minimizar la influencia en un conjunto de datos, mientras que las porciones seleccionadas de un genoma de referencia pueden “ponderarse al alza” 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 algunas implementaciones, un conjunto de datos se divide entre una variable predeterminada (por ejemplo, variable de ponderación). A menudo, se selecciona una variable predeterminada (por ejemplo, función objetivo minimizada, Phi) para ponderar distintas partes de un conjunto de datos de manera diferente (por ejemplo, aumentar la influencia de determinados tipos de datos mientras se reduce la influencia de otros tipos de datos).

45 En determinadas implementaciones, una etapa de procesamiento puede comprender una o más manipulaciones matemáticas y/o estadísticas. Cualquier manipulación matemática y/o estadística adecuada, sola o en combinación, puede usarse para analizar y/o manipular un conjunto de datos descrito en el presente documento. Puede usarse cualquier número adecuado de manipulaciones matemáticas y/o estadísticas. En algunas implementaciones, un conjunto de datos puede manipularse matemática y/o estadísticamente 1 o más, 5 o más, 10 o más o incluso 20 o más veces. Los ejemplos no limitativos de manipulaciones matemáticas y estadísticas que pueden usarse incluyen suma, resta, multiplicación, división, funciones algebraicas, estimadores de mínimos cuadrados, ajuste de curvas, ecuaciones diferenciales, polinomios racionales, polinomios dobles, polinomios ortogonales, puntuaciones z, valores de p, valores de chi, valores de phi, análisis de niveles de pico, determinación de ubicaciones de bordes de pico, cálculo de razones de áreas de pico, análisis de la mediana de nivel cromosómico, 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 niveles, niveles medios, distribución de recuentos dentro de una región genómica, representación relativa de especies de ácido nucleico, similares o combinaciones de los mismos.

65 En algunas implementaciones, una etapa de procesamiento puede comprender el uso de uno o más algoritmos estadísticos. Cualquier algoritmo estadístico adecuado, solo o en combinación, puede usarse para analizar y/o manipular un conjunto de datos descrito en el presente documento. Puede usarse cualquier número adecuado de algoritmos estadísticos. En algunas implementaciones, un conjunto de datos puede analizarse usando 1 o más, 5 o más, 10 o más o incluso 20 o más algoritmos estadísticos. Los ejemplos no limitativos de algoritmos estadísticos

adecuados para su uso con los métodos descritos en el presente documento incluyen árboles de decisión, valores contranulos, comparaciones múltiples, prueba ómnibus, problema de Behrens-Fisher, remuestreo de tipo *bootstrapping*, método de Fisher para combinar pruebas independientes de significación, hipótesis nula, error tipo I, error tipo II, prueba exacta, prueba Z de una muestra, prueba Z de dos muestras, prueba de la t de una muestra, prueba de la t para datos emparejados, prueba de la t agrupada de dos muestras que tienen varianzas iguales, prueba de la t no agrupada de dos muestras que tienen varianzas desiguales, prueba z de una proporción, prueba z de dos proporciones agrupadas, prueba z de dos proporciones no agrupadas, prueba de chi cuadrado de una muestra, prueba F de dos muestras para determinar la igualdad de varianzas, intervalo de confianza, intervalo creíble, significación, metaanálisis, regresión lineal simple, regresión lineal robusta, similares o combinaciones de los anteriores. Los ejemplos no limitativos de variables o características del conjunto de datos que pueden analizarse usando algoritmos estadísticos incluyen recuentos sin procesar, recuentos filtrados, recuentos normalizados, alturas de pico, anchuras de pico, bordes de pico, tolerancias laterales, valores de p, mediana de niveles, niveles medios, distribución de recuentos dentro de una región genómica, representación relativa de especies de ácido nucleico, similares o combinaciones de los mismos.

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

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

En algunas implementaciones, las porciones se pueden filtrar según una medida de error (por ejemplo, desviación estándar, error estándar, varianza calculada, valor de p, error absoluto medio (EAM), desviación absoluta promedio y/o desviación absoluta media (DAM)). En determinadas implementaciones, una medida de error se refiere a la variabilidad del recuento. En algunas implementaciones, las porciones se filtran según la variabilidad del recuento. En determinadas implementaciones, la variabilidad del recuento es una medida de error determinado para los recuentos mapeados en una porción (es decir, una porción) de un genoma de referencia para múltiples muestras (por ejemplo, múltiples muestras obtenidas de múltiples sujetos, por ejemplo, 50 o más, 100 o más, 500 o más 1000 o más, 5000 o más o 10.000 o más sujetos). En algunas implementaciones, las porciones con una variabilidad de recuento por encima de un rango superior predeterminado se filtran (por ejemplo, se excluyen de la consideración). En algunas implementaciones, un rango superior predeterminado es un valor de la DAM igual o superior a aproximadamente 50, aproximadamente 52, aproximadamente 54, aproximadamente 56, aproximadamente 58, aproximadamente 60, aproximadamente 62, aproximadamente 64, aproximadamente 66, aproximadamente 68, aproximadamente 70, aproximadamente 72, aproximadamente de 74 o igual o superior a aproximadamente 76. En algunas implementaciones, las porciones con una variabilidad de recuento por debajo de un rango inferior predeterminado se filtran (por ejemplo, se excluyen de la consideración). En algunas implementaciones, un rango inferior predeterminado es un valor de la DAM igual o inferior a aproximadamente 40, aproximadamente 35, aproximadamente 30, aproximadamente 25, aproximadamente 20, aproximadamente 15, aproximadamente 10, aproximadamente 5, aproximadamente 1, o igual o inferior a aproximadamente 0. En algunas implementaciones, las porciones con una variabilidad de recuento fuera de un rango predeterminado se filtran (por ejemplo, se excluyen de la consideración). En algunas implementaciones, un rango predeterminado es un valor de la DAM superior a cero e inferior a

aproximadamente 76, inferior a aproximadamente 74, inferior a aproximadamente 73, inferior a aproximadamente 72, inferior a aproximadamente 71, inferior a aproximadamente 70, inferior a aproximadamente 69, inferior a aproximadamente 68, inferior a aproximadamente 67, inferior a aproximadamente 66, inferior a aproximadamente 65, inferior a aproximadamente 64, inferior a aproximadamente 62, inferior a aproximadamente 60, inferior a aproximadamente 58, inferior a aproximadamente 56, inferior a aproximadamente 54, inferior a aproximadamente 52 o inferior a aproximadamente 50. En algunas implementaciones, un rango predeterminado es un valor de la DAM superior a cero e inferior a aproximadamente 67,7. En algunas implementaciones, se seleccionan porciones con una variabilidad de recuento dentro de un rango predeterminado (por ejemplo, se usan para determinar la presencia o ausencia de una variación genética).

En algunas implementaciones, la variabilidad del recuento de porciones representa una distribución (por ejemplo, una distribución normal). En algunas implementaciones, las porciones se seleccionan dentro de un cuantil de la distribución. En algunas implementaciones, se seleccionan las porciones dentro de un cuantil igual o inferior a aproximadamente el 99,9 %, 99,8 %, 99,7 %, 99,6 %, 99,5 %, 99,4 %, 99,3 %, 99,2 %, 99,1 %, 99,0 %, 98,9 %, 98,8 %, 98,7 %, 98,6 %, 98,5 %, 98,4 %, 98,3 %, 98,2 %, 98,1 %, 98,0 %, 97 %, 96 %, 95 %, 94 %, 93 %, 92 %, 91 %, 90 %, 85 %, 80 %, o igual o inferior a un cuantil de aproximadamente el 75 % para la distribución. En algunas implementaciones, se seleccionan porciones dentro de un cuantil del 99 % de la distribución de la variabilidad del recuento. En algunas implementaciones, las porciones con DAM > 0 y DAM < 67,725 están dentro del cuantil del 99 % y se seleccionan, lo que resulta en la identificación de un conjunto de porciones estables de un genoma de referencia.

En el presente documento y en la solicitud de patente internacional n.º PCT/US12/59123 (WO2013/052913), se proporcionan ejemplos no limitativos de filtrado de porciones con respecto a PERUN, por ejemplo. Las porciones pueden filtrarse basándose en, o basándose en parte en, una medida de error. Una medida de error que comprende valores absolutos de desviación, tal como un factor R, puede usarse para la eliminación o ponderación de porciones en determinadas implementaciones. Un factor R, en algunas implementaciones, se define como la suma de las desviaciones absolutas de los valores de recuento predichos con respecto a las mediciones reales dividida por los valores de recuento predichos con respecto a las mediciones reales. Aunque puede usarse una medida de error que comprende valores absolutos de desviación, puede usarse alternativamente una medida de error adecuada. En determinadas implementaciones puede usarse una medida de error que no comprende valores absolutos de desviación, tal como una dispersión basada en cuadrados. En algunas implementaciones, las porciones se filtran o se ponderan según una medida de mapeabilidad (por ejemplo, una puntuación de mapeabilidad). A veces, una porción se filtra o se pondera según un número relativamente bajo de lecturas de secuencia mapeadas en la porción (por ejemplo, 0, 1, 2, 3, 4, 5 lecturas mapeadas en la porción). Las porciones pueden filtrarse o ponderarse según el tipo de análisis que se realice. Por ejemplo, para el análisis de aneuploidía del cromosoma 13, 18 y/o 21, los cromosomas sexuales pueden filtrarse, y pueden analizarse solo autosomas, o un subconjunto de autosomas.

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

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

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

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

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

En algunas implementaciones, los conjuntos de datos se procesan usando uno o más análisis de nivel de pico, análisis de anchura de pico, análisis de ubicación de borde de pico, tolerancias laterales de pico, similares, derivaciones de los mismos o combinaciones de los anteriores, como parte de o después de procesar y/o manipular los conjuntos de datos. En algunas implementaciones, se genera un gráfico de perfiles de datos procesados usando uno o más análisis de nivel 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 nivel de pico, análisis de anchura de pico, análisis de ubicación de borde de pico, tolerancias laterales de pico, similares, derivaciones de los mismos o combinaciones de los anteriores.

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

En algunas implementaciones, el análisis y procesamiento de datos pueden incluir el uso de una o más suposiciones. Puede usarse un número o tipo adecuado de suposiciones para analizar o procesar un conjunto de datos. Los ejemplos no limitativos de suposiciones que pueden usarse para el procesamiento y/o análisis de datos incluyen ploidía materna, contribución fetal, prevalencia de determinadas secuencias en una población de referencia, origen étnico, prevalencia de una afección médica seleccionada en miembros de la familia emparentados, paralelismo entre perfiles de recuento sin procesar de diferentes pacientes y/o ejecuciones después de la normalización de GC y enmascaramiento de repetición (por ejemplo, GCRM), 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. La expresión “perfil de recuento normalizado”, tal como se usa en el presente documento, se refiere a un perfil generado usando recuentos normalizados. Los ejemplos de métodos que pueden usarse para generar recuentos normalizados y perfiles de recuento normalizados se describen en el presente documento. Tal como se indicó, las lecturas de secuencia mapeadas que se han contado pueden normalizarse con respecto a los recuentos de muestras de prueba o recuentos

de muestras de referencia. En algunas implementaciones, un perfil de recuento normalizado puede presentarse como una representación gráfica.

Normalización LOESS

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 algunas implementaciones, aplica un modelo de LOESS a la relación entre el recuento de fragmentos (por ejemplo, lecturas de secuencia, recuentos) y composición de GC para porciones de un genoma de referencia. 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.

Normalización PERUN

Una metodología de normalización 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) descrita en el presente documento y en la Publicación de solicitud de patente internacional n.º WO2013/052913. 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.

En determinadas implementaciones, la metodología PERUN incluye el cálculo de un nivel de sección genómica para porciones de un genoma de referencia a partir de (a) recuentos de lectura de secuencias mapeados en una porción de un genoma de referencia para una muestra de prueba, (b) sesgo experimental (por ejemplo, sesgo de GC) para la muestra de prueba, y (c) uno o más parámetros de ajuste (por ejemplo, estimaciones de ajuste) para una relación ajustada entre (i) el sesgo experimental para una porción de un genoma de referencia en el que se mapean las lecturas de secuencia y (ii) los recuentos de secuencia lecturas mapeados en la porción. El sesgo experimental para cada una de las porciones de un genoma de referencia puede determinarse a través de múltiples muestras según una relación ajustada para cada muestra entre (i) los recuentos de lecturas de secuencia mapeadas en cada una de las porciones de un genoma de referencia, y (ii) una característica de mapeo para cada una de las porciones de un genoma de referencia. Esta relación ajustada para cada muestra puede ensamblarse para múltiples muestras en tres dimensiones. El conjunto puede ordenarse según el sesgo experimental en determinadas implementaciones, aunque la metodología PERUN puede ponerse en práctica sin ordenar el conjunto según el sesgo experimental. La relación ajustada para cada muestra y la relación ajustada para cada porción del genoma de referencia pueden ajustarse independientemente a una función lineal o función no lineal mediante un proceso de ajuste adecuado conocido en la técnica.

Normalización de regresión híbrida

En algunas implementaciones, se utiliza un método de normalización híbrida. En algunas implementaciones, un método de normalización híbrida reduce el sesgo (por ejemplo, sesgo de GC). Una normalización híbrida, en algunas implementaciones, comprende (i) un análisis de una relación de dos variables (por ejemplo, recuentos y contenido de GC) y (ii) la selección y aplicación de un método de normalización según el análisis. Una normalización híbrida, en determinadas implementaciones, comprende (i) una regresión (por ejemplo, un análisis de regresión) y (ii) la selección y aplicación de un método de normalización según la regresión. En algunas implementaciones, los recuentos obtenidos para una primera muestra (por ejemplo, un primer conjunto de muestras) se normalizan mediante un método diferente al de los recuentos obtenidos de otra muestra (por ejemplo, un segundo conjunto de muestras). En algunas implementaciones, los recuentos obtenidos para una primera muestra (por ejemplo, un primer conjunto de muestras) se normalizan mediante un primer método de normalización y los recuentos obtenidos a partir de una segunda muestra (por ejemplo, un segundo conjunto de muestras) se normalizan mediante un segundo método de normalización. Por ejemplo, en determinadas implementaciones, un primer método de normalización comprende el uso de una regresión lineal y un segundo método de normalización comprende el uso de una regresión no lineal (por ejemplo, una regresión LOESS, GC-LOESS, LOWESS, suavizado LOESS).

En algunas implementaciones, se utiliza un método de normalización híbrida para normalizar las lecturas de secuencia mapeadas en porciones de un genoma o cromosoma (por ejemplo, recuentos, recuentos mapeados, lecturas mapeadas). En determinadas implementaciones, los recuentos sin procesar se normalizan y, en algunas implementaciones, los recuentos ajustados, ponderados, filtrados o previamente normalizados se normalizan

mediante un método de normalización híbrida. En determinadas implementaciones, se normalizan los niveles de sección genómica o las puntuaciones Z. En algunas implementaciones, los recuentos mapeados en porciones seleccionadas de un genoma o cromosoma se normalizan mediante un enfoque de normalización híbrida. Los recuentos pueden referirse a una medida adecuada de lecturas de secuencia mapeadas en porciones de un genoma, cuyos ejemplos no limitativos incluyen recuentos sin manipular (por ejemplo, recuentos sin procesar), recuentos normalizados (por ejemplo, normalizados mediante PERUN, ChAI, normalización de componentes principales, o un método adecuado), niveles de porción (por ejemplo, niveles promedio, niveles medios, mediana de los niveles o similares), puntuaciones Z, similares o combinaciones de los mismos. Los recuentos pueden ser recuentos sin procesar o recuentos procesados de una o más muestras (por ejemplo, una muestra de prueba, una muestra de una mujer embarazada). En algunas implementaciones, los recuentos se obtienen de una o más muestras obtenidas de uno o más sujetos.

En algunas implementaciones, se selecciona un método de normalización (por ejemplo, el tipo de método de normalización) según una regresión (por ejemplo, un análisis de regresión) y/o un coeficiente de correlación. Un análisis de regresión se refiere a una técnica estadística para estimar una relación entre variables (por ejemplo, recuentos y contenido de GC). En algunas implementaciones, se genera una regresión según los recuentos y una medida del contenido de GC para cada porción de múltiples porciones de un genoma de referencia. Se puede utilizar una medida adecuada del contenido de GC, cuyos ejemplos no limitativos incluyen una medida del contenido de guanina, citosina, adenina, timina, purina (GC) o pirimidina (AT o ATU), temperatura de fusión (T_f) (por ejemplo, temperatura de desnaturalización, temperatura de combinación, temperatura de hibridación), una medida de energía libre, similares o combinaciones de los mismos. Una medida del contenido de guanina (G), citosina (C), adenina (A), timina (T), purina (GC) o pirimidina (AT o ATU) se puede expresar como una relación o un porcentaje. En algunas implementaciones se utiliza cualquier relación o porcentaje adecuado, cuyos ejemplos no limitativos incluyen GC/AT, GC/nucleótido total, GC/A, GC/T, AT/nucleótido total, AT/GC, AT/G, AT/C, G/A, C/A, G/T, G/A, G/AT, C/T, similares o combinaciones de los mismos. En algunas implementaciones, una medida del contenido de GC es una relación o un porcentaje de GC con respecto al contenido total de nucleótidos. En algunas implementaciones, una medida del contenido de GC es una relación o un porcentaje de GC con respecto al contenido total de nucleótidos para lecturas de secuencia mapeadas en una porción del genoma de referencia. En determinadas implementaciones, el contenido de GC se determina según y/o a partir de lecturas de secuencia mapeadas en cada porción de un genoma de referencia y las lecturas de secuencia se obtienen de una muestra (por ejemplo, una muestra obtenida de una mujer embarazada). En algunas implementaciones, una medida del contenido de GC no se determina según y/o a partir de las lecturas de secuencia. En determinadas implementaciones, se determina una medida del contenido de GC para una o más muestras obtenidas de uno o más sujetos.

En algunas implementaciones, la generación de una regresión comprende generar un análisis de regresión o un análisis de correlación. Se puede usar una regresión adecuada, cuyos ejemplos no limitativos incluyen un análisis de regresión (por ejemplo, un análisis de regresión lineal), un análisis de bondad de ajuste, un análisis de correlación de Pearson, una correlación de rango, una fracción de varianza no explicada, un análisis de eficiencia del modelo de Nash-Sutcliffe, validación del modelo de regresión, reducción proporcional de la pérdida, desviación cuadrática media, similares o una combinación de los mismos. En algunas implementaciones, se genera una línea de regresión. En determinadas implementaciones, la generación de una regresión comprende generar una regresión lineal. En determinadas implementaciones, la generación de una regresión comprende generar una regresión no lineal (por ejemplo, una regresión LOESS, una regresión LOWESS).

En algunas implementaciones, una regresión determina la presencia o ausencia de una correlación (por ejemplo, una correlación lineal), por ejemplo, entre recuentos y una medida del contenido de GC. En algunas implementaciones, se genera una regresión (por ejemplo, una regresión lineal) y se determina un coeficiente de correlación. En algunas implementaciones, se determina un coeficiente de correlación adecuado, cuyos ejemplos no limitativos incluyen un coeficiente de determinación, un valor de R^2 , un coeficiente de correlación de Pearson, o similar.

En algunas implementaciones, se determina la bondad de ajuste para una regresión (por ejemplo, un análisis de regresión, una regresión lineal). La bondad de ajuste a veces se determina mediante análisis visual o matemático. Una evaluación a veces incluye determinar si la bondad de ajuste es mayor para una regresión no lineal o para una regresión lineal. En algunas implementaciones, un coeficiente de correlación es una medida de la bondad de ajuste. En algunas implementaciones, una evaluación de la bondad de ajuste para una regresión se determina según un coeficiente de correlación y/o un valor de corte del coeficiente de correlación. En algunas implementaciones, una evaluación de la bondad de ajuste comprende comparar un coeficiente de correlación con un valor de corte del coeficiente de correlación. En algunas implementaciones, una evaluación de la bondad de ajuste de una regresión es indicativa de una regresión lineal. Por ejemplo, en determinadas implementaciones, la bondad de ajuste es mayor para una regresión lineal que para una regresión no lineal, y la evaluación de la bondad de ajuste es indicativa de una regresión lineal. En algunas implementaciones, una evaluación es indicativa de una regresión lineal, y se utiliza una regresión lineal para normalizar los recuentos. En algunas implementaciones, una evaluación de la bondad de ajuste de una regresión es indicativa de una regresión no lineal. Por ejemplo, en determinadas implementaciones, la bondad de ajuste es mayor para una regresión no lineal que para una regresión lineal, y la evaluación de la bondad de ajuste es indicativa de una regresión no lineal. En algunas implementaciones, una evaluación es indicativa de una regresión no lineal, y se utiliza una regresión no lineal para normalizar los recuentos.

En algunas implementaciones, una evaluación de la bondad de ajuste es indicativa de una regresión lineal cuando un coeficiente de correlación es igual o superior a un valor de corte del coeficiente de correlación. En algunas implementaciones, una evaluación de la bondad de ajuste es indicativa de una regresión no lineal cuando un coeficiente de correlación es inferior a un valor de corte del coeficiente de correlación. En algunas implementaciones, se predetermina un valor de corte del coeficiente de correlación. En algunas implementaciones, un valor de corte del coeficiente de correlación es de aproximadamente 0,5 o mayor, aproximadamente 0,55 o mayor, aproximadamente 0,6 o mayor, aproximadamente 0,65 o mayor, aproximadamente 0,7 o mayor, aproximadamente 0,75 o mayor, aproximadamente 0,8 o mayor o aproximadamente 0,85 o mayor.

Por ejemplo, en determinadas implementaciones, se usa un método de normalización que comprende una regresión lineal cuando un coeficiente de correlación es igual o superior a aproximadamente 0,6. En determinadas implementaciones, los recuentos de una muestra (por ejemplo, recuentos por porción de un genoma de referencia, recuentos por porción) se normalizan según una regresión lineal cuando un coeficiente de correlación es igual o superior a un valor de corte del coeficiente de correlación de 0,6; de lo contrario, los recuentos se normalizan según una regresión no lineal (por ejemplo, cuando el coeficiente es inferior a un valor de corte del coeficiente de correlación de 0,6). En algunas implementaciones, un proceso de normalización comprende generar una regresión lineal o una regresión no lineal para (i) los recuentos y (ii) el contenido de GC, para cada porción de múltiples porciones de un genoma de referencia. En determinadas implementaciones, se usa un método de normalización que comprende una regresión no lineal (por ejemplo, un método LOWESS o LOESS) cuando un coeficiente de correlación es inferior a un valor de corte del coeficiente de correlación de 0,6. En algunas implementaciones, se usa un método de normalización que comprende una regresión no lineal (por ejemplo, un método LOWESS) cuando un coeficiente de correlación (por ejemplo, un coeficiente de correlación) es inferior a un valor de corte del coeficiente de correlación de aproximadamente 0,7, inferior a aproximadamente 0,65, inferior a aproximadamente 0,6, inferior a aproximadamente 0,55 o inferior a aproximadamente 0,5. Por ejemplo, en algunas implementaciones se usa un método de normalización que comprende una regresión no lineal (por ejemplo, un método LOWESS o LOESS) cuando un coeficiente de correlación es inferior a un valor de corte del coeficiente de correlación de aproximadamente 0,6.

En algunas implementaciones, se selecciona un tipo específico de regresión (por ejemplo, una regresión lineal o no lineal) y, después de generar la regresión, los recuentos se normalizan restando la regresión a los recuentos. En algunas implementaciones, la resta de una regresión a los recuentos proporciona recuentos normalizados con un sesgo reducido (por ejemplo, sesgo de GC). En algunas implementaciones, se resta una regresión lineal a los recuentos. En algunas implementaciones, se resta una regresión no lineal (por ejemplo, una regresión LOESS, GC-LOESS, LOWESS) a los recuentos. Se puede utilizar cualquier método adecuado para restar una línea de regresión a los recuentos. Por ejemplo, si los recuentos x se derivan de la porción i (por ejemplo, una porción i) que comprende un contenido de GC de 0,5 y una línea de regresión determina los recuentos y con un contenido de GC de 0,5, entonces $x - y =$ recuentos normalizados para la porción i . En algunas implementaciones, los recuentos se normalizan antes y/o después de restar una regresión. En algunas implementaciones, se utilizan recuentos normalizados mediante un enfoque de normalización híbrida para generar niveles de sección genómica, puntuaciones Z , niveles y/o perfiles de un genoma o un segmento del mismo. En determinadas implementaciones, los recuentos normalizados mediante un enfoque de normalización híbrida se analizan mediante métodos descritos en el presente documento para determinar la presencia o ausencia de una variación genética (por ejemplo, en un feto).

En algunas implementaciones, un método de normalización híbrida comprende filtrar o ponderar una o más porciones antes o después de la normalización. Se puede utilizar un método adecuado para filtrar porciones, incluidos los métodos de filtrado de porciones (por ejemplo, porciones de un genoma de referencia) descritos en el presente documento. En algunas implementaciones, las porciones (por ejemplo, porciones de un genoma de referencia) se filtran antes de aplicar un método de normalización híbrida. En algunas implementaciones, solo se normalizan mediante una normalización híbrida los recuentos de lecturas de secuenciación mapeadas en porciones seleccionadas (por ejemplo, porciones seleccionadas según la variabilidad del recuento). En algunas implementaciones, los recuentos de lecturas de secuenciación mapeadas en porciones filtradas de un genoma de referencia (por ejemplo, porciones filtradas según la variabilidad del recuento) se eliminan antes de utilizar un método de normalización híbrida. En algunas implementaciones, un método de normalización híbrida comprende seleccionar o filtrar porciones (por ejemplo, porciones de un genoma de referencia) según un método adecuado (por ejemplo, un método descrito en el presente documento). En algunas implementaciones, un método de normalización híbrida comprende seleccionar o filtrar porciones (por ejemplo, porciones de un genoma de referencia) según un valor de incertidumbre para los recuentos mapeados en cada una de las porciones para múltiples muestras de prueba. En algunas implementaciones, un método de normalización híbrida comprende seleccionar o filtrar porciones (por ejemplo, porciones de un genoma de referencia) según la variabilidad del recuento. En algunas implementaciones, un método de normalización híbrida comprende seleccionar o filtrar porciones (por ejemplo, porciones de un genoma de referencia) según el contenido de GC, elementos repetitivos, secuencias repetitivas, intrones, exones, similares o una combinación de los mismos.

Por ejemplo, en algunas implementaciones se analizan múltiples muestras de múltiples sujetos de sexo femenino gestantes y se selecciona un subconjunto de porciones (por ejemplo, porciones de un genoma de referencia) según la variabilidad del recuento. En determinadas implementaciones se utiliza una regresión lineal para determinar un coeficiente de correlación para (i) recuentos y (ii) contenido de GC, para cada una de las porciones seleccionadas

para una muestra obtenida de un sujeto de sexo femenino gestante. En algunas implementaciones, se determina un coeficiente de correlación que es superior a un valor de corte de correlación predeterminado (por ejemplo, de aproximadamente 0,6), una evaluación de la bondad de ajuste es indicativa de una regresión lineal y los recuentos se normalizan restando la regresión lineal a los condes. En determinadas implementaciones, se determina un coeficiente de correlación que es inferior a un valor de corte de correlación predeterminado (por ejemplo, de aproximadamente 0,6), una evaluación de la bondad de ajuste es indicativa de una regresión no lineal, se genera una regresión LOESS y los recuentos se normalizan restando la regresión LOESS a los recuentos.

Perfiles

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

En algunas implementaciones, un perfil es representativo de la totalidad de un conjunto de datos y, en determinadas implementaciones, un perfil es representativo de una parte o subconjunto de un conjunto de datos. Es decir, un perfil a veces incluye o se genera a partir de puntos de datos representativos de datos que no se han filtrado para eliminar cualquier dato y, a veces, un perfil incluye o se genera a partir de puntos de datos representativos de datos que se han filtrado para eliminar datos no deseados. En algunas implementaciones, un punto de datos en un perfil representa los resultados de la manipulación de datos para una porción. En determinadas implementaciones, un punto de datos en un perfil incluye resultados de la manipulación de datos para grupos de porciones. En algunas implementaciones, los grupos de porciones pueden ser adyacentes entre sí y, en determinadas implementaciones, los grupos de porciones 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: porciones basadas en el tamaño, porciones basadas en características de secuencia (por ejemplo, contenido de GC, contenido de AT, posición en un cromosoma (por ejemplo, brazo corto, brazo largo, centrómero, telómero) y similares), niveles de expresión, cromosoma, similares o combinaciones de los mismos. En algunas implementaciones, puede generarse un perfil a partir de puntos de datos obtenidos de otro perfil (por ejemplo, perfil de datos normalizado renormalizado a un valor de normalización diferente para generar un perfil de datos renormalizado). En determinadas implementaciones, un perfil generado a partir de los puntos de datos obtenidos a partir de otro perfil reduce el número de puntos de datos y/o la complejidad del conjunto de datos. Reducir el número de puntos de datos y/o la complejidad de un conjunto de datos facilita a menudo la interpretación de los datos y/o facilita proporcionar un resultado.

Un perfil (por ejemplo, un perfil genómico, un perfil cromosómico, un perfil de un segmento de un cromosoma) suele ser una colección de recuentos normalizados o no normalizados para dos o más porciones. Un perfil suele incluir al menos un nivel (por ejemplo, un nivel de sección genómica), y suele comprender dos o más niveles (por ejemplo, un perfil suele tener múltiples niveles). Un nivel generalmente corresponde a un conjunto de porciones que tienen aproximadamente los mismos recuentos o recuentos normalizados. Los niveles se describen con mayor detalle en el presente documento. En determinadas implementaciones, un perfil comprende una o más porciones, que pueden ser ponderadas, eliminadas, filtradas, normalizadas, ajustadas, promediadas, derivadas como media, sumadas, restadas, procesadas o transformadas mediante cualquier combinación de las mismas. Un perfil suele comprender recuentos normalizados mapeados en porciones que definen dos o más niveles, donde los recuentos se normalizan aún más según una de los niveles mediante un método adecuado. A menudo, los recuentos de un perfil (por ejemplo, un nivel de perfil) se asocian con un valor de incertidumbre.

Un perfil que comprende una o más niveles a veces se rellena (por ejemplo, relleno de huecos). El relleno (por ejemplo, relleno de huecos) se refiere a un proceso de identificación y ajuste de niveles en un perfil que se deben a microdeleciones maternas o duplicaciones maternas (por ejemplo, variaciones en el número de copias). En algunas implementaciones se rellenan niveles que se deben a microduplicaciones fetales o microdeleciones fetales. Las microduplicaciones o microdeleciones en un perfil pueden, en algunas implementaciones, aumentar o disminuir artificialmente el nivel general de un perfil (por ejemplo, un perfil de un cromosoma), lo que lleva a determinaciones positivas falsas o negativas falsas de una aneuploidía cromosómica (por ejemplo, una trisomía). En algunas implementaciones, los niveles en un perfil que se deben a microduplicaciones y/o eliminaciones se identifican y ajustan

(por ejemplo, se rellenan y/o se eliminan) mediante un proceso que a veces se denomina relleno o relleno de huecos. En determinadas implementaciones, un perfil comprende uno o más primeros niveles que son significativamente diferentes de un segundo nivel dentro del perfil, cada uno de los primeros niveles comprende 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 uno o más de los primeros niveles se ajustan.

Un perfil que comprende uno o más niveles puede incluir un primer nivel y un segundo nivel. En algunas implementaciones un primer nivel es diferente (por ejemplo, significativamente diferente) que un segundo nivel. En algunas implementaciones, un primer nivel comprende un primer conjunto de porciones, un segundo nivel comprende un segundo conjunto de porciones, y el primer conjunto de porciones no es un subconjunto del segundo conjunto de porciones. En determinadas implementaciones, un primer conjunto de porciones es diferente de un segundo conjunto de porciones a partir del cual se determinan un primer y un segundo nivel. En algunas implementaciones, un perfil puede tener múltiples primeros niveles que son diferentes (por ejemplo, significativamente diferentes, por ejemplo, tienen un valor significativamente diferente) que un segundo nivel dentro del perfil. En algunas implementaciones, un perfil comprende uno o más primeros niveles que son significativamente diferentes que un segundo nivel dentro del perfil, y se ajustan uno o más de los primeros niveles. En algunas implementaciones, un perfil comprende uno o más primeros niveles que son significativamente diferentes de un segundo nivel dentro del perfil, cada uno de los primeros niveles comprende 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 uno o más de los primeros niveles se ajustan. En algunas implementaciones, un primer nivel dentro de un perfil se elimina del perfil o se ajusta (por ejemplo, se rellena). Un perfil puede comprender múltiples niveles que incluyen uno o más primeros niveles significativamente diferentes de uno o más segundos niveles y a menudo la mayoría de los niveles en un perfil son segundos niveles, segundos niveles que son aproximadamente iguales entre sí. En algunas implementaciones, más del 50 %, más del 60 %, más del 70 %, más del 80 %, más del 90 % o más del 95 % de los niveles de un perfil son segundos niveles.

Algunas veces, un perfil se visualiza como una representación gráfica. Por ejemplo, pueden representarse gráficamente y visualizarse uno o más niveles que representen recuentos (por ejemplo, recuentos normalizados) de porciones. Los ejemplos no limitativos de gráficos de perfiles que pueden generarse incluyen el recuento sin procesar (por ejemplo, perfil de recuento sin procesar o perfil sin procesar), recuento normalizado, ponderado en porciones, puntuación z, valor de p, razón de área frente a ploidía ajustada, mediana de nivel frente a razón entre la fracción fetal ajustada y medida, componentes principales, similares o combinaciones de los mismos. Los gráficos de perfiles permiten la visualización de los datos manipulados, en algunas implementaciones. En determinadas implementaciones, puede usarse un gráfico de perfiles para proporcionar un resultado (por ejemplo, razón de área frente a ploidía ajustada, mediana de nivel frente a razón entre fracción fetal ajustada y medida, componentes principales). Las expresiones “representación gráfica de perfil de recuento en bruto” o “representación gráfica de perfil sin procesar”, tal y como se utilizan en el presente documento, se refieren a una representación gráfica de recuentos en cada porción de una región normalizados con respecto a los recuentos totales de una región (por ejemplo, genoma, porción, cromosoma, porciones cromosómicas de un genoma de referencia o un segmento de un cromosoma). En algunas implementaciones, puede generarse un perfil usando un procedimiento de ventana estática y, en determinadas implementaciones, puede generarse un perfil usando un procedimiento de ventana deslizante.

A veces, un perfil generado para un sujeto de prueba se compara con un perfil generado para uno o más sujetos de referencia, para facilitar la interpretación de manipulaciones matemáticas y/o estadísticas de un conjunto de datos y/o para proporcionar un resultado. En algunas implementaciones, se genera un perfil basándose en una o más suposiciones iniciales (por ejemplo, contribución materna de ácido nucleico (por ejemplo, fracción materna), contribución fetal de ácido nucleico (por ejemplo, fracción fetal), ploidía de la muestra de referencia, similares o combinaciones de los mismos). En determinadas implementaciones, un perfil de prueba se centra a menudo alrededor de un valor predeterminado representativo de la ausencia de una variación genética y se desvía a menudo de un valor predeterminado en áreas correspondientes a la ubicación genómica en la que está ubicada la variación genética en el sujeto de prueba, si el sujeto de prueba presentase la variación genética. En los sujetos de prueba en riesgo de, o que padecen, una afección médica asociada con una variación genética, se espera que el valor numérico para una porción seleccionada varíe significativamente con respecto al valor predeterminado para ubicaciones genómicas no afectadas. Dependiendo de las suposiciones iniciales (por ejemplo, ploidía fija o ploidía optimizada, fracción fetal fija o fracción fetal optimizada o combinaciones de las mismas) el umbral predeterminado o valor de punto de corte o rango umbral de valores indicativos de la presencia o ausencia de una variación genética puede variar mientras todavía proporciona un resultado útil para determinar la presencia o ausencia de una variación genética. En algunas implementaciones, un perfil es indicativo de y/o representativo de un fenotipo.

A modo de ejemplo no limitativo, los perfiles normalizados de recuento de muestras y/o de referencia pueden obtenerse a partir de datos de lectura de secuencias sin procesar mediante (a) calcular la mediana de los recuentos de referencia para cromosomas, porciones o segmentos de los mismos seleccionados a partir de un conjunto de referencias que se sabe que no portan una variación genética, (b) eliminar las porciones no informativas de los recuentos sin procesar de las muestras de referencia (por ejemplo, filtrado); (c) normalizar los recuentos de referencia para todas las porciones restantes de un genoma de referencia con respecto al número residual total de recuentos (por ejemplo, la suma de los recuentos restantes después de eliminar las porciones no informativas de un genoma de referencia) para el

cromosoma seleccionado de la muestra de referencia o ubicación genómica seleccionada, lo que genera de este modo un perfil de sujeto de referencia normalizado; (d) eliminar las porciones correspondientes de la muestra del sujeto de prueba; y (e) normalizar los recuentos de sujetos de prueba restantes para una o más ubicaciones genómicas seleccionadas con respecto a la suma de la mediana de los recuentos de referencia residuales para el cromosoma o cromosomas que contienen las ubicaciones genómicas seleccionadas, generando de esta manera un perfil de sujeto de prueba normalizado. En determinadas implementaciones, una etapa de normalización adicional con respecto a todo el genoma, reducida por las porciones filtradas en (b), puede incluirse entre (c) y (d).

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

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

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

Niveles

En algunas implementaciones, se atribuye un valor (por ejemplo, un número, un valor cuantitativo) a un nivel. Un nivel puede determinarse mediante un método, una operación o un procedimiento matemático adecuado (por ejemplo, un nivel procesado). A menudo, un nivel es, o se deriva de, recuentos (por ejemplo, recuentos normalizados) para un conjunto de porciones. En algunas implementaciones, el nivel de una porción es sustancialmente igual al número total de recuentos mapeados en una porción (por ejemplo, recuentos, recuentos normalizados). Con frecuencia, un nivel 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. En algunas implementaciones, un nivel 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, nivel medio), sumados, restados, transformados o combinaciones de los mismos. En algunas implementaciones, un nivel comprende recuentos que están normalizados (por ejemplo, recuentos normalizados de porciones). Un nivel puede ser para recuentos normalizados mediante un procedimiento adecuado, los ejemplos no limitativos de los cuales incluyen normalización basada en porciones, normalización por contenido de GC, normalización por recuento de mediana, regresión lineal y no lineal por mínimos cuadrados, LOESS (por ejemplo, LOESS de GC), LOWESS, PERUN, ChAI, normalización de componentes principales, RM, GCRM, cQn, similares y/o combinaciones de los mismos. Un nivel puede comprender recuentos normalizados o cantidades relativas de recuentos. En algunas implementaciones un nivel es para recuentos o recuentos normalizados de dos o más porciones que se promedian, y el nivel se denomina nivel medio. En algunas implementaciones un nivel es para un conjunto de porciones que tienen un recuento medio o media de recuentos normalizados que se denomina nivel medio. En algunas implementaciones se deriva un nivel para porciones que comprenden recuentos sin procesar y/o filtrados. En algunas implementaciones, un nivel se basa en recuentos que están sin procesar. En algunas implementaciones, un nivel está asociado con un valor de incertidumbre (por ejemplo, una desviación estándar, una DAM). En algunas implementaciones, un nivel se representa mediante una puntuación Z o un valor de p . En el presente documento, un nivel para una o más porciones es sinónimo de un “nivel de sección genómica”.

En el presente documento, un nivel para una o más porciones es sinónimo de un “nivel de sección genómica”. El término “nivel”, tal como se usa en el presente documento, es a veces sinónimo del término “elevación”. Una determinación del significado del término “nivel” puede determinarse a partir del contexto en el que se usa. Por ejemplo, el término “nivel”, cuando se usa en el contexto de secciones genómicas, perfiles, lecturas y/o recuentos a menudo significa una elevación. El término “nivel”, cuando se usa en el contexto de una sustancia o composición (por ejemplo, nivel de ARN, nivel de plexación) a menudo se refiere a una cantidad. El término “nivel”, cuando se usa en el contexto de incertidumbre (por ejemplo, nivel de error, nivel de confianza, nivel de desviación, nivel de incertidumbre) a menudo se refiere a una cantidad.

Los recuentos normalizados o no normalizados para dos o más niveles (por ejemplo, dos o más niveles en un perfil) pueden, algunas veces, manipularse matemáticamente (por ejemplo, sumarse, multiplicarse, promediarse, normalizarse, similares o una combinación de los mismos) según los niveles. Por ejemplo, los recuentos normalizados o no normalizados para dos o más niveles pueden normalizarse según una, algunas o todos los niveles de un perfil. En algunas implementaciones, los recuentos normalizados o no normalizados de todos los niveles en un perfil se normalizan según un nivel del perfil. En algunas implementaciones, los recuentos normalizados o no normalizados de un primer nivel en un perfil se normalizan según los recuentos normalizados o no normalizados de un segundo nivel en el perfil.

Los ejemplos no limitativos de un nivel (por ejemplo, un primer nivel, un segundo nivel) son un nivel para un conjunto de porciones que comprenda recuentos procesados, un nivel para un conjunto de porciones que comprenda una media, mediana o promedio de recuentos, un nivel para un conjunto de porciones que comprenda recuentos normalizados, similares o cualquier combinación de los mismos. En algunas implementaciones, un primer nivel y un segundo nivel en un perfil se derivan de recuentos de porciones mapeadas en el mismo cromosoma. En algunas implementaciones, un primer nivel y un segundo nivel en un perfil se derivan de recuentos de porciones mapeadas en cromosomas diferentes.

En algunas implementaciones, un nivel se determina a partir de recuentos normalizados o no normalizados mapeados en una o más porciones. En algunas implementaciones, un nivel se determina a partir de recuentos normalizados o no normalizados mapeados en dos o más porciones, donde los recuentos normalizados para cada porción suelen ser aproximadamente los mismos. Puede haber variación en los recuentos (por ejemplo, recuentos normalizados) en un conjunto de porciones para un nivel. En un conjunto de porciones para un nivel puede haber una o más porciones que tengan recuentos que sean significativamente diferentes que en otras porciones del conjunto (por ejemplo, picos y/o caídas). Cualquier número adecuado de recuentos normalizados o no normalizados asociados a cualquier número adecuado de porciones puede definir un nivel.

En algunas implementaciones, uno o más niveles pueden determinarse a partir de recuentos normalizados o no normalizados de todas o algunas de las porciones de un genoma. A menudo, un nivel puede determinarse a partir de todos o algunos de los recuentos normalizados o no normalizados de un cromosoma, o de un segmento del mismo. En algunas implementaciones, dos o más recuentos derivados de dos o más porciones (por ejemplo, un conjunto de porciones) determinan un nivel. En algunas implementaciones, dos o más recuentos (por ejemplo, recuentos de dos o más porciones) determinan un nivel. En algunas implementaciones, los recuentos de 2 a aproximadamente 100.000 porciones determinan un nivel. En algunas implementaciones, recuentos de 2 a aproximadamente 50.000, de 2 a aproximadamente 40.000, de 2 a aproximadamente 30.000, de 2 a aproximadamente 20.000, de 2 a aproximadamente 10.000, de 2 a aproximadamente 5000, de 2 a aproximadamente 2500, de 2 a aproximadamente 1250, de 2 a aproximadamente 1000, de 2 a aproximadamente 500, de 2 a aproximadamente 250, de 2 a aproximadamente 100 o de 2 a aproximadamente 60 porciones determinan un nivel. En algunas implementaciones, los recuentos de aproximadamente 10 a aproximadamente 50 porciones determinan un nivel. En algunas implementaciones, los recuentos de aproximadamente 20 a aproximadamente 40 o más porciones determinan un nivel. En algunas implementaciones, un nivel 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 porciones. En algunas implementaciones, un nivel corresponde a un conjunto de porciones (por ejemplo, un conjunto de porciones de un genoma de referencia, un conjunto de porciones de un cromosoma o un conjunto de porciones de un segmento de un cromosoma).

En algunas implementaciones, se determina un nivel para recuentos normalizados o no normalizados de porciones que son contiguas. En algunas implementaciones, las porciones (por ejemplo, un conjunto de porciones) que son contiguas representan segmentos vecinos de un genoma o segmentos vecinos de un cromosoma o gen. Por ejemplo, dos o más porciones contiguas, cuando se alinean uniendo las porciones extremo con extremo, pueden representar un conjunto de secuencias de una secuencia de ADN más larga que cada porción. Por ejemplo, dos o más porciones contiguas pueden representar un genoma intacto, cromosoma, gen, intrón, exón o segmento del mismo. En algunas implementaciones, un nivel se determina a partir de una colección (por ejemplo, un conjunto) de porciones contiguas y/o porciones no contiguas.

Resultado

Los métodos descritos en el presente documento pueden proporcionar una determinación de la presencia o ausencia de una variación genética (por ejemplo, aneuploidía fetal) para una muestra, proporcionando de ese modo un resultado (por ejemplo, proporcionando de ese modo un resultado determinante de la presencia o ausencia de una variación genética (por ejemplo, aneuploidía fetal)). Una variación genética incluye a menudo una ganancia, una pérdida y/o alteración (por ejemplo, duplicación, delección, fusión, inserción, mutación, reorganización, sustitución o metilación aberrante) de información genética (por ejemplo, cromosomas, segmentos de cromosomas, regiones polimórficas, regiones translocadas, secuencias de nucleótidos alteradas, similares o combinaciones de los anteriores) que dan como resultado un cambio detectable en el genoma o la información genética de un sujeto de prueba con respecto a una referencia. La presencia o ausencia de una variación genética puede determinarse transformando, analizando y/o manipulando lecturas de secuencias que hayan sido mapeadas en porciones (por ejemplo, recuentos, recuentos de porciones genómicas de un genoma de referencia). La determinación de un resultado, en algunas implementaciones, comprende analizar el ácido nucleico de una mujer embarazada. En determinadas implementaciones, se determina un resultado según los recuentos (por ejemplo, recuentos normalizados, densidades de lectura, perfiles de densidades de lectura) obtenidos de una mujer embarazada donde los recuentos son de ácido nucleico obtenido de la mujer embarazada.

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

En algunas implementaciones, la presencia o ausencia de una variación genética (por ejemplo, una aneuploidía fetal) se determina mediante un método descrito en el presente documento, mediante un método conocido en la técnica o mediante una combinación de los mismos. La presencia o ausencia de una variación genética se determina generalmente a partir de recuentos de lecturas de secuencia mapeadas en porciones de un genoma de referencia.

Algunas veces, las densidades de lectura de una referencia son para una muestra de ácido nucleico de la misma mujer embarazada de la cual se obtiene una muestra de prueba. En determinadas implementaciones, las densidades de lectura de una referencia son para una muestra de ácido nucleico de una o más mujeres embarazadas diferentes a las mujeres de las que se obtuvo una muestra de prueba. En algunas implementaciones, las densidades de lectura y/o los perfiles de densidad de lectura de un primer conjunto de porciones de un sujeto de prueba se comparan con las densidades de lectura y/o los perfiles de densidad de lectura de un segundo conjunto de porciones, donde el segundo conjunto de porciones es diferente del primer conjunto de porciones. En algunas implementaciones, las densidades de lectura y/o los perfiles de densidad de lectura de un primer conjunto de porciones de un sujeto de prueba se comparan con las densidades de lectura y/o los perfiles de densidad de lectura de un segundo conjunto de porciones, donde el segundo conjunto de porciones es del sujeto de prueba o de un sujeto de referencia que no es el sujeto de prueba. En un ejemplo no limitativo, cuando un primer conjunto de porciones se encuentra en el cromosoma 21 o en un segmento del mismo, un segundo conjunto de porciones suele encontrarse en otro cromosoma (por ejemplo, el cromosoma 1, el cromosoma 13, el cromosoma 14, el cromosoma 18, el cromosoma 19, un segmento del mismo o una combinación de los anteriores). Una referencia está ubicada a menudo en un cromosoma o segmento del mismo que es normalmente euploide. Por ejemplo, el cromosoma 1 y el cromosoma 19 son a menudo euploides en fetos debido a una alta tasa de mortalidad fetal precoz asociada con aneuploidías del cromosoma 1 y del cromosoma 19. Se puede generar y/o comparar una medida de incertidumbre entre las densidades de lectura y/o los perfiles de densidad de lectura de un sujeto de prueba y una referencia. La presencia o ausencia de una variación genética (por ejemplo, aneuploidía fetal) a veces se determina sin comparar las densidades de lectura y/o los perfiles de densidad de lectura de un ensayo sujeto a una referencia.

En determinadas implementaciones, una referencia comprende densidades de lectura y/o perfil de lectura para el mismo conjunto de porciones que para un sujeto de prueba, donde las densidades de lectura para la referencia son de una o más muestras de referencia (por ejemplo, a menudo múltiples muestras de referencia de múltiples sujetos de referencia). Una muestra de referencia es a menudo de una o más mujeres embarazadas diferentes a una mujer de la cual se obtiene una muestra de prueba.

Puede generarse una medida de incertidumbre de las densidades de lectura y/o los perfiles de lectura de un sujeto y/o referencia de prueba. En algunas implementaciones, se determina una medida de incertidumbre para las densidades de lectura y/o los perfiles de lectura de un sujeto de prueba. En algunas implementaciones, se determina una medida de incertidumbre para las densidades de lectura y/o los perfiles de lectura de un sujeto de referencia. En algunas implementaciones, se determina una medida de incertidumbre a partir de un perfil de densidad de lectura completo o un subconjunto de porciones dentro de un perfil de densidad de lectura completo.

En algunas implementaciones, las muestras de referencia son euploides para un segmento seleccionado de un genoma, y se evalúa una medida de incertidumbre entre un perfil de prueba y un perfil de referencia para el segmento

seleccionado. En algunas implementaciones, la determinación de la presencia o ausencia de una variación genética es según el número de desviaciones (por ejemplo, medidas de desviaciones, DAM) entre un perfil de prueba y un perfil de referencia para un segmento seleccionado de un genoma (por ejemplo, un cromosoma o segmento del mismo). En algunas implementaciones, la presencia de una variación genética se determina cuando el número de desviaciones entre un perfil de prueba y un perfil de referencia es superior a aproximadamente 1, superior a aproximadamente 1,5, superior a aproximadamente 2, superior a aproximadamente 2,5, superior a aproximadamente 2,6, superior a aproximadamente 2,7, superior a aproximadamente 2,8, superior a aproximadamente 2,9, superior a aproximadamente 3, superior a aproximadamente 3,1, superior a aproximadamente 3,2, superior a aproximadamente 3,3, superior a aproximadamente 3,4, superior a aproximadamente 3,5, superior a aproximadamente 4, superior a aproximadamente 5 o superior a aproximadamente 6. Por ejemplo, a veces un perfil de prueba y un perfil de referencia difieren en más de 3 medidas de desviación (por ejemplo, 3 sigma, 3 DAM) y se determina la presencia de una variación genética. En algunas implementaciones, un perfil de prueba obtenido de una mujer embarazada es superior a un perfil de referencia en más de 3 medidas de desviación (por ejemplo, 3 sigma, 3 DAM) y se determina la presencia de una aneuploidía cromosómica fetal (por ejemplo, una trisomía fetal). A menudo, una desviación mayor de tres entre un perfil de prueba y un perfil de referencia es indicativa de un sujeto de prueba no euploide (por ejemplo, presencia de una variación genética) para un segmento seleccionado de un genoma. Un perfil de prueba significativamente mayor que un perfil de referencia para un segmento seleccionado de un genoma, cuya referencia es euploide para el segmento seleccionado, a veces es determinante de una trisomía. En algunas implementaciones, un perfil de densidad de lectura obtenido de una mujer embarazada es inferior a un perfil de referencia para un segmento seleccionado, en más de 3 medidas de desviación (por ejemplo, 3 sigma, 3 DAM) y se determina la presencia de una aneuploidía cromosómica fetal (por ejemplo, una monosomía fetal). Los perfiles de prueba significativamente por debajo de un perfil de referencia, cuyo perfil de referencia es indicativo de euploidía, algunas veces son determinantes de una monosomía.

En algunas implementaciones, la ausencia de una variación genética se determina cuando el número de desviaciones entre un perfil de prueba y un perfil de referencia para un segmento seleccionado de un genoma es inferior a aproximadamente 3,5, inferior a aproximadamente 3,4, inferior a aproximadamente 3,3, inferior a aproximadamente 3,2, inferior a aproximadamente 3,1, inferior a aproximadamente 3,0, inferior a aproximadamente 2,9, inferior a aproximadamente 2,8, inferior a aproximadamente 2,7, inferior a aproximadamente 2,6, inferior a aproximadamente 2,5, inferior a aproximadamente 2,0, inferior a aproximadamente 1,5 o inferior a aproximadamente 1,0. Por ejemplo, a veces un perfil de prueba difiere de un perfil de referencia en menos de 3 medidas de desviación (por ejemplo, 3 sigma, 3 DAM) y se determina la ausencia de una variación genética. En algunas implementaciones, un perfil de prueba obtenido de una mujer embarazada difiere de un perfil de referencia en menos de 3 medidas de desviación (por ejemplo, 3 sigma, 3 DAM) y se determina la ausencia de aneuploidía cromosómica fetal (por ejemplo, euploidía fetal). En algunas implementaciones, por ejemplo, una desviación inferior a tres entre los perfiles de prueba y los perfiles de referencia (por ejemplo, 3-sigma para la desviación estándar) suele ser indicativa de un segmento de un genoma que es euploide (por ejemplo, ausencia de una variación genética). Puede representarse gráficamente y visualizarse una medida de la desviación entre los perfiles de prueba de una muestra de prueba y los perfiles de referencia de uno o más sujetos de referencia (por ejemplo, una representación gráfica de la puntuación z).

Cualquier otra referencia adecuada puede factorizarse con perfiles de prueba para determinar la presencia o ausencia de una variación genética (o determinación de euploides o no euploides) para una región de prueba (por ejemplo, un segmento de un genoma que se prueba) de una muestra de prueba. En algunas implementaciones, se puede tener en cuenta una determinación de la fracción fetal con recuentos de lecturas de secuencia (por ejemplo, densidades de lectura) para determinar la presencia o ausencia de una variación genética. Por ejemplo, las densidades de lectura y/o los perfiles de densidad de lectura pueden normalizarse según la fracción fetal antes de una comparación y/o determinar un resultado. Puede usarse un procedimiento adecuado para cuantificar la fracción fetal, los ejemplos no limitativos de los mismos incluyen un procedimiento de espectrometría de masas, procedimiento de secuenciación o combinación de los mismos.

En algunas implementaciones, una determinación de la presencia o ausencia de una variación genética (por ejemplo, una aneuploidía fetal) se determina según una zona de identificación. En determinadas implementaciones, se realiza una identificación (por ejemplo, una identificación que determina la presencia o ausencia de una variación genética, por ejemplo, un resultado) cuando un valor (por ejemplo, un perfil de densidad de lectura y/o una medida de incertidumbre) o colección de valores cae dentro de un rango predefinido (por ejemplo, una zona, una zona de identificación). En algunas implementaciones, una zona de identificación se define según una colección de valores (por ejemplo, perfiles de densidad de lectura y/o medidas de incertidumbre) que se obtienen de la misma muestra de paciente. En determinadas implementaciones, una zona de identificación se define según una colección de valores que se derivan del mismo cromosoma o segmento del mismo. En algunas implementaciones, una zona de identificación basada en una determinación de variación genética se define según una medida de incertidumbre (por ejemplo, alto nivel de confianza, por ejemplo, baja medida de incertidumbre) y/o una fracción fetal.

En algunas implementaciones, una zona de identificación se define según una determinación de variación genética y una fracción fetal de aproximadamente 2,0 % o superior, aproximadamente 2,5 % o superior, aproximadamente 3 % o superior, aproximadamente 3,25 % o superior, aproximadamente 3,5 % o superior, aproximadamente 3,75 % o superior, o aproximadamente 4,0 % o superior. Por ejemplo, en algunas implementaciones se hace una identificación

de que un feto comprende una trisomía 21 basándose en una comparación de un perfil de prueba y un perfil de referencia donde una muestra de prueba de la que se derivó el perfil de prueba comprende una determinación de fracción fetal del 2 % o superior o del 4 % o superior para una muestra de prueba obtenida de una mujer embarazada portadora de un feto. Por ejemplo, en algunas implementaciones se hace una identificación de que un feto es euploide basándose en una comparación de un perfil de prueba y un perfil de referencia donde una muestra de prueba de la que se derivó el perfil de prueba comprende una determinación de fracción fetal del 2 % o superior o del 4 % o superior para una muestra de prueba obtenida de una mujer embarazada portadora de un feto. En algunas implementaciones, una zona de identificación se define por un nivel de confianza de aproximadamente el 99 % o superior, aproximadamente el 99,1 % o superior, aproximadamente el 99,2 % o superior, aproximadamente el 99,3 % o superior, aproximadamente el 99,4 % o superior, aproximadamente el 99,5 % o superior, aproximadamente el 99,6 % o superior, aproximadamente el 99,7 % o superior, aproximadamente el 99,8 % o superior, o aproximadamente el 99,9 % o superior. En algunas implementaciones, se realiza una identificación sin utilizar una zona de identificación. En algunas implementaciones, se realiza una identificación utilizando una zona de identificación y datos o información adicionales. En algunas implementaciones, se realiza una identificación en función de una comparación sin el uso de una zona de identificación. En algunas implementaciones, se realiza una identificación basada en el examen visual de un perfil (por ejemplo, examen visual de densidades de lectura).

En algunas implementaciones, una zona de no identificación es donde no se realiza una identificación. En algunas implementaciones, una zona de no identificación se define por un valor o conjunto de valores que indican baja precisión, alto riesgo, alto error, bajo nivel de confianza, alta medida de incertidumbre, similares o una combinación de los mismos. En algunas implementaciones, una zona de no identificación se define, en parte, por una fracción fetal de aproximadamente el 5 % o inferior, aproximadamente el 4 % o inferior, aproximadamente el 3 % o inferior, aproximadamente el 2,5 % o inferior, aproximadamente el 2,0 % o inferior, aproximadamente el 1,5 % o inferior, o aproximadamente el 1,0 % o inferior.

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

Un resultado algunas veces es un fenotipo. Un resultado a veces es un fenotipo con un nivel asociado de confianza (por ejemplo, una medida de incertidumbre, por ejemplo, un feto es positivo para la trisomía 21 con un nivel de confianza del 99 %, un sujeto de prueba es negativo para un cáncer asociado con una variación genética a un nivel de confianza del 95 %). Diferentes métodos para generar valores de resultado a veces pueden producir diferentes

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

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

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

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

Tal como se mencionó anteriormente, un resultado puede caracterizarse como un verdadero positivo, verdadero negativo, falso positivo o falso negativo. La expresión “verdadero positivo”, tal como se usa en el presente documento, se refiere a un sujeto diagnosticado correctamente de una variación genética. La expresión “falso positivo”, tal como se usa en el presente documento, se refiere a un sujeto mal identificado como que tiene una variación genética. La expresión “verdadero negativo”, tal como se usa en el presente documento, se refiere a un sujeto identificado correctamente como que no tiene una variación genética. La expresión “falso negativo”, tal como se usa en el presente documento, se refiere a un sujeto mal identificado como que no tiene una variación genética. Dos medidas de rendimiento para cualquier método dado pueden calcularse basándose en las razones de estas apariciones: (i) un valor de sensibilidad, que 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.

En determinadas implementaciones, uno o más de sensibilidad, especificidad y/o nivel de confianza se expresan como un porcentaje. En algunas implementaciones, el porcentaje, independientemente para cada variable, es mayor de aproximadamente el 90 % (por ejemplo, aproximadamente el 90, 91, 92, 93, 94, 95, 96, 97, 98 o el 99 %, o mayor del 99 % (por ejemplo, aproximadamente el 99,5 %, o mayor, aproximadamente el 99,9 % o mayor, aproximadamente el 99,95 % o mayor, aproximadamente el 99,99 % o mayor)). El coeficiente de variación (CV) en algunas

implementaciones se expresa como un porcentaje, y algunas veces el porcentaje es de aproximadamente el 10 % o menos (por ejemplo, aproximadamente el 10, 9, 8, 7, 6, 5, 4, 3, 2 o el 1 %, o menos del 1 % (por ejemplo, aproximadamente el 0,5 % o menos, aproximadamente el 0,1 % o menos, aproximadamente el 0,05 % o menos, aproximadamente el 0,01 % o menos)). Una probabilidad (por ejemplo, que un resultado particular no se deba al azar) en determinadas implementaciones se expresa como una puntuación Z, un valor de p o los resultados de una prueba de la t. En algunas implementaciones, una varianza medida, intervalo de confianza, sensibilidad, especificidad y similares (por ejemplo, denominados colectivamente parámetros de confianza) para un resultado pueden generarse usando una o más manipulaciones de procesamiento de datos descritas en el presente documento. Los ejemplos específicos de generación de resultados y los niveles de confianza asociados se describen en el apartado de ejemplos y en la solicitud internacional de patente nº. PCT/US12/59123 (WO2013/052913).

El término “sensibilidad”, tal como se usa en el presente documento, se refiere al número de verdaderos positivos dividido entre el número de verdaderos positivos más el número de falsos negativos, donde la sensibilidad (sens) puede estar dentro del rango de $0 \leq \text{sens} \leq 1$. El término “especificidad”, tal como se usa en el presente documento, se refiere al número de verdaderos negativos dividido entre el número de verdaderos negativos más el número de falsos positivos, donde la sensibilidad (espec) puede estar dentro del rango de $0 \leq \text{espec} \leq 1$. En algunas implementaciones, a veces se selecciona un método que tenga una sensibilidad y una especificidad iguales a uno, o del 100 %, o cercanas a uno (por ejemplo, entre aproximadamente el 90 % y aproximadamente el 99 %). En algunas implementaciones, se selecciona un método que tiene una sensibilidad igual a 1 o el 100 % y, en determinadas implementaciones, se selecciona un método que tiene una sensibilidad próxima a 1 (por ejemplo, una sensibilidad de aproximadamente el 90 %, una sensibilidad de aproximadamente el 91 %, una sensibilidad de aproximadamente el 92 %, una sensibilidad de aproximadamente el 93 %, una sensibilidad de aproximadamente el 94 %, una sensibilidad de aproximadamente el 95 %, una sensibilidad de aproximadamente el 96 %, una sensibilidad de aproximadamente el 97 %, una sensibilidad de aproximadamente el 98 % o una sensibilidad de aproximadamente el 99 %). En algunas implementaciones se selecciona un método que tiene una especificidad equivalente a 1 o el 100 % y, en determinadas implementaciones, se selecciona un método que tiene una especificidad próxima a 1 (por ejemplo, una especificidad de aproximadamente el 90 %, una especificidad de aproximadamente el 91 %, una especificidad de aproximadamente el 92 %, una especificidad de aproximadamente el 93 %, una especificidad de aproximadamente el 94 %, una especificidad de aproximadamente el 95 %, una especificidad de aproximadamente el 96 %, una especificidad de aproximadamente el 97 %, una especificidad de aproximadamente el 98 % o una especificidad de aproximadamente el 99 %).

En algunas implementaciones, se determina la presencia o ausencia de una variación genética (por ejemplo, aneuploidía cromosómica) en un feto. En dichas implementaciones, se determina la presencia o ausencia de una variación genética fetal (por ejemplo, aneuploidía cromosómica fetal).

En determinadas implementaciones, se determina la presencia o ausencia de una variación genética (por ejemplo, aneuploidía cromosómica) para una muestra. En dichas implementaciones, se determina la presencia o ausencia de una variación genética en el ácido nucleico de la muestra (por ejemplo, aneuploidía cromosómica). En algunas implementaciones, una variación detectada o no detectada reside en la muestra de ácido nucleico de una fuente, pero no en la muestra de ácido nucleico de otra fuente. Los ejemplos no limitativos de fuentes incluyen ácido nucleico placentario, ácido nucleico fetal, ácido nucleico materno, ácido nucleico de células cancerosas, ácido nucleico de células no cancerosas, similares y combinaciones de los mismos. En ejemplos no limitativos, una variación genética particular detectada o no detectada (i) reside en el ácido nucleico placentario pero no en el ácido nucleico fetal y no en el ácido nucleico materno; (ii) reside en el ácido nucleico fetal, pero no en el ácido nucleico materno; o (iii) reside en el ácido nucleico materno, pero no en el ácido nucleico fetal.

La presencia o ausencia de una variación genética y/o afección médica asociada (por ejemplo, un resultado) a menudo se proporciona por un módulo de resultados. La presencia o ausencia de una variación genética (por ejemplo, una aneuploidía, una aneuploidía fetal, una variación del número de copias) es, en algunas implementaciones, identificada por un módulo de resultados o por una máquina que comprende un módulo de resultados. Un módulo de resultados puede especializarse para determinar una variación genética específica (por ejemplo, una trisomía, una trisomía 21, una trisomía 18). Por ejemplo, un módulo de resultados que identifica una trisomía 21 puede ser diferente y/o diferente de un módulo de resultados que identifica una trisomía 18. En algunas implementaciones, se requiere un módulo de resultados o una máquina 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). En determinadas implementaciones, un resultado se transfiere desde un módulo de resultados a un módulo de visualización donde el módulo de visualización proporciona un resultado.

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). Normalmente, se proporciona un resultado a un profesional de atención sanitaria (por ejemplo, técnico o gerente de laboratorio; médico o asistente). En determinadas implementaciones, se proporciona un resultado en un medio visual adecuado (por ejemplo, un periférico o componente de una máquina, por ejemplo, una impresora o pantalla). En algunas implementaciones, un determinante del resultado de la presencia o ausencia de una variación genética se proporciona a un profesional

sanitario en forma de un informe, y en determinadas implementaciones el informe comprende una visualización de un valor de resultado y un parámetro de confianza asociado. Generalmente, un resultado puede mostrarse en un formato adecuado que facilita la determinación de la presencia o ausencia de una variación genética y/o afección médica. Los ejemplos no limitativos de formatos adecuados para usar para informar y/o visualizar conjuntos de datos o para informar sobre un resultado incluyen datos digitales, un gráfico, un gráfico 2D, un gráfico 3D y un gráfico 4D, una imagen (por ejemplo, un jpg, mapa de bits (por ejemplo, bmp), pdf, tiff, gif, raw, png, similares o un formato adecuado) un pictograma, un gráfico, 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 de lectura de secuencia de ácido nucleico, o similares, en una representación de un ácido nucleico celular de un sujeto, en determinadas implementaciones. Por ejemplo, el análisis de lecturas de secuencia de ácido nucleico de un sujeto y la generación de un resultado y/o perfil cromosómico puede considerarse una transformación de fragmentos de lectura de secuencia relativamente pequeños en una representación de una estructura cromosómica relativamente grande. En algunas implementaciones, un resultado procede de una transformación de lecturas de secuencia de un sujeto (por ejemplo, una mujer embarazada), en una representación de una estructura existente (por ejemplo, un genoma, un cromosoma o segmento del mismo) presente en el sujeto (por ejemplo, un ácido nucleico materno y/o fetal). En algunas implementaciones, un resultado comprende una transformación de lecturas de secuencia de un primer sujeto (por ejemplo, una mujer embarazada), en una representación compuesta de estructuras (por ejemplo, un genoma, un cromosoma o segmento del mismo), y una segunda transformación de la representación compuesta que produce una representación de una estructura presente en un primer sujeto (por ejemplo, una mujer embarazada) y/o un segundo sujeto (por ejemplo, un feto).

Resultados relacionados con los cromosomas sexuales

En algunas implementaciones, un resultado se refiere a una variación genética de un cromosoma sexual. Se describen variaciones genéticas de los cromosomas sexuales, por ejemplo, en la publicación de solicitud de patente internacional n.º WO 2013/192562. En algunas implementaciones, un resultado es una determinación del cariotipo del cromosoma sexual, la detección de una aneuploidía cromosómica sexual y/o la determinación del sexo fetal. Algunas afecciones de aneuploidía cromosómica sexual (SCA) incluyen, pero no se limitan a, síndrome de Turner [45,X], trisomía X [47,XXX], síndrome de Klinefelter [47,XXY] y síndrome [47,XYY] (a veces denominado síndrome Jacobs).

Las evaluaciones de las variaciones cromosómicas sexuales, en algunas implementaciones, se basan en una segregación de las transformaciones de recuento de lecturas de secuencia para el cromosoma X y el cromosoma Y. Las transformaciones de recuento de lecturas de secuencia pueden incluir, por ejemplo, representaciones de cromosoma X y representaciones de cromosoma Y y/o puntuaciones Z basadas en tales representaciones. Una representación bidimensional de las transformaciones de recuento de lectura de secuencias de nucleótidos (por ejemplo, puntuaciones Z basadas en recuentos de lectura normalizados por PERUN o recuentos de lectura normalizados de componentes principales) para el cromosoma X frente al cromosoma Y para un grupo de muestras que tienen diversos cariotipos (por ejemplo, XX, XY, XXX, X, XXY, XYY) genera un campo plano de puntos de gráfico que pueden ser tallados en regiones, cada uno específico para un cariotipo particular. La determinación de un cariotipo del cromosoma sexual, por ejemplo, para una muestra dada puede lograrse determinando en qué región del campo plano el punto de gráfico para esa muestra cae.

Ciertos métodos descritos en el presente documento pueden ser útiles para generar gráficos que tienen regiones bien definidas (por ejemplo, con límites afilados, alta resolución) para variaciones de cariotipo particulares. Los métodos que pueden ayudar a generar gráficos de alta resolución incluyen la normalización de recuento de lectura de secuencia, la selección de porciones informativas (es decir, “de no identificación”) y la normalización adicional de los niveles del cromosoma X y del cromosoma Y. La normalización de las lecturas de secuencia y la normalización adicional de los niveles se describe en el presente documento y puede incluir la normalización PERUN y/o la normalización de los componentes principales, por ejemplo, de las lecturas de secuencia mapeadas en el cromosoma X y/o el cromosoma Y y/o los niveles (por ejemplo, representaciones cromosómicas) para el cromosoma X y/o Y. La selección de porciones informativas para el cromosoma X y el cromosoma Y se describe, por ejemplo, en la publicación de solicitud de patente internacional n.º WO 2013/192562, Y puede incluir, por ejemplo, evaluación de parámetros de filtrado tales como parámetros de validación cruzada, capacidad de mapeo, repetibilidad y/o separación masculina frente a hembra.

Uso de resultados

Un profesional sanitario, u otro individuo cualificado, que recibe un informe que comprende uno o más resultados determinantes de la presencia o ausencia de una variación genética, puede usar los datos visualizados en el informe para realizar una identificación en relación con el estado del paciente o sujeto de prueba. El profesional sanitario puede realizar una recomendación basada en el resultado proporcionado, en algunas implementaciones. Un profesional sanitario o individuo cualificado puede proporcionar a un paciente o sujeto de prueba una identificación o puntuación

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

La expresión “proporcionar un resultado” y equivalentes gramaticales de la misma, tal como se usa en el presente documento puede referirse además a un método para obtener tal información incluyendo, sin limitación, obtener la información de un laboratorio (por ejemplo, un archivo de laboratorio). Un archivo de laboratorio puede generarse por un laboratorio que lleva a cabo uno o más ensayos o una o más etapas de procesamiento de datos para determinar la presencia o ausencia de la afección médica. El laboratorio puede estar en la misma ubicación o en una ubicación diferente (por ejemplo, en otro país) que el personal que identifica la presencia o ausencia de la afección 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 determinadas implementaciones.

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

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

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

Variaciones genéticas y afecciones médicas

La presencia o ausencia de una varianza genética puede determinarse usando un método, una máquina o un aparato descrito en el presente documento. En determinadas implementaciones, la presencia o ausencia de una o más variaciones genéticas se determina según un resultado proporcionado mediante los métodos, máquinas y aparatos descritos en el presente documento. Una variación genética es generalmente un fenotipo genético particular presente en determinados individuos y a menudo una variación genética está presente en una subpoblación estadísticamente significativa de individuos. En algunas implementaciones, una variación genética es una anomalía cromosómica (por ejemplo, aneuploidía, duplicación de uno o más cromosomas, pérdida de uno o más cromosomas), anomalía cromosómica parcial o mosaicismo (por ejemplo, pérdida o ganancia de uno o más segmentos de un cromosoma), translocaciones, inversiones, cada una de las cuales se describe con mayor detalle en el presente documento. Los ejemplos no limitativos de variaciones genéticas incluyen una o más deleciones (por ejemplo, microdeleciones),

duplicaciones (por ejemplo, microduplicaciones), inserciones, mutaciones, polimorfismos (por ejemplo, polimorfismos de un solo nucleótido), fusiones, repeticiones (por ejemplo, repeticiones cortas en tándem), distintos sitios de metilación, distintos patrones de metilación, similares y combinaciones de los mismos. Una inserción, repetición, delección, duplicación, mutación o polimorfismo puede ser de cualquier longitud, y en algunas implementaciones, tiene aproximadamente 1 base o par de bases (pb) a aproximadamente 250 megabases (Mb) de longitud. En algunas implementaciones, una inserción, repetición, delección, duplicación, mutación o polimorfismo tiene de aproximadamente 1 base o par de bases (pb) a aproximadamente 50.000 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, 1000 kb, 5000 kb o 10.000 kb de longitud).

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

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

Una variación genética es, algunas veces, una inserción. Una inserción es, algunas veces, la adición de uno o más pares de bases de nucleótidos en una secuencia de ácido nucleico. Una inserción es, algunas veces, una microinserción. En determinadas implementaciones, una inserción comprende la adición de un segmento de un cromosoma en un genoma, cromosoma o segmento de los mismos. En determinadas implementaciones, una inserción comprende la adición de un alelo, un gen, un intrón, un exón, cualquier región no codificante, cualquier región codificante, segmento de los mismos o combinación de los mismos en un genoma o segmento de los mismos. En determinadas implementaciones, una inserción comprende la adición (por ejemplo, inserción) de ácido nucleico de origen desconocido en un genoma, cromosoma o segmento de los mismos. En determinadas implementaciones, una inserción comprende la adición (por ejemplo, inserción) de una sola base.

Tal como se usa en el presente documento, una “variación del número de copias” es generalmente una clase o un tipo de variación genética o aberración cromosómica. Una variación del número de copias puede ser una delección (por ejemplo, microdelección), duplicación (por ejemplo, una microduplicación) o inserción (por ejemplo, una microinserción). A menudo, el prefijo “micro”, tal como se usa en el presente documento, algunas veces es un segmento de ácido nucleico con una longitud menor de 5 Mb. Una variación del número de copias puede incluir una o más delecciones (por ejemplo, microdelección), duplicaciones y/o inserciones (por ejemplo, una microduplicación, microinserción) de un segmento de un cromosoma. En determinadas implementaciones, una duplicación comprende una inserción. En determinadas implementaciones, una inserción es una duplicación. En determinadas implementaciones, una inserción no es una duplicación.

En algunas implementaciones, una variación del número de copias es una variación del número de copias fetal. A menudo, una variación del número de copias fetal es una variación del número de copias en el genoma de un feto. En algunas implementaciones, una variación del número de copias es una variación del número de copias materno y/o fetal. En algunas implementaciones, una variación del número de copias materno y/o fetal es una variación del número de copias dentro del genoma de una mujer embarazada (por ejemplo, una mujer que porta un feto), una mujer que da a luz o una mujer capaz de portar un feto. Una variación del número de copias puede ser una variación del número de copias heterocigotas en la que la variación (por ejemplo, una duplicación o delección) está presente en un alelo de un genoma. Una variación del número de copias puede ser una variación del número de copias homocigotas en la que la variación está presente en ambos alelos de un genoma. En algunas implementaciones, una variación del número de copias es una variación del número de copias heterocigotas u homocigotas fetal. En algunas implementaciones, una variación del número de copias es una variación del número de copias heterocigotas u homocigotas materno y/o fetal.

Algunas veces, una variación del número de copias está presente en un genoma materno y un genoma fetal, un genoma materno y no en un genoma fetal, o un genoma fetal y no en un genoma materno.

“Ploidía” es una referencia al número de cromosomas presentes en un feto o su madre. En determinadas implementaciones, la “ploidía” es lo mismo que “ploidía cromosómica”. En seres humanos, por ejemplo, los cromosomas autosómicos están presentes a menudo en pares. Por ejemplo, en ausencia de una variación genética, la mayoría de los seres humanos tienen dos de cada cromosoma autosómico (por ejemplo, cromosomas 1-22). La presencia del complemento normal de 2 cromosomas autosómicos en un ser humano se denomina a menudo euploide o diploide. “Microploidía” tiene un significado similar a ploidía. “Microploidía” se suele referir a la ploidía de un segmento de un cromosoma. El término “microploidía” a veces es una referencia a la presencia o ausencia de una variación del número de copias (por ejemplo, una delección, duplicación y/o una inserción) dentro de un cromosoma (por ejemplo, una delección, duplicación o inserción homocigota o heterocigota, similares o ausencia de las mismas).

En algunas implementaciones, la microploidía de un feto coincide con la microploidía de la madre del feto (por ejemplo, el sujeto femenino gestante). En determinadas implementaciones, la microploidía de un feto coincide con la microploidía de la madre del feto y tanto la madre como el feto portan la misma variación del número de copias heterocigotas, la variación del número de copias homocigotas o ambos son euploides. En determinadas implementaciones, la microploidía de un feto es diferente de la microploidía de la madre del feto. Por ejemplo, a veces la microploidía de un feto es heterocigota para una variación del número de copias, la madre es homocigota para una variación del número de copias y la microploidía del feto no coincide (por ejemplo, no es igual) con la microploidía de la madre para la variación del número de copias especificada.

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

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

Sexo del feto

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

En determinados casos, puede ser beneficioso determinar el sexo de un feto en el útero. Por ejemplo, un paciente (por ejemplo, mujer embarazada) con antecedentes familiares de uno o más trastornos ligados al sexo puede desear determinar el sexo del feto que porta para ayudar a evaluar el riesgo de que el feto herede tal trastorno. Los trastornos ligados al sexo incluyen, sin limitación, trastornos ligados al cromosoma X e Y. Los trastornos ligados al cromosoma X incluyen trastornos recesivos ligados al cromosoma X y dominantes ligados al cromosoma X. Los ejemplos de trastornos recesivos ligados al cromosoma X incluyen, sin limitación, trastornos inmunitarios (por ejemplo, enfermedad granulomatosa crónica (CYBB), síndrome de Wiskott-Aldrich, inmunodeficiencia combinada severa ligada al cromosoma X, agammaglobulinemia ligada al cromosoma X, síndrome de hiper-IgM tipo 1, IPEX, enfermedad linfoproliferativa ligada al cromosoma X, deficiencia de properdina), trastornos hematológicos (por ejemplo, hemofilia A, hemofilia B, anemia sideroblástica ligada al cromosoma X), trastornos endocrinos (por ejemplo, síndrome de insensibilidad a andrógenos/enfermedad de Kennedy, síndrome de Kallmann KAL1, hipoplasia suprarrenal congénita ligada al cromosoma X), trastornos metabólicos (por ejemplo, deficiencia de ornitina transcarbamilasa, síndrome oculocerebrorenal, adrenoleucodistrofia, deficiencia de glucosa-6-fosfato deshidrogenasa, deficiencia de piruvato deshidrogenasa, enfermedad de Danon/glucogenosis tipo lib, enfermedad de Fabry, síndrome de Hunter, síndrome de Lesch-Nyhan, enfermedad de Menkes/síndrome de asta occipital), trastornos del sistema nervioso (por ejemplo, síndrome de Coffin-Lowry, síndrome MASA, síndrome de alfa talasemia con retraso mental ligado al cromosoma X, síndrome de Siderius con retraso mental ligado al cromosoma X, daltonismo, albinismo ocular, enfermedad de Norrie, coroideremia, enfermedad de Charcot-Marie-Tooth (CMTX2-3), enfermedad de Pelizaeus-Merzbacher, SMAX2),

trastornos de la piel y tejidos relacionados (por ejemplo, disqueratosis congénita, displasia ectodérmica hipohidrótica (DEH), ictiosis ligada al cromosoma X, distrofia corneal endotelial ligada al cromosoma X), trastornos neuromusculares (por ejemplo, distrofia muscular de Becker/Duchenne, miopatía centronuclear (MTM1), síndrome de Conradi-Hünemann, distrofia muscular de Emery-Dreifuss 1), trastornos urológicos (por ejemplo, síndrome de Alport, enfermedad de Dent, diabetes insípida nefrótica ligada al cromosoma X), trastornos óseos/dentales (por ejemplo, AMELX amelogénesis imperfecta), y otros trastornos (por ejemplo, síndrome de Barth, síndrome de McLeod, síndrome de Smith-Fineman-Myers, síndrome de Simpson-Golabi-Behmel, síndrome de Mohr-Tranebjærg, síndrome nasodigitoacústico). Los ejemplos de trastornos dominantes ligados al cromosoma X incluyen, sin limitación, hipofosfatemia ligada al cromosoma X, hipoplasia dérmica focal, síndrome del cromosoma X frágil, síndrome de Aicardi, incontinencia pigmentaria, síndrome de Rett, síndrome CHILD, síndrome de Lujan-Fryns y síndrome bucofaciodigital 1. Los ejemplos de trastornos ligados al cromosoma Y incluyen, sin limitación, esterilidad masculina, retinitis pigmentosa y azoospermia.

Anomalías cromosómicas

En algunas implementaciones, la presencia o ausencia de una anomalía cromosómica fetal puede determinarse usando un método, una máquina o un aparato descrito en el presente documento. Las anomalías cromosómicas incluyen, sin limitación, una ganancia o pérdida de un cromosoma completo o una región de un cromosoma que comprende uno o más genes. Las anomalías cromosómicas incluyen monosomías, trisomías, polisomías, pérdida de heterocigosidad, translocaciones, 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. La expresión “anomalía cromosómica” o “aneuploidía”, como se usa en el presente documento, se refiere a una desviación entre la estructura del cromosoma en cuestión y un cromosoma homólogo normal. El término “normal” se refiere al cariotipo predominante o patrón de bandas que se encuentra en individuos sanos de una especie particular, por ejemplo, un genoma euploide (por ejemplo, diploide en humanos, por ejemplo, 46,XX o 46,XY). Dado que los diferentes organismos tienen complementos cromosómicos ampliamente variables, el término “aneuploidía” no se refiere a un número particular de cromosomas, sino más bien a la situación en la que el contenido cromosómico dentro de una célula o células dadas de un organismo es anómalo. En algunas implementaciones, el término “aneuploidía” en el presente documento se refiere a un desequilibrio de material genético provocado por una pérdida o ganancia de un cromosoma completo o parte de un cromosoma. Un “aneuploidía” puede referirse a una o más deleciones y/o inserciones de un segmento de un cromosoma. El término “euploide”, en algunas implementaciones, se refiere a un complemento normal de cromosomas.

El término “monosomía”, tal como se usa en el presente documento, se refiere a la falta de un cromosoma del complemento normal. La monosomía parcial puede producirse en translocaciones o deleciones desequilibradas, en las cuales solo un segmento del cromosoma está presente en una sola copia. La monosomía de cromosomas sexuales (45, X) provoca el síndrome de Turner, por ejemplo. El término “disomía” se refiere a la presencia de dos copias de un cromosoma. Para organismos tales como seres humanos que tienen dos copias de cada cromosoma (aquellos que son diploides o “euploides”), la disomía es la condición normal. Para organismos que normalmente tienen tres o más copias de cada cromosoma (aquellos que son triploides o superiores), la disomía es un estado cromosómico aneuploide. En la disomía uniparental, ambas copias de un cromosoma provienen del mismo progenitor (sin contribución del otro progenitor).

El término “trisomía”, tal como se usa en el presente documento, se refiere a la presencia de tres copias, en lugar de dos copias, de un cromosoma particular. La presencia de un cromosoma 21 extra, que se encuentra en el síndrome de Down humano, se denomina “trisomía 21”. La trisomía 18 y la trisomía 13 son otras dos trisomías autosómicas humanas. La trisomía de los cromosomas sexuales puede observarse en mujeres (por ejemplo, 47, XXX en el síndrome triple X) o en hombres (por ejemplo, 47, XXY en el síndrome de Klinefelter; o 47, XYY en el síndrome de Jacobs). En algunas implementaciones, una trisomía es una duplicación de la mayor parte o la totalidad de un autosoma. En determinadas implementaciones, una trisomía es una aneuploidía cromosómica completa que produce tres casos (por ejemplo, tres copias) de un tipo particular de cromosoma (por ejemplo, en lugar de dos instancias (por ejemplo, un par) de un tipo particular de cromosoma para un euploide).

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

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.

Las expresiones “monosomía parcial” y “trisomía parcial”, tal como se usan en el presente documento, se refieren a un desequilibrio de material genético provocado por la pérdida o ganancia de parte de un cromosoma. Una monosomía parcial o trisomía parcial puede resultar de una translocación desequilibrada, en 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 el segmento que existe en el cromosoma derivado) y solo una copia de parte del otro cromosoma implicado en el cromosoma derivado.

El término “mosaicismo”, tal como se usa en el presente documento, se refiere a aneuploidía en algunas células, pero no en todas las células, de un organismo. Determinadas anomalías cromosómicas pueden existir como anomalías cromosómicas con mosaicismo y sin mosaicismo. Por ejemplo, determinados individuos con trisomía 21 tienen síndrome de Down y algunos tienen síndrome de Down sin mosaicismo. Diferentes mecanismos pueden conducir al mosaicismo. Por ejemplo, (i) un cigoto inicial puede tener tres cromosomas 21, lo que normalmente daría como resultado 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.

Las tablas 1A y 1B presentan una lista no limitativa de afecciones, síndromes y/o anomalías cromosómicas que pueden identificarse potencialmente mediante los métodos, máquinas y/o un aparato descritos en el presente documento. La tabla 1B es de la base de datos DECIPHER del 6 de octubre de 2011 (por ejemplo, versión 5.1, basada en posiciones mapeadas en GRCh37; disponible en el localizador uniforme de recursos (URL, por sus siglas en inglés) decipher.sanger.ac.uk).

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) 1 p	neuroblastoma
2	monosomía-trisomía 2q	retraso del crecimiento, retraso del desarrollo y mental, y anomalías físicas menores
3	monosomía-trisomía (somática)	Linfoma no hodgkiniano
4	monosomía-trisomía (somática)	Leucemia no linfocítica aguda (LNLA)
5	5p	Maullido de gato; síndrome de Lejeune
5	monosomía-trisomía (somática) 5q	síndrome mielodisplásico

Cromosoma	Anomalia	Asociación con la enfermedad
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)	ALL o ANLL
11	11 p-	Aniridia; tumor de Wilms
11	11q-	Síndrome de Jacobsen
11	monosomía (somática)-trisomía	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	trisomía	Trisomía 16 completa Trisomía 16 de mosaico
16	deleción de 16q13.3	Rubenstein-Taybi
3	monosomía-trisomía (somática)	carcinomas de células renales papilares (malignos)
17	17p-(somático)	Síndrome p17 en neoplasias malignas mieloides
17	deleción de 17q 11.2	Smith-Magenis
17	17q13.3	Miller-Dieker
17	monosomía-trisomía (somática)	adenomas corticales renales
17	trisomía 17p11.2-12	Síndrome de Charcot-Marie-Tooth tipo 1; HNPP
18	18p-	Síndrome de monosomía parcial 18p o síndrome de Grouchy-Lamy-Thieffry

Cromosoma	Anomalia	Asociación con la enfermedad
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 conotruncales 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	
microdeleción recurrente de 16p13.11 (locus de susceptibilidad a trastornos neurocognitivos)	16	15.504.454	16.284.248	0,78	
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
locus de susceptibilidad 1q21.1 para la trombocitopenia-síndrome de radio ausente (TAR)	1	145.401.253	145.928.123	0,53	3

Síndrome	Cromosoma	Inicio	Fin	Intervalo (Mb)	Grado
síndrome de delección 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 delección distal de 22q11.2	22	22.115.848	23.696.229	1,58	
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	
síndrome de delección de 2q33.1	2	196.925.089	205.206.940	8,28	1
monosomía de 2q37	2	239.954.693	243.102.476	3,15	1
síndrome 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	
Predisposición genética a las parálisis por presión (HNPP)	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	Cromosoma	Inicio	Fin	Intervalo (Mb)	Grado
Síndrome de Prader-Willi (Tipo 1)	15	22.876.632	28.557.186	5,68	1
Síndrome de Prader-Willi (tipo 2)	15	23.758.390	28.557.186	4,80	1
QRD (quistes renales y diabetes)	17	34.907.366	36.076.803	1,17	
Síndrome de Rubenstein-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 1 de mano/pie dividido (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 algunas implementaciones, la presencia o ausencia de preeclampsia se determina usando un método, una máquina o un aparato descrito en el presente documento. La preeclampsia es una afección en la que surge hipertensión durante el embarazo (por ejemplo, hipertensión inducida por el embarazo) y se asocia con cantidades significativas de proteína en la orina. En determinadas implementaciones, 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 (floc) 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 ADNfloc antes del inicio de los síntomas puede deberse a hipoxia/reoxigenación dentro del espacio intervilloso que conduce al estrés oxidativo tisular y al aumento de la apoptosis y necrosis

placentaria. Además de la evidencia de mayor descarga de ADNfnc en la circulación materna, también existe evidencia de menor aclaramiento renal de ADNfnc en la preeclampsia. Dado que la cantidad de ADN fetal se determina actualmente mediante la cuantificación de secuencias específicas del cromosoma Y, los enfoques alternativos, tales como la medición del ADN libre de células total o el uso de marcadores epigenéticos fetales independientes del sexo, tales como la metilación del ADN, ofrecen una alternativa. EL ARN libre de células de origen placentario es otro biomarcador alternativo que puede usarse para analizar y diagnosticar preeclampsia en la práctica clínica. EL ARN fetal se asocia con partículas placentarias subcelulares que lo protegen frente a la degradación. Los niveles de ARN fetal a veces son diez veces mayores en mujeres embarazadas con preeclampsia en comparación con los controles y, por tanto, es un biomarcador alternativo que puede usarse para analizar y diagnosticar preeclampsia en la práctica clínica.

Patógenos

En algunas implementaciones, la presencia o ausencia de una afección patógena se determina mediante un método, una máquina o un 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, las máquinas y los aparatos proporcionados en el presente documento pueden usarse para determinar la presencia o ausencia de un patógeno. A menudo, los patógenos poseen ácido nucleico con características únicas para un patógeno particular, tal como, por ejemplo, estado epigenético y/o una o más variaciones, duplicaciones y/o deleciones de secuencias. Por tanto, los métodos proporcionados en el presente documento pueden usarse para identificar un patógeno particular o variante de patógenos (por ejemplo, cepa).

Cánceres

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

Se puede usar software para realizar una o más etapas en los procesos descritos 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, como se describe con mayor detalle de aquí en adelante.

Máquinas, software e interfaces

Ciertos procesos y métodos descritos en el presente documento a menudo no pueden realizarse sin un ordenador, procesador, software, módulo u otro aparato. Los métodos descritos en el presente documento normalmente son métodos implementados por ordenador, y una o más porciones de un método a veces las realizan uno o más procesadores (por ejemplo, microprocesadores), ordenadores o aparatos controlados por microprocesadores. En algunas implementaciones uno o más o todos los métodos de procesamiento conocidos o descritos en el presente documento (por ejemplo, mapeo, compresión de datos, determinaciones de estimación del sesgo local del genoma, determinaciones de relación, comparaciones de relaciones, normalización de recuento, densidad de lectura y/o generaciones de perfiles de densidad de lectura, PCA, ajustes de perfil, filtrado de porciones, ponderación de porción, comparaciones de perfiles, puntuación de perfil, determinación de un resultado, similares o combinaciones de los mismos) se realizan por un procesador, un microprocesador, un ordenador, junto con memoria y/o por un aparato controlado por microprocesador. Las implementaciones relacionadas con los métodos descritos en este documento son aplicables generalmente al mismo procedimiento o a procedimientos relacionados implementados por instrucciones en los sistemas, aparatos y productos de programa informático descritos en el presente documento. En algunas implementaciones, los procedimientos y métodos descritos en el presente documento (por ejemplo, cuantificación, recuento y/o determinación de lecturas de secuencia, recuentos, niveles y/o perfiles de secuencia) se realizan mediante métodos automatizados. En algunas implementaciones, una o más etapas y un método descrito en el presente documento se llevan a cabo mediante un procesador y/o un ordenador, y/o se llevan a cabo junto con la

memoria. En algunas implementaciones, un método automatizado se incorpora en software, módulos, procesadores, periféricos y/o una máquina que comprende similares, que determinan lecturas de secuencia, recuentos, mapeo, etiquetas de secuencia mapeadas, niveles, perfiles, normalizaciones, comparaciones, establecimiento de rango, categorización, ajustes, representación gráfica, resultados, transformaciones e identificaciones. Tal como se usa en el presente documento, software se refiere a instrucciones de programas legibles por ordenador que, cuando se ejecutan por un procesador, realizan operaciones informáticas, tal como se describe en el presente documento.

Las lecturas de secuencia, recuentos, densidades de lectura y perfiles de densidad de lectura derivados de un sujeto de prueba (por ejemplo, un paciente, una mujer embarazada) y/o de un sujeto de referencia pueden analizarse y procesarse adicionalmente para determinar la presencia o ausencia de una variación genética. Las lecturas de secuencias, recuentos, niveles y/o perfiles se denominan, algunas veces, “datos” o “conjuntos de datos”. En algunas implementaciones, los datos o conjuntos de datos pueden caracterizarse por una o más características o variables (por ejemplo, basadas en secuencia [por ejemplo, contenido de GC, secuencia de nucleótidos específica, similares], específicas de función [por ejemplo, genes expresados, genes de cáncer, similares], basadas en la ubicación [específicas del genoma, específicas del cromosoma, porción o específicas de una porción], similares y combinaciones de los mismos). En determinadas implementaciones, los datos o conjuntos de datos pueden organizarse en una matriz que tiene dos o más dimensiones basándose en una o más características o variables. Los datos organizados en matrices pueden organizarse usando cualquier característica o variable adecuada. Un ejemplo no limitativo de datos en una matriz incluye datos organizados por edad materna, ploidía materna y contribución fetal. En determinadas implementaciones, los conjuntos de datos caracterizados por una o más características o variables a veces se procesan después del recuento.

Los aparatos (múltiples aparatos, también denominados en el presente documento en plural como aparatos), software e interfaces pueden usarse para realizar 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 de variación estadística, comparaciones, etapas iterativas, algoritmos de validación y representaciones gráficas, por ejemplo. En algunas implementaciones, un usuario puede introducir un conjunto de datos como información de entrada, un usuario puede descargar uno o más conjuntos de datos mediante un medio de *hardware* adecuado (por ejemplo, unidad flash) y/o un usuario puede enviar un conjunto de datos de un sistema a otro para el procesamiento posterior y/o proporcionar un resultado (por ejemplo, enviar datos de lectura de secuencia de un secuenciador a un sistema informático para el mapeo de lecturas de secuencia; enviar datos de secuencia mapeados en un sistema informático para procesar y producir un resultado y/o informe).

Típicamente, un sistema comprende uno o más aparatos. En algunas implementaciones, un aparato es una máquina. En algunas implementaciones, un aparato comprende una máquina. Un aparato puede comprender uno o más de memoria, uno o más procesadores y/o instrucciones. Cuando un sistema incluye dos o más aparatos, algunos o todos los aparatos pueden estar ubicados en la misma ubicación, algunos o todos los aparatos pueden estar ubicados en ubicaciones diferentes, todos los aparatos pueden estar ubicados en una ubicación y/o todos los aparatos pueden estar ubicados en ubicaciones diferentes. Cuando un sistema incluye dos o más aparatos, algunos o todos los aparatos pueden estar ubicados en la misma ubicación que un usuario, algunos o todos los aparatos pueden estar ubicados en una ubicación distinta a un usuario, todos los aparatos pueden estar ubicados en la misma ubicación que el usuario y/o todos los aparatos pueden estar ubicados en una o más ubicaciones diferentes al usuario. Los aparatos de un sistema descrito en el presente documento pueden interactuar con uno o más servidores informáticos remotos y/u ordenadores (por ejemplo, una nube, un servicio informático en la nube) mediante un método adecuado. El término “nube”, como se usa en el presente documento, se refiere, en parte, a dos o más ordenadores (por ejemplo, a menudo una pluralidad de ordenadores) conectados a través de una red de comunicación en tiempo real (por ejemplo, una internet) que puede realizar una función centralizada (por ejemplo, un método descrito en el presente documento) donde porciones de la función son compartidas por una pluralidad de ordenadores en una red. Una “nube” a menudo puede ejecutar uno o más programas (por ejemplo, programas de software, módulos) en una pluralidad de ordenadores conectados al mismo tiempo. En algunas implementaciones, un sistema y/o un aparato descrito en el presente documento comprende una nube (por ejemplo, un servidor en la nube, un ordenador en la nube, un servicio informático en la nube). Una o más funciones de un sistema y/o un aparato descrito en el presente documento pueden realizarse por una nube. Los datos y/o información se pueden transferir a, y/o desde un aparato y una nube usando un método adecuado. El término “ordenador”, como se usa en el presente documento, se refiere a un dispositivo eléctrico, artificial que comprende un microprocesador que puede realizar operaciones aritméticas y lógicas. Un ordenador a veces comprende instrucciones, software (por ejemplo, módulos), memoria, una pantalla, uno o más periféricos y/o un medio de almacenamiento. En algunas implementaciones, una máquina comprende un ordenador. En algunas implementaciones, una máquina es un ordenador. Un ordenador a menudo interactúa y/o está conectado a otros ordenadores (por ejemplo, una internet, una red, una nube).

Un sistema a veces comprende un aparato informático o un aparato de secuenciación, o un aparato informático y un aparato de secuenciación (es decir, una máquina de secuenciación y/o una máquina informática). Un aparato de secuenciación generalmente se configura para recibir ácido nucleico físico y generar señales correspondientes a bases de nucleótidos del ácido nucleico. Un aparato de secuenciación a menudo se “carga” con una muestra que comprende

ácido nucleico y el ácido nucleico de la muestra cargada en el aparato de secuenciación generalmente se somete a un proceso de secuenciación de ácido nucleico. El término “cargar un aparato de secuenciación” como se usa en el presente documento se refiere a poner en contacto una porción de un aparato de secuenciación (por ejemplo, una celda de flujo) con una muestra de ácido nucleico, cuya porción del aparato de secuenciación se configura para recibir una muestra para realizar un proceso de secuenciación de ácido nucleico. En algunas implementaciones, un aparato de secuenciación se carga con una variante de un ácido nucleico de muestra. Una variante a veces se produce mediante un proceso que modifica el ácido nucleico de muestra a una forma adecuada para secuenciar el ácido nucleico (por ejemplo, mediante ligación (por ejemplo, añadir adaptadores a los extremos del ácido nucleico de muestra mediante ligación), amplificación, digestión de restricción, similares o combinaciones de los mismos). Un aparato de secuenciación a menudo se configura, en parte, para realizar un método de secuenciación de ADN adecuado que genera señales (por ejemplo, señales electrónicas, señales de detector, imágenes, similares, o combinaciones de las mismas) correspondientes a bases de nucleótidos del ácido nucleico cargado.

Una o más señales correspondientes a cada base de una secuencia de ADN a menudo se procesan y/o transforman en identificaciones base (por ejemplo, una base de nucleótidos específica, por ejemplo, guanina, citosina, timina, uracilo, adenina y similares) mediante un proceso adecuado. Una colección de identificaciones base derivadas de un ácido nucleico cargado a menudo se procesan y/o ensamblan en una o más lecturas de secuencia. En implementaciones en las que múltiples ácidos nucleicos de muestra se secuencian a la vez (es decir, multiplexación), se puede utilizar un proceso de demultiplexación adecuado para lecturas particulares asociadas con el ácido nucleico de muestra del que se originaron. Las lecturas de secuencia pueden alinearse mediante un proceso adecuado a un genoma de referencia y pueden contarse las lecturas alineadas con porciones del genoma de referencia, como se describe en el presente documento.

Un aparato de secuenciación a veces está asociado con y/o comprende uno o más aparatos informáticos en un sistema. El uno o más aparatos informáticos a veces están configurados para realizar uno o más de los siguientes procesos: generar identificaciones base a partir de señales de aparato de secuenciación, lecturas de ensamblaje (por ejemplo, lecturas de generación), lecturas de multiplexación, alineación de lecturas a un genoma de referencia, recuento de lecturas alineadas a porciones genómicas en el genoma de referencia, y similares. El uno o más aparatos informáticos a veces se configuran para realizar uno o más de los siguientes procesos adicionales: normalizar recuentos de lectura (por ejemplo, reducir o eliminar sesgos), generar una o más determinaciones (por ejemplo, determinar la fracción fetal, ploidía fetal, sexo fetal, recuento cromosómico fetal, resultado, presencia o ausencia de una variación genética (por ejemplo, presencia o ausencia de una aneuploidía cromosómica fetal (por ejemplo, trisomía del cromosoma 13, 18 y/o 21)) y similares.

En algunas implementaciones, un aparato informático está asociado con un aparato de secuenciación, y en determinadas implementaciones, un aparato informático realiza la mayoría o todos los siguientes procesos: generar identificaciones base a partir de señales de aparato de secuenciación, lecturas de ensamblaje, lecturas de demultiplexado, lecturas de alineación y recuento de las lecturas alineadas con porciones genómicas de un genoma de referencia, normalizar recuentos de lectura y generar uno o más resultados (por ejemplo, fracción fetal, presencia o ausencia de una variación genética particular). En las últimas implementaciones, en las que un aparato informático se asocia con un aparato de secuenciación, el aparato informático a menudo incluye uno o más procesadores (por ejemplo, microprocesadores) y memoria que tiene instrucciones que son llevadas a cabo por el uno o más procesadores para realizar los procesos. En algunas implementaciones, un aparato informático puede ser un dispositivo informático único o de múltiples núcleos local en el aparato de secuenciación (por ejemplo, ubicado en la misma ubicación (por ejemplo, la misma dirección, el mismo edificio, el mismo piso, la misma habitación o similares)). En algunas implementaciones, un aparato informático está integrado con el aparato de secuenciación.

En algunas implementaciones, múltiples aparatos informáticos en un sistema están asociados con un aparato de secuenciación, y un subconjunto de los procesos totales realizados por el sistema puede asignarse o dividirse entre un aparato informático particular en el sistema. Los subconjuntos del número total de procesos pueden dividirse entre dos o más aparatos informáticos, o grupos de los mismos, en cualquier combinación adecuada. En determinadas implementaciones, generar identificaciones base a partir de señales de aparato de secuenciación, ensamblar lecturas y demultiplexar lecturas se realiza por un primer aparato informático o grupo del mismo, alinear y contar lecturas mapeadas en porciones de un genoma de referencia se realiza por un segundo aparato informático o grupo del mismo, y normalizar recuentos de lectura y proporcionar uno o más resultados se realiza por un tercer aparato informático o grupo del mismo. En sistemas que comprenden dos o más aparatos informáticos o grupos de los mismos, cada aparato informático particular puede incluir memoria, uno o más procesadores o una combinación de los mismos. Un sistema de aparato multiinformático a veces incluye uno o más servidores adecuados locales a un aparato de secuenciación, y a veces incluye uno o más servidores adecuados no locales al aparato de secuenciación (por ejemplo, servidores web, servidores en línea, servidores de aplicaciones, servidores de archivos remotos, servidores en la nube (por ejemplo, entorno en la nube, computación en la nube)).

El aparato en diferentes configuraciones de sistema puede generar diferentes tipos de datos de salida. Por ejemplo, un aparato de secuenciación puede emitir señales de base y los datos de salida de señal de base pueden transferirse a un aparato informático que convierte los datos de señal de base en identificaciones base. En algunas implementaciones, las identificaciones base son datos de salida de un aparato informático y se transfieren a otro

aparato informático para generar lecturas de secuencia. En determinadas implementaciones, las identificaciones base no son datos de salida de un aparato particular, y en su lugar, se utilizan en el mismo aparato que recibió señales de base del aparato de secuenciación para generar lecturas de secuencia. En algunas implementaciones, un aparato recibe señales de base del aparato de secuenciación, genera identificaciones base, lecturas de secuencia y demultiplexa lecturas de secuencia, y emite lecturas de secuencia demultiplexadas para una muestra que puede transferirse a otro aparato o grupo de las mismas que alinea las lecturas de secuencia con un genoma de referencia. En algunas implementaciones, un aparato o grupo de los mismos puede emitir lecturas de secuencia alineadas mapeadas en porciones de un genoma de referencia (por ejemplo, archivos SAM o BAM), y dichos datos de salida pueden transferirse a un segundo aparato informático o grupo del mismo que normaliza las lecturas de secuencia (por ejemplo, normaliza los recuentos de las lecturas de secuencia) y genera un resultado (por ejemplo, fracción fetal y/o presencia o ausencia de una trisomía fetal). Los datos de salida de un aparato pueden transferirse a un segundo aparato de cualquier manera adecuada. Por ejemplo, los datos de salida de un aparato a veces se colocan en un dispositivo de almacenamiento físico y el dispositivo de almacenamiento se transporta y conecta a un segundo aparato al que se transfieren los datos de salida. Los datos de salida a veces se almacenan por un aparato en una base de datos, y un segundo aparato accede a los datos de salida de la misma base de datos.

Un sistema a veces comprende una máquina de reducción de sesgo. Una máquina de reducción de sesgo a veces comprende uno o más ordenadores. En algunas implementaciones, una máquina de reducción de sesgo mapea lecturas de secuencia y/o comprime lecturas (por ejemplo, lecturas de secuencia mapeadas). Una máquina de reducción de sesgo a veces comprime lecturas de secuencia en un formato comprimido adecuado (por ejemplo, un formato BReads). En algunas implementaciones, una máquina de reducción de sesgo genera densidades de lectura, perfiles de densidad, perfiles de densidad de lectura ajustados y/o resultados. Una o más funciones de una máquina de reducción de sesgo pueden ser realizadas por una red y/o una nube (por ejemplo, red informática en la nube). Una máquina de reducción de sesgo puede interactuar con múltiples servidores (por ejemplo, servidores en la nube) que comprenden microprocesadores, medios de memoria y almacenamiento, módulos, datos y/o información (por ejemplo, referencias, lecturas de secuencia de referencia, densidades de lectura de referencia, perfiles de densidad de referencia y similares) y/o software. Una máquina de reducción de sesgo puede transferir datos y/o información a una nube donde se realizan una o más funciones de una máquina de reducción de sesgo. Los datos procesados y/o información pueden transferirse a una máquina de reducción de sesgo desde una nube.

Un sistema a veces comprende una máquina de secuenciación y una máquina de reducción de sesgo donde una máquina de secuenciación genera lecturas de secuencia de ácido nucleico de muestra, a veces mapea lecturas de secuencia y proporciona y/o transfiere lecturas de secuencia no mapeadas o mapeadas en una máquina de reducción de sesgo. Una máquina de secuenciación puede proporcionar o transferir lecturas a una máquina de reducción de sesgo mediante cualquier método adecuado. Una máquina de secuenciación y una máquina de reducción de sesgo a veces se conectan entre sí mediante una interfaz de hardware adecuada. En algunas implementaciones, una máquina de secuenciación y una máquina de reducción de sesgo están conectadas a una red y/o a una nube. En algunas implementaciones, una máquina de secuenciación y una máquina de reducción de sesgo están conectadas entre sí por red y/o una nube. Algunos o todos los métodos y/o funciones de una máquina de secuenciación y/o una máquina de reducción de sesgo pueden realizarse por una nube. Una máquina de secuenciación puede transferir lecturas mediante el uso de un medio legible por ordenador transitorio y/o no transitorio a una máquina de reducción de sesgo. Por ejemplo, las lecturas de secuencia pueden transferirse por señales digitales o analógicas transmitidas por cables cableados y/o señales inalámbricas. En algunas implementaciones, las lecturas de secuencia se transfieren desde una máquina de secuenciación a una máquina de reducción de sesgo usando medio de almacenamiento legible por ordenador no transitorio.

Una máquina de reducción de sesgo puede comprender uno o más módulos descritos en el presente documento que pueden llevar a cabo algunas o todas las funciones de una máquina de reducción de sesgo. En algunas implementaciones, una máquina de reducción de sesgo comprende un módulo de compresión y lleva a cabo la función de un módulo de compresión. En algunas implementaciones, una máquina de reducción de sesgo comprende uno o más de un módulo de densidad de sesgo, un módulo de relación, un módulo de corrección de sesgo y/o un módulo de corrección multivariante. Una máquina de corrección de sesgo puede usar uno o más módulos para eliminar el sesgo (por ejemplo, sesgo de GC) de lecturas y/o proporcionar recuentos normalizados de lecturas de muestra. En algunas implementaciones, una máquina de corrección de sesgo comprende uno o más de un módulo de distribución, un módulo de filtrado y/o un módulo de generación de perfil. Una máquina de corrección de sesgo puede procesar a menudo lecturas de secuencia de un conjunto de entrenamiento o referencia, así como lecturas de secuencia de una muestra de prueba. En algunas implementaciones, una máquina de corrección de sesgo comprende uno o más de un módulo de estadísticas de PCA y/o un módulo de ponderación de porción. Una máquina de corrección de sesgo utiliza a menudo lecturas mapeadas y múltiples módulos y proporciona densidades de lectura, perfiles de densidad y/o perfiles de densidad de lectura ajustados a un módulo de puntuación, un usuario final, un periférico de ordenador (por ejemplo, una pantalla, una impresora) o a una máquina generadora de resultados. En algunas implementaciones, una máquina de reducción de sesgo proporciona un resultado. A veces, una máquina de reducción de sesgo no proporciona un resultado. En algunas implementaciones, una máquina de reducción de sesgo comprende una máquina generadora de resultados. A veces, una máquina de reducción de sesgo transfiere lecturas normalizadas, densidades de lectura, perfiles de densidad y/o perfiles de densidad de lectura ajustados a una máquina generadora de resultados. Una máquina de reducción de sesgo puede transferir datos y/o información (por ejemplo, perfiles de

densidad de lectura) a una máquina generadora de resultados mediante cualquier método adecuado. En algunas implementaciones, un sistema comprende una o más de una máquina de secuenciación, una máquina de reducción de sesgo y/o una máquina generadora de resultados. Una máquina generadora de resultados puede recibir recuentos normalizados de lecturas, densidades de lectura, perfiles de densidad y/o perfiles de densidad de lectura ajustados a partir de una máquina de corrección de sesgo. Una máquina generadora de resultados a menudo proporciona una identificación o un resultado (por ejemplo, una determinación de la presencia o ausencia de una variación genética). Una máquina generadora de resultados a menudo proporciona una identificación o un resultado a un usuario final y/o un periférico de ordenador (por ejemplo, una pantalla, una impresora). Una máquina generadora de resultados a veces comprende uno o más de un módulo de filtrado, módulo de distribución, un módulo de generación de perfil, módulo de estadísticas de PCA, módulo de ponderación de porción, módulo de puntuación y/o uno o más módulos adecuados.

En algunas implementaciones, un usuario interactúa con un aparato (por ejemplo, un aparato informático, un aparato de secuenciación). En algunas implementaciones, un usuario puede colocar una consulta en un sistema, ordenador o módulo que, después, puede adquirir un conjunto de datos por medio de acceso a Internet, (por ejemplo, una nube) y en determinadas implementaciones, puede solicitarse a un procesador programable que adquiera un conjunto de datos adecuado basándose en parámetros dados. Un procesador programable también puede solicitar a un usuario que seleccione una o más opciones de conjuntos de datos seleccionadas por el procesador basándose en parámetros dados. Un procesador programable puede solicitar a un usuario que seleccione una o más opciones de conjuntos de datos seleccionadas por el procesador basándose en la información que se encuentra a través de Internet, otra información interna o externa o similares.

Las opciones pueden elegirse para seleccionar una o más selecciones de características de datos, uno o más algoritmos estadísticos, uno o más algoritmos de análisis estadístico, uno o más algoritmos de significación estadística, etapas iterativas, uno o más algoritmos de validación y una o más representaciones gráficas de métodos, aparatos o programas informáticos.

Los sistemas abordados en el presente documento pueden comprender componentes generales de sistemas informáticos tales como, por ejemplo, servidores de red, sistemas portátiles, sistemas de escritorio, sistemas de mano, asistentes digitales personales, quioscos informáticos y similares. Un sistema informático puede comprender uno o más medios de entrada tales como un teclado, una pantalla táctil, un ratón, reconocimiento de voz u otros medios para permitir que el usuario introduzca datos en el sistema. Un sistema puede comprender además una o más salidas, incluyendo, aunque no de forma limitativa, 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 algunas implementaciones, un módulo de visualización procesa, transforma y/o transfiere datos y/o información a un medio visual adecuado para su presentación en un sistema de representación adecuado (por ejemplo, un monitor, LED, LCD, CRT, similares o combinaciones de los mismos), una impresora, un periférico o dispositivo. En determinadas implementaciones, un módulo de visualización proporciona una visualización de un relación, un perfil o resultado. Los ejemplos no limitantes de un medio visual y/o pantalla adecuados incluyen un cuadro, representación, gráfico, similares o combinaciones de los mismos. En algunas implementaciones, un módulo de visualización procesa, transforma datos y/o información en una representación visual de un genoma fetal y/o materno o un segmento del mismo (por ejemplo, un cromosoma o parte del mismo). En algunas implementaciones, se requiere un módulo de visualización o una máquina que incluya un módulo de visualización para proporcionar una visualización adecuada.

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

Un sistema puede incluir una interfaz de comunicaciones en algunas implementaciones. Una interfaz de comunicaciones permite la transferencia de software y datos entre un sistema informático y uno o más dispositivos externos. Los ejemplos no limitativos de interfaces de comunicaciones incluyen un módem, una interfaz de red (tal como una tarjeta Ethernet), un puerto de comunicaciones, una ranura y tarjeta PCMCIA, y similares. El software y los datos transferidos por medio de una interfaz de comunicaciones generalmente están en forma de señales, que pueden ser señales electrónicas, electromagnéticas, ópticas y/u otras señales capaces de recibirse por una interfaz de

comunicaciones. A menudo, se proporcionan señales a una interfaz de comunicaciones mediante un canal. Un canal transporta a menudo señales y puede implementarse usando hilo o cable, fibra óptica, una línea telefónica, un enlace de teléfono celular, un enlace de RF y/u otros canales de comunicaciones. Por tanto, en un ejemplo, puede usarse una interfaz de comunicaciones para recibir información de señal que puede detectarse por un módulo de detección de señal.

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

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

Un sistema puede incluir un software útil para realizar un procedimiento descrito en el presente documento, y el software puede incluir uno o más módulos para realizar tales procedimientos (por ejemplo, módulo de secuenciación, módulo de corrección de sesgo, módulo de visualización). El término “software” se refiere a instrucciones de programas legibles por ordenador que, cuando se ejecutan por un ordenador, realizan operaciones informáticas. Las instrucciones ejecutables por el uno o más procesadores a veces se proporcionan como código ejecutable, que cuando se ejecutan, pueden hacer que uno o más procesadores implementen un método descrito en el presente documento.

Un módulo descrito en el presente documento puede existir como software, e instrucciones (por ejemplo, procesos, rutinas, subrutinas) incorporadas en el software pueden implementarse o realizarse por un procesador. Por ejemplo, un módulo (por ejemplo, un módulo de software) puede formar parte de un programa que realiza un procedimiento o tarea particular. El término “módulo” se refiere a una unidad funcional autónoma que puede usarse en un aparato o sistema de software más grande. Un módulo puede comprender un conjunto de instrucciones para llevar a cabo una función del módulo por uno o más microprocesadores. Se pueden implementar instrucciones de un módulo en un entorno informático usando un lenguaje de programación adecuado, un software adecuado y/o un código escrito en un lenguaje adecuado (por ejemplo, un lenguaje de programación informático conocido en la técnica) y/o sistema operativo, cuyos ejemplos no limitativos incluyen UNIX, Linux, Oracle, Windows, Ubuntu, ActionScript, C, C++, C#, Haskell, Java, JavaScript, Objective-C, Perl, Python, Ruby, Smalltalk, SQL, Visual Basic, COBOL, Fortran, UML, HTML (por ejemplo, con PHP), PGP, G, R, S, similares o combinaciones de los mismos. En algunas implementaciones, un módulo descrito en el presente documento comprende código (por ejemplo, secuencia de comandos) escrito en S o R que utiliza un paquete adecuado (por ejemplo, un paquete S, un paquete R). R, el código fuente de R, los programas de R, los paquetes de R y la documentación de R están disponibles para su descarga desde un CRAN o un sitio espejo de CRAN (The Comprehensive R Archive Network (CRAN) [en línea], [recuperado el 24-04-2013], obtenido de Internet <URL:*><http://cran.us.r-project.org/><>). CRAN es una red de servidores ftp y web en todo el mundo que almacenan versiones idénticas y actualizadas de código y documentación para R.

Un módulo puede transformar información y/o datos. La información y/o los datos pueden estar en una forma adecuada. Por ejemplo, la información y/o los datos pueden ser digitales o analógicos. En determinadas implementaciones, la información y/o los datos pueden ser paquetes, bytes, caracteres o bits. En algunas implementaciones, la información y/o los datos pueden ser cualquier información o dato recopilado, ensamblado o utilizable. Los ejemplos no limitativos de información y/o datos incluyen un medio adecuado, archivos, imágenes, vídeo, sonido (por ejemplo, frecuencias, audibles o no audibles), números, constantes, valores, objetos, tiempo, texto, funciones, código informático, mapas, referencias, secuencias, lecturas, lecturas mapeadas, densidades de lectura, perfiles de densidad de lectura, rangos, umbrales, visualizaciones, representaciones, resultados, transformaciones, lo similar o combinaciones de los mismos. Un módulo puede aceptar o recibir información y/o datos, transformar la información y/o los datos en una segunda forma y proporcionar o transferir la segunda forma a una máquina, periférico, componente u otro módulo. Un módulo puede realizar una o más de las siguientes funciones no limitativas: mapear lecturas de secuencia, comprimir un archivo (por ejemplo, datos de lectura mapeados), filtrar porciones, seleccionar porciones, realizar un PCA, proporcionar componentes principales, ajustar densidades de lectura y/o los perfiles de densidad de lectura, ponderar porciones, puntuar, proporcionar recuentos, ensamblar porciones, normalizar recuentos, proporcionar estimados de sesgo de genoma local, proporcionar frecuencias de sesgo, proporcionar densidades de lectura, proporcionar perfiles de densidad de lectura, proporcionar una zona de identificación y/o una zona de no identificación, proporcionar una medida de incertidumbre, proporcionar o determinar los rangos esperados (por ejemplo, rangos de umbral y niveles umbral), representar y/o determinar un resultado, por ejemplo. Un procesador puede, en determinadas implementaciones, llevar a cabo las instrucciones en un módulo. En algunas

implementaciones, se requiere que uno o más procesadores lleven a cabo instrucciones en un módulo o grupo de módulos. Un módulo puede proporcionar información y/o datos a otro módulo, aparato o fuente y puede recibir información y/o datos de otro módulo, aparato o fuente.

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

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

El software a menudo se proporciona en un producto de programa que contiene instrucciones de programa grabadas en un medio legible por ordenador (por ejemplo, un medio legible por ordenador no transitorio), incluidos, aunque no de forma limitativa, medios magnéticos que incluyen disquetes, discos duros y cintas magnéticas; y medios ópticos que incluyen discos CD-ROM, discos DVD, discos magnetoópticos, unidades de estado sólido, unidades flash, RAM, ROM, BUS, disquetes, similares y otros medios similares en los que se pueden grabar las instrucciones del programa. En la implementación en línea, un servidor y un sitio web mantenidos por una organización pueden estar configurados para proporcionar descargas de software a usuarios remotos, o los usuarios remotos pueden acceder a un sistema remoto mantenido por una organización para acceder de manera remota al software. El software puede obtener o recibir información de entrada. El software puede incluir un módulo que obtiene o recibe de manera específica datos (por ejemplo, un módulo receptor de datos que recibe datos de lectura de secuencias y/o datos de lectura mapeados) y puede incluir un módulo que procesa de manera específica los datos (por ejemplo, un módulo de procesamiento que procesa los datos recibidos (por ejemplo, filtra, normaliza, proporciona un resultado y/o informe). Los términos “obtener” y “recibir” información de entrada se refieren a recibir datos (por ejemplo, lecturas de secuencia, lecturas mapeadas) mediante medios de comunicación por ordenador desde un sitio local, o remoto, entrada de datos por seres humanos, o cualquier otro método para recibir datos. La información de entrada puede generarse en la misma ubicación en la que se recibe, o puede generarse en una ubicación diferente y transmitirse a la ubicación de recepción. En algunas implementaciones, la información de entrada se modifica antes de procesarse (por ejemplo, se coloca en un formato susceptible al procesamiento (por ejemplo, tabulado)).

- El software puede incluir uno o más algoritmos en determinadas implementaciones. Un algoritmo puede usarse para procesar datos y/o proporcionar un resultado o informe según una secuencia finita de instrucciones. Un algoritmo es a menudo una lista de instrucciones definidas para completar una tarea. Partiendo de un estado inicial, las instrucciones pueden describir un cálculo que avanza a través de una serie definida de estados sucesivos, terminando eventualmente en un estado final. La transición de un estado al siguiente no es necesariamente determinista (por ejemplo, algunos algoritmos incorporan aleatoriedad). A modo de ejemplo, y sin limitación, un algoritmo puede ser un
- 5 algoritmo de búsqueda, algoritmo de clasificación, algoritmo de fusión, algoritmo numérico, algoritmo gráfico, algoritmo de cadena, algoritmo de modelado, algoritmo de genométrica computacional, algoritmo combinatorio, algoritmo de aprendizaje automático, algoritmo de criptografía, algoritmo de compresión de datos, algoritmo de análisis y similares.
- 10 Un algoritmo puede incluir un algoritmo o dos o más algoritmos que funcionan en combinación. Un algoritmo puede ser de cualquier clase de complejidad adecuada y/o complejidad parametrizada. Puede usarse un algoritmo para el cálculo y/o procesamiento de datos y, en algunas implementaciones, puede usarse en un enfoque determinista o probabilístico/predictivo. Puede implementarse un algoritmo en un entorno informático usando un lenguaje de programación adecuado, son ejemplos no limitativos de los mismos C, C++, Java, Perl, R, S, Python, Fortran, y
- 15 similares. En algunas implementaciones, un algoritmo puede configurarse o modificarse para incluir margen de errores, análisis estadístico, significación estadística una medida de incertidumbre y/o comparaciones con otra información o conjuntos de datos (por ejemplo, aplicable cuando se usa una red neural o algoritmo de agrupamiento).
- En determinadas implementaciones, pueden implementarse varios algoritmos para su uso en software. Estos algoritmos pueden entrenarse con datos sin procesar en algunas implementaciones. Para cada nueva muestra de datos sin procesar, los algoritmos entrenados pueden producir un conjunto o resultado de datos procesados representativos. Un conjunto de datos procesados algunas veces tiene una complejidad reducida en comparación con el conjunto de datos original que se procesó.
- 20 Basándose en un conjunto procesado, el rendimiento de un algoritmo entrenado puede evaluarse basándose en la sensibilidad y especificidad, en algunas implementaciones. Un algoritmo con la mayor sensibilidad y/o especificidad puede identificarse y usarse, en determinadas implementaciones.
- En determinadas implementaciones, los datos simulados (o simulación) pueden ayudar al procesamiento de datos, por ejemplo, entrenando un algoritmo o someter a prueba un algoritmo. En algunas implementaciones, los datos simulados incluyen varios muestreos hipotéticos de agrupamientos diferentes de lecturas de secuencia. Los datos simulados pueden basarse en lo que podría esperarse de una población real o pueden sesgarse para someter a prueba un algoritmo y/o asignar una clasificación correcta. Los datos simulados también se denominan en el presente documento datos “virtuales”. Las simulaciones pueden realizarse mediante un programa informático en determinadas
- 30 implementaciones. Una etapa posible en el uso de un conjunto de datos simulados es evaluar la confianza de un resultado identificado, por ejemplo, cuán bien coincide un muestreo aleatorio o representa mejor los datos originales. Un enfoque consiste en calcular un valor de probabilidad (valor de p) que estima la probabilidad de una muestra aleatoria con mejor puntuación que las muestras seleccionadas. En algunas implementaciones, puede evaluarse un modelo empírico, en el cual se supone que al menos una muestra coincide con una muestra de referencia (con o sin variaciones resueltas). En algunas implementaciones, otra distribución, tal como una distribución de Poisson, por
- 40 ejemplo, puede usarse para definir la distribución de probabilidad.
- Un sistema puede incluir uno o más procesadores en determinadas implementaciones. Un procesador puede conectarse a un bus de comunicaciones. Un sistema informático puede incluir una memoria principal, a menudo memoria de acceso aleatorio (RAM), y puede incluir además una memoria secundaria. La memoria en algunas implementaciones comprende un medio de almacenamiento no transitorio legible por ordenador. La memoria secundaria puede incluir, por ejemplo, una unidad de disco duro y/o una unidad de almacenamiento extraíble, que representa una unidad de disquete, una unidad de cinta magnética, una unidad de disco óptico, tarjeta de memoria y similares. Una unidad de almacenamiento extraíble a menudo lee y/o escribe en una unidad de almacenamiento extraíble. Los ejemplos no limitativos de unidades de almacenamiento extraíbles incluyen un disquete, una cinta magnética, un disco óptico y similares, que pueden leerse y escribirse, por ejemplo, en una unidad de almacenamiento extraíble. Una unidad de almacenamiento extraíble puede incluir un medio de almacenamiento utilizable por ordenador que tiene almacenados un software y/o datos informáticos.
- 45 Un procesador puede implementar software en un sistema. En algunas implementaciones, un procesador puede programarse para realizar automáticamente una tarea descrita en el presente documento que un usuario podría realizar. En consecuencia, un procesador, o algoritmo conducido por tal procesador, puede requerir poca o ninguna supervisión o entrada de un usuario (por ejemplo, el software puede programarse para implementar una función automáticamente). En algunas implementaciones, la complejidad de un procedimiento es tan grande que una sola
- 50 persona o grupo de personas no podría realizar el procedimiento en un marco de tiempo lo suficientemente corto para determinar la presencia o ausencia de una variación genética.
- En algunas implementaciones, la memoria secundaria puede incluir otros medios similares para permitir que los programas informáticos u otras instrucciones se carguen en un sistema informático. Por ejemplo, un sistema puede
- 65 incluir una unidad de almacenamiento extraíble y un dispositivo de interfaz. Los ejemplos no limitativos de tales sistemas incluyen un cartucho de programa e interfaz de cartucho (tales como los que se encuentran en dispositivos

de videojuegos), un chip de memoria extraíble (tal como una EPROM o PROM) y un enchufe asociado, y otras unidades de almacenamiento extraíbles e interfaces que permiten que el software y los datos se transfieran desde la unidad de almacenamiento extraíble a un sistema informático.

Una entidad puede generar recuentos de lecturas de secuencia, mapear las lecturas de secuencia en porciones, contar las lecturas mapeadas y utilizar las lecturas mapeadas contadas en un método, sistema, máquina o producto de programa informático descrito en el presente documento, en algunas implementaciones. Los recuentos de lecturas de secuencia mapeadas en porciones a veces se transfieren por una entidad a una segunda entidad para su uso por la segunda entidad en un método, sistema, máquina o producto de programa informático descrito en el presente documento, en determinadas implementaciones.

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

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

La FIG. 11 ilustra un ejemplo no limitativo de un entorno informático 510 en el que pueden implementarse varios sistemas, métodos, algoritmos y estructuras de datos descritos en el presente documento. El entorno informático 510 es solo un ejemplo de un entorno informático adecuado y no pretende sugerir ninguna limitación en cuanto al alcance del uso o la funcionalidad de los sistemas, métodos y estructuras de datos descritos en el presente documento. Tampoco debe interpretarse que el entorno informático 510 tiene alguna dependencia o requisito relacionado con uno o una combinación de componentes ilustrados en el entorno informático 510. En determinadas implementaciones, se puede utilizar un subconjunto de LOS sistemas, métodos y estructuras de datos que se muestran en la FIG. 11. Los sistemas, métodos y estructuras de datos descritos en el presente documento son operativos con muchos otros entornos o configuraciones de sistemas informáticos de uso general o especial. Los ejemplos de sistemas, entornos y/o configuraciones informáticos conocidos que pueden ser adecuados incluyen, aunque no de forma limitativa, ordenadores personales, servidores, clientes ligeros, clientes pesados, dispositivos portátiles o manuales, sistemas multiprocesador, sistemas basados en microprocesador, decodificadores, electrónica de consumo programable, PC de red, miniordenadores, ordenadores centrales, entornos informáticos distribuidos que incluyen cualquiera de los sistemas o dispositivos anteriores, y similares.

El entorno operativo 510 de la FIG. 11 incluye un dispositivo informático de uso general en forma de ordenador 520, que incluye una unidad de procesamiento 521, una memoria de sistema 522 y un bus 523 del sistema que acopla operativamente varios componentes del sistema, incluida la memoria del sistema 522, a la unidad de procesamiento 521. Puede haber solo una o puede haber más de una unidad de procesamiento 521, de modo que el procesador del ordenador 520 incluya una única unidad de procesamiento central (CPU) o una pluralidad de unidades de procesamiento, comúnmente denominadas un entorno de procesamiento en paralelo. La ordenador 520 puede ser un ordenador convencional, un ordenador distribuido o cualquier otro tipo de ordenador.

El bus 523 del sistema puede ser cualquiera de varios tipos de estructuras de bus que incluyen un bus de memoria o un controlador de memoria, un bus periférico y un bus local que usa cualquiera de una variedad de arquitecturas de bus.

La memoria del sistema también puede denominarse simplemente memoria, e incluye memoria de solo lectura (ROM) 524 y memoria de acceso aleatorio (RAM). Un sistema básico de entrada/salida (BIOS) 526, que contiene las rutinas básicas que ayudan a transferir información entre elementos de dentro del ordenador 520, tal como durante el arranque, se almacena en la ROM 524. El ordenador 520 puede incluir además una interfaz 527 de unidad de disco duro para leer y escribir en un disco duro, que no se muestra, una unidad de disco magnético 528 para leer o escribir en un disco magnético extraíble 529, y una unidad de disco óptico 530 para leer desde o escribir en un disco óptico extraíble 531 tal como un CD ROM u otro medio óptico.

La unidad de disco duro 527, la unidad de disco magnético 528 y la unidad de disco óptico 530 están conectadas al bus 523 del sistema mediante una interfaz 532 de unidad de disco duro, una interfaz 533 de unidad de disco magnético y una interfaz 534 de unidad de disco óptico, respectivamente. Las unidades y sus medios legibles por ordenador asociados proporcionan almacenamiento no volátil de instrucciones legibles por ordenador, estructuras de datos, módulos de programa y otros datos para el ordenador 520. En el entorno operativo, se puede utilizar cualquier tipo de medio legible por ordenador que pueda almacenar datos a los que pueda acceder un ordenador, tales como casetes magnéticos, tarjetas de memoria flash, discos de vídeo digital, cartuchos de Bernoulli, memorias de acceso aleatorio (RAM), memorias de solo lectura (ROM) y similares.

Se pueden almacenar varios módulos de programa en el disco duro, disco magnético 529, disco óptico 531, ROM 524 o RAM, incluido un sistema operativo 535, uno o más programas de aplicación 536, otros módulos de programa 537 y datos de programa 538. Un usuario puede introducir comandos e información en la ordenador personal 520 a través de dispositivos de entrada como un teclado 540 y un dispositivo de puntero 542. Otros dispositivos de entrada (no mostrados) pueden incluir un micrófono, una palanca universal, un controlador de juegos, una antena parabólica, un escáner o similares. Estos y otros dispositivos de entrada a menudo se conectan a la unidad de procesamiento 521 a través de una interfaz 546 de puerto de serie que está acoplada al bus del sistema, pero se pueden conectar mediante otras interfaces, tales como un puerto paralelo, un puerto de juegos o un bus en serie universal (USB). Un monitor 547 u otro tipo de dispositivo de visualización también está conectado al bus 523 del sistema a través de una interfaz, como un adaptador de vídeo 548. Además del monitor, los ordenadores suelen incluir otros dispositivos de salida periféricos (no se muestran), como altavoces e impresoras.

El ordenador 520 puede funcionar en un entorno de red utilizando conexiones lógicas a uno o más ordenadores remotes, como el ordenador remoto 549. Estas conexiones lógicas pueden lograrse mediante un dispositivo de comunicación acoplado a o una parte del ordenador 520, o de otras maneras. El ordenador remoto 549 puede ser otro ordenador, un servidor, un enrutador, un PC de red, un cliente, un dispositivo par u otro nodo de red común, y generalmente incluye muchos o todos los elementos descritos anteriormente en relación con el ordenador 520, aunque solo se ha ilustrado un dispositivo 550 de almacenamiento de memoria en la FIG. 11. Las conexiones lógicas representadas en la FIG. 11 incluyen una red de área local (LAN) 551 y una red de área amplia (WAN) 552. Dichos entornos de red son comunes en las redes de oficina, las redes informáticas de toda la empresa, las intranets e Internet, que son todos tipos de redes.

Cuando se usa en un entorno de red LAN, el ordenador 520 se conecta a la red local 551 a través de una interfaz 553 de red o adaptador, que es un tipo de dispositivo de comunicaciones. Cuando se usa en un entorno de red WAN, el ordenador 520 suele incluir un módem 554, un tipo de dispositivo de comunicaciones o cualquier otro tipo de dispositivo de comunicaciones para establecer comunicaciones a través de la red de área amplia 552. El módem 554, que puede ser interno o externo, está conectado al bus 523 del sistema a través de la interfaz 546 del puerto de serie. En un entorno de red, los módulos de programa representados en relación con el ordenador personal 520, o porciones del mismo, pueden almacenarse en el dispositivo de almacenamiento de memoria remota. Se aprecia que las conexiones de red mostradas son ejemplos no limitativos, y que se pueden usar otros dispositivos de comunicaciones para establecer un enlace de comunicaciones entre ordenadores.

En algunas implementaciones, un sistema comprende uno o más microprocesadores y memoria, cuya memoria comprende instrucciones ejecutables por el uno o más microprocesadores y cuyas instrucciones ejecutables por el uno o más microprocesadores se configuran para (a) generar una relación entre (i) estimaciones de sesgo del genoma local y (ii) frecuencias de sesgo para lecturas de secuencia de una muestra de prueba, lo que genera de este modo una relación de sesgo de muestra, donde las lecturas de secuencia son de ácido nucleico circulante libre de células de la muestra de prueba, y las lecturas de secuencia se asignan a un genoma de referencia, (b) comparar la relación de sesgo de muestra y una relación de sesgo de referencia, lo que genera de este modo una comparación, donde la relación de sesgo de referencia es entre (i) estimaciones de sesgo de genoma local y (ii) las frecuencias de sesgo para una referencia y (c) normalizar los recuentos de las lecturas de secuencia para la muestra según la comparación determinada en (b), mediante la cual el sesgo en las lecturas de secuencias para la muestra se reduce.

En algunas implementaciones, un sistema comprende uno o más microprocesadores y memoria, cuya memoria comprende instrucciones ejecutables por el uno o más microprocesadores y cuyas instrucciones ejecutables por el uno o más microprocesadores se configuran para (a) generar una relación entre (i) densidades de guanina y citosina (GC) y (ii) frecuencias de densidades para lecturas de secuencia de una muestra de prueba, lo que genera de este modo una relación de densidad de GC de muestra, donde las lecturas de secuencia son de ácido nucleico circulante libre de células de la muestra de prueba, y las lecturas de secuencia se asignan a un genoma de referencia, (b) comparar la relación de densidad de GC de muestra y una relación de densidad de GC de referencia, lo que genera de este modo una comparación, donde la relación de densidad de GC de referencia es entre (i) densidades de GC y (ii) las frecuencias de densidad de GC para una referencia y (c) normalizar los recuentos de las lecturas de secuencia para la muestra según la comparación determinada en (b), mediante la cual el sesgo en las lecturas de secuencias para la muestra se reduce.

En algunas implementaciones, un sistema comprende uno o más microprocesadores y memoria, cuya memoria comprende instrucciones ejecutables por uno o más microprocesadores y cuyas instrucciones ejecutables por uno o más microprocesadores se configuran para (a) filtrar, según una distribución de densidad de lectura, porciones de un genoma de referencia, lo que proporciona de este modo un perfil de densidad de lectura para una muestra de prueba que comprende densidades de lectura de porciones filtradas, donde las densidades de lectura se determinan usando lecturas de secuencia de ácido nucleico libre de células circulantes de una muestra de prueba de una mujer embarazada mapeada a un genoma de referencia y la distribución de densidad de lectura se determina para densidades de lectura de porciones para múltiples muestras, (b) ajustar, usando un microprocesador, el perfil de densidad de lectura para la muestra de prueba según uno o más componentes principales, cuyos componentes principales se obtienen de un conjunto de muestras euploides conocidas mediante un análisis de componentes principales, lo que proporciona de este modo un perfil de muestra de prueba que comprende densidades de lectura ajustadas, (c) comparar el perfil de muestra de prueba con un perfil de referencia, lo que proporciona de este modo una comparación y (d) determinar la presencia o ausencia de una aneuploidía cromosómica para la muestra de prueba según la comparación.

En algunas implementaciones, se presentan en el presente documento un medio de almacenamiento legible por ordenador no transitorio que comprende un programa ejecutable almacenado en el mismo. En algunas implementaciones, un medio de almacenamiento legible por ordenador no transitorio que comprende un programa ejecutable almacenado en el mismo comprende un producto de programa informático. En algunas implementaciones, un medio de almacenamiento legible por ordenador no transitorio que comprende un programa ejecutable almacenado en el mismo se refiere al software. Un producto de programa informático es a menudo software. En algunas implementaciones presentadas en el presente documento hay un medio de almacenamiento legible por ordenador no transitorio que comprende un programa ejecutable almacenado en el mismo, donde el programa ordena a un microprocesador que realice lo siguiente: (a) generar una relación entre (i) densidades de guanina y citosina (GC) y (ii) frecuencias de densidades para lecturas de secuencia de una muestra de prueba, lo que genera de este modo una relación de densidad de GC de muestra, donde las lecturas de secuencia son de ácido nucleico circulante libre de células de la muestra de prueba, y las lecturas de secuencia se asignan a un genoma de referencia, (b) comparar la relación de densidad de GC de muestra y una relación de densidad de GC de referencia, lo que genera de este modo una comparación, donde la relación de densidad de GC de referencia es entre (i) densidades de GC y (ii) las frecuencias de densidad de GC para una referencia y (c) normalizar los recuentos de las lecturas de secuencia para la muestra según la comparación determinada en (b), mediante la cual el sesgo en las lecturas de secuencias para la muestra se reduce.

También se presentan en el presente documento, en algunas implementaciones, un medio de almacenamiento legible por ordenador no transitorio que comprende un programa ejecutable almacenado en el mismo, donde el programa ordena a un microprocesador que realice lo siguiente: (a) filtrar, según una distribución de densidad de lectura, porciones de un genoma de referencia, lo que proporciona de este modo un perfil de densidad de lectura para una muestra de prueba que comprende densidades de lectura de porciones filtradas, donde las densidades de lectura comprenden lecturas de secuencia de ácido nucleico libre de células circulantes de una muestra de prueba de una mujer embarazada y la distribución de densidad de lectura se determina para densidades de lectura de porciones para múltiples muestras, (b) ajustar el perfil de densidad de lectura para la muestra de prueba según uno o más componentes principales, cuyos componentes principales se obtienen de un conjunto de muestras euploides conocidas mediante un análisis de componentes principales, lo que proporciona de este modo un perfil de muestra de prueba que comprende densidades de lectura ajustadas, (c) comparar el perfil de muestra de prueba con un perfil de referencia, lo que proporciona de este modo una comparación y (d) determinar la presencia o ausencia de una aneuploidía cromosómica para la muestra de prueba según la comparación.

Módulos

Se pueden utilizar uno o más módulos en un método descrito en el presente documento, cuyos ejemplos no limitantes incluyen un módulo de compresión, módulo de secuenciación, módulo de mapeo, módulo de filtrado, módulo de densidad de sesgo, módulo de relación, módulo de corrección de sesgo, módulo de corrección multivariante, módulo de distribución, módulo de generación de perfil, módulo de estadísticas de PCA, módulo de ponderación de porción, módulo de puntuación, módulo de resultados, módulo de visualización, similares o una combinación de los mismos. En algunas implementaciones, un módulo es un medio legible por ordenador no transitorio que comprende un conjunto de instrucciones (por ejemplo, un producto de programa informático, por ejemplo, software, un programa), donde el conjunto de instrucciones dirige uno o más microprocesadores para realizar una función. En algunas implementaciones, un módulo comprende instrucciones en forma de código informático adecuado (por ejemplo, código fuente). Un código fuente a veces comprende un programa. El código informático a veces comprende uno o más archivos (por ejemplo, archivos de texto). El código informático puede almacenarse en un medio de almacenamiento no transitorio adecuado (por ejemplo, en memoria, por ejemplo, en el disco duro de un ordenador). Los archivos de código informático a menudo están dispuestos en un árbol de directorio (por ejemplo, un árbol de origen). El código informático de un módulo puede escribirse en un lenguaje de programación adecuado no limitante de ejemplos de los cuales incluyen lenguaje de programación C, básico, R, R++, S, java, HTML, similares, o combinaciones de los mismos. En algunas implementaciones, un programa principal adecuado actúa como un intérprete para un código de ordenador. En algunas implementaciones, un módulo comprende y/o tiene acceso a la memoria. Los módulos a veces

son controlados por un microprocesador. En determinadas implementaciones, un módulo o una máquina que comprende uno o más módulos, recopila, ensambla, recibe, obtiene, accede, recupera, proporciona y/o transfiere datos y/o información hacia o desde otro módulo, máquina, componente, periférico u operador de una máquina. En algunas implementaciones, los datos y/o la información (por ejemplo, lecturas de secuencia, recuentos, etc.) se proporcionan a un módulo mediante una máquina que comprende uno o más de los siguientes: una o más celdas de flujo, una cámara, un detector (por ejemplo, un fotodetector, un fotocélula, un detector eléctrico (por ejemplo, un detector de modulación de amplitud, un detector de modulación de frecuencia y fase, un detector de bucle de bloqueo de fase), un contador, un sensor (por ejemplo, un sensor de presión, temperatura, volumen, flujo, peso), un dispositivo de manejo de fluidos, una impresora, una pantalla (por ejemplo, un LED, LCT o CRT), similares o combinaciones de los mismos. Algunas veces un operador de una máquina proporciona una constante, un valor umbral, una fórmula o un valor predeterminado a un módulo. Un módulo a menudo se configura para transferir datos y/o información hacia o desde otro módulo o máquina. Un módulo puede recibir datos y/o información de otro módulo, cuyos ejemplos no limitantes incluyen un módulo de compresión, módulo de secuenciación, módulo de mapeo, módulo de filtrado, módulo de densidad de sesgo, módulo de relación, módulo de corrección de sesgo, módulo de corrección multivariante, módulo de distribución, módulo de generación de perfil, módulo de estadísticas de PCA, módulo de ponderación de porción, módulo de puntuación, módulo de resultados, módulo de visualización, similares o una combinación de los mismos. Un módulo puede manipular y/o transformar información y/o datos. Los datos y/o información derivada o transformada por un módulo pueden transferirse a otra máquina y/o módulo adecuado, cuyos ejemplos no limitantes incluyen un módulo de compresión, módulo de secuenciación, módulo de mapeo, módulo de filtrado, módulo de densidad de sesgo, módulo de relación, módulo de corrección de sesgo, módulo de corrección multivariante, módulo de distribución, módulo de generación de perfil, módulo de estadísticas de PCA, módulo de ponderación de porción, módulo de puntuación, módulo de resultados, módulo de visualización, similares o una combinación de los mismos. Una máquina que comprende un módulo puede comprender al menos un procesador. En algunas implementaciones, los datos y/o la información son recibidos y/o proporcionados por una máquina que comprende un módulo. Una máquina que comprende un módulo puede incluir un procesador (por ejemplo, uno o más procesadores) cuyo procesador puede realizar y/o implementar una o más instrucciones (por ejemplo, procesos, rutinas y/o subrutinas) de un módulo. En algunas implementaciones, un módulo funciona con uno o más procesadores externos (por ejemplo, una red interna o externa, un servidor, dispositivo de almacenamiento y/o una red de almacenamiento (por ejemplo, una nube)). En algunas implementaciones, un sistema (por ejemplo, una implementación de un sistema mostrado en la Figura 10) comprende uno o más de un módulo de compresión, módulo de secuenciación, módulo de mapeo, módulo de filtrado, módulo de densidad de sesgo, módulo de relación, módulo de corrección de sesgo, módulo de corrección multivariante, módulo de distribución, módulo de generación de perfil, módulo de estadísticas de PCA, módulo de ponderación de porción, módulo de puntuación, módulo de resultados, módulo de visualización, similares o una combinación de los mismos.

Transformaciones

Tal como se mencionó anteriormente, los datos a veces se transforman de una forma a otra. Los términos “transformado/a”, “transformación” y derivaciones gramaticales o equivalentes de los mismos, tal como se usan en el presente documento, se refieren a una alteración de los datos de un material de partida físico (por ejemplo, ácido nucleico de muestra de sujeto de referencia y/o sujeto de prueba) en una representación digital del material de partida físico (por ejemplo, datos de lectura de secuencia), y en algunas implementaciones, incluye una transformación adicional en uno o más valores numéricos o representaciones gráficas de la representación digital que pueden usarse para proporcionar un resultado. En determinadas implementaciones, el uno o más valores numéricos y/o representaciones gráficas de datos representados digitalmente pueden usarse para representar el aspecto del genoma físico de un sujeto de prueba (por ejemplo, representar virtualmente y/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 métodos pueden transformar material de partida físico en un valor numérico o representación gráfica, o una representación del aspecto físico del genoma de un sujeto de prueba.

En algunas implementaciones, los métodos y sistemas en el presente documento transforman una mezcla de una multitud de fragmentos de polinucleótidos encontrados en la sangre de una mujer embarazada en una o más representaciones de estructuras microscópicas y/o submicroscópicas específicas (por ejemplo, un cromosoma, o segmento del mismo) presentes en células fetales, maternas o placentarias. Estos fragmentos de polinucleótidos generalmente se originan a partir de diferentes células y tejidos (por ejemplo, músculo, placenta, fetal, por ejemplo, músculo, corazón, hígado, linfocitos, tumor), cromosomas diferentes y elementos genéticos y/o ubicaciones diferentes (por ejemplo, regiones centroméricas, elementos repetitivos, regiones ricas en GC, regiones hipervariables, genes diferentes, elementos reguladores diferentes, intrones, exones y similares). En algunas implementaciones, un sistema descrito en el presente documento transforma fragmentos de polinucleótidos, mediante el uso de una máquina de secuenciación, en lecturas de secuencia. En algunas implementaciones, un sistema descrito en el presente documento transforma lecturas de secuencia, cuyas lecturas de secuencia comprenden sesgo, a recuentos de secuencia normalizados, densidades de lectura y/o perfiles. Las lecturas de secuencia a menudo se transforman en recuentos de secuencia normalizados, densidades de lectura y/o perfiles en los que el sesgo se reduce significativamente, a menudo mediante el uso de una máquina de reducción de sesgo y/o uno o más procesos y/o módulos adecuados (por

ejemplo, un módulo de mapeo, módulo de densidad de sesgo, módulo de relación, módulo de corrección de sesgo y/o módulo de corrección multivariante). Las lecturas de secuencia normalizadas y las densidades de lectura y/o los perfiles de densidad de lectura generados a partir de lecturas de secuencia normalizadas que tienen un sesgo reducido son útiles para generar un resultado más seguro. Las lecturas de secuencia a menudo se alteran mediante una transformación que cambia los parámetros de lectura de secuencia específicos y reduce el sesgo, proporcionando así lecturas de secuencia normalizadas que a veces se transforman en perfiles y resultados.

En algunas implementaciones, la transformación de un conjunto de datos facilita proporcionar un resultado al reducir la complejidad de los datos y/o la dimensionalidad de los datos. La complejidad del conjunto de datos a veces se reduce durante el procedimiento de transformación de un material de partida físico en una representación virtual del material de partida (por ejemplo, lecturas de secuencia representativas del material de partida físico). Una característica o variable adecuada puede usarse para reducir la complejidad y/o dimensionalidad del conjunto de datos. Los ejemplos no limitativos de características que pueden seleccionarse para su uso como característica objetiva 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.

Ejemplos

Los siguientes ejemplos se proporcionan únicamente a modo de ilustración y no a modo de limitación. Por lo tanto, los ejemplos expuestos a continuación ilustran determinadas implementaciones y no limitan la tecnología. Los expertos en la técnica reconocerán fácilmente una variedad de parámetros no críticos que podrían cambiarse o modificarse para producir resultados esencialmente iguales o similares.

Ejemplo 1: ChAI

ChAI es un sistema ilustrativo para determinar la presencia o ausencia de una aneuploidía cromosómica en un feto a partir de lecturas de secuencia obtenidas de un sujeto de prueba (por ejemplo, una mujer embarazada). Un ejemplo de un diagrama de flujo del sistema para ChAI se muestra en la Figura 10A y 10B. Se obtuvieron lecturas de secuencia de un sujeto de prueba hembra gestante y uno o más sujetos de referencia a veces denominados en el presente documento como un conjunto de entrenamiento. Los sujetos femeninos gestantes del conjunto de entrenamiento tenían fetos que fueron euploides según lo confirmado por otros métodos de prueba.

Las lecturas de secuencia se comprimieron primero de un formato SAM o BAM a un formato de lecturas binarias (formato BReads) que permitió que ChAI se ejecutara mucho más rápidamente. El formato BReads almacena la ubicación genómica de cada lectura, que incluye una posición de cromosoma y par de bases determinada según un genoma de referencia y descarta otra información. Un archivo BReads comienza con un recuento de las lecturas contenidas. Esto mejora los tiempos de carga eliminando la necesidad de reasignaciones de memoria. El valor se almacenó en disco como una matriz de cuatro bytes. A continuación, las lecturas se almacenaron usando un formato de 5 bytes, uno para la ordinal cromosoma (índice cero de 1-22, X, Y, M) y cuatro para la posición cromosómica. Los archivos de BReads se cargaron leyendo primero el recuento de lectura de secuencia de los primeros cuatro bytes. Cada lectura de secuencia se carga entonces cinco bytes a la vez, indicando el primer byte una ordinaria cromosómica y la siguiente cuatro conversión a la posición entera. El muestreo aleatorio de lecturas se puede realizar rápidamente usando comandos de salto de disco a índices de lectura específicos.

Como ejemplo, el uso de disco de diferentes formatos se compara con la utilización de disco de formato de BReads en la Tabla 1 para 17.673.732 lecturas mapeadas.

Table1: Uso de disco para diferentes formatos basados en una muestra con 17.673.732 lecturas.

Formato	Uso de espacio
SAM	4,0 GB
Posiciones de lectura mapeadas	247 MB
Posiciones de lectura GZip	97 MB

Formato	Uso de espacio
BReads	85 MB

El formato BReads fue aproximadamente 50 x más pequeño que el archivo SAM original y usó aproximadamente 12 % menos espacio que un formato GZip. Las cajas también tenían la ventaja de almacenar el número de lecturas en la cabeza para una asignación de memoria de un solo tiempo, y pueden muestrearse rápidamente ya que las lecturas no tienen que leerse en orden. Estas características no fueron posibles con los otros formatos.

Modelado de sesgo de GC

A continuación, se aprendieron modelos de sesgo de GC para cada muestra. Las muestras que se designaron para el entrenamiento se usaron, en parte, para crear un filtro de porción y para aprender otros sesgos del genoma que no están bien contabilizados por el sesgo de GC solo. Finalmente, se usaron las estadísticas de entrenamiento para filtrar y puntuar las muestras de prueba.

ChAI modeló un sesgo de GC usando estimaciones de densidad del contenido local de GC. La densidad de GC se estimó a partir de un genoma de referencia mediante el uso de una función de núcleo tal como el núcleo de Epanechnikov (Figura 1). Otros núcleos también son apropiados, incluyendo un núcleo gaussiano o triponderal. El ancho de banda se seleccionó como 200 pb, sin embargo, el parámetro de ancho de banda es flexible.

Usando un núcleo, se estimó la densidad de GC en la resolución de pares de bases en un genoma de referencia (por ejemplo, como se muestra en la Figura 2). Usando las estimaciones de densidad de GC de la referencia, se determinó el contenido local de GC de cada lectura de una muestra. La distribución de las estimaciones de densidad de GC para la muestra se comparó a continuación con la distribución a través del genoma de referencia completo para determinar el sesgo de GC (Figura 3). Se descartaron lecturas y valores de referencia que se correlacionan con regiones ricas en AT (densidad de GC = 0). La diferencia entre la distribución de densidad de GC de una muestra y una referencia se modelizó mediante el uso de un polinomio, ajuste en una relación logarítmica de la densidad de la distribución de referencia dividida por la densidad de la distribución de la muestra (Figura 4). El modelo se ajustó de manera ponderada, con cada peso tomado como valor de densidad de distribución de la muestra para un valor de densidad de GC determinado. Esto garantizó que las colas de la distribución no impulsen excesivamente el ajuste. Se pueden usar otros modelos de ajuste, tales como un modelo de regresión cuantil o distribuciones parametrizadas como sea apropiado para la distribución de sesgo.

Usando el modelo de GC ajustado, cada recuento de una lectura de secuencia para una muestra se ponderó para ajustar su sobrerepresentación o subrepresentación en comparación con la referencia. Al incorporar estos pesos en la estimación de la densidad de lectura, el algoritmo ChAI fue capaz de corregir los sesgos de GC.

Corrección de sesgo multidimensional

El sesgo de GC fue solo uno de varios sesgos que afectan los patrones de lectura en un genoma. Los sesgos adicionales a veces se modelizaron y corrigieron mediante el uso de un modelo multivariado generalizado para estimar los pesos leídos. Esta corrección se realizó de la siguiente manera:

1. Se estimaron los valores de sesgo de N para una muestra de prueba y un genoma de referencia en cada uno de un subconjunto de posiciones genómicas.

2. La densidad de los valores de sesgo se modeló usando un núcleo de suavizado N-dimensional o una función paramétrica apropiada.

3. La relación logarítmica se calculó para un conjunto de valores de densidad tomados de las densidades de referencia y de prueba.

4. La relación logarítmica de densidad se modeló usando los puntos elegidos con un modelo multivariado (por ejemplo, un polinomio de 3.º orden ponderado para cada dimensión).

5. El modelo se usó para estimar la relación de la frecuencia de una lectura dada en comparación con la referencia, y se asignó el peso apropiado.

Filtrado de porciones

Las muestras se puntuaron para anomalías cromosómicas basándose en la representación de lecturas de secuencia (por ejemplo, recuentos) en el genoma. Esta representación se determinó usando una función de densidad, similar a la usada para la estimación de GC local. El núcleo de densidad de lectura tiene generalmente un ancho de banda mucho mayor, siendo el valor predeterminado de 50.000 pb. Cada recuento de una lectura contribuye a la densidad

un valor igual a su peso del modelo de sesgo de GC. La densidad de lectura se puede evaluar en cualquiera o en todos los pares de bases, pero para el rendimiento computacional solo se usaron determinadas ubicaciones. Estas posiciones se denominaron “porciones”. Las porciones pueden ubicarse siempre que sea más importante estimar la densidad de lectura. Para la clasificación de las porciones cromosómicas, inicialmente (por ejemplo, antes del filtrado) se espaciaron uniformemente a través del genoma. Cada porción compuesta por una ventana de 50.000 pb y, antes de filtrar, se superpuso a la siguiente porción adyacente por 25.000 pb.

Algunas porciones incluyen regiones genómicas pobremente mapeadas que condujeron a perturbaciones extremas en la densidad de lectura de muestra a muestra. ChAI identificó y eliminó estas porciones mediante un proceso de filtrado mediante el uso de un conjunto de entrenamiento. Las porciones que mostraron grandes desviaciones en la mediana (por ejemplo, la Figura 5A) y/o los valores de DAM (por ejemplo, la Figura 5B) se eliminaron de la consideración. El umbral de estas desviaciones se tomó como cualquier valor fuera de los cuartiles de la población de entrenamiento en más de cuatro veces el rango entre cuartiles (Figura 5). Este umbral se puede ajustar con precisión para maximizar el rendimiento de la prueba para un conjunto específico de parámetros de ChAI.

Entrenamiento y puntuación

Usando solo lecturas que se correlacionan con porciones filtradas, se calculó el perfil de densidad de lectura del genoma de cada muestra. Las muestras que eran parte del conjunto de entrenamiento se usaron a continuación para estimar las estadísticas de entrenamiento que se usaron para puntuar el conjunto de prueba. Estas estadísticas consistieron en porciones medianas, componentes principales y distribuciones nulas para la estadística de prueba de puntuación. Los medianos de porción y componentes principales se usaron para modelar sesgos de lectura de todo el genoma que pueden estar presentes a partir de cualquier número de artefactos biológicos y técnicos (Figura 6A-C). Para minimizar el impacto de los valores de la porción extrema en el resto de la muestra, cada valor que estaba fuera de 4xIQR a través de las otras porciones de una muestra se recortó a 4xIQR.

Las muestras de prueba se corrigieron para sesgos ocultos restando primero los valores medianos entrenados de los valores de la porción de prueba. Los componentes de los valores de muestra que se correlacionan con los componentes principales entrenados superiores también se eliminaron. Esto se hizo modelando los valores de la porción usando regresión lineal multivariante basándose en los términos de componentes principales (Figura 7). Los valores predichos por el modelo se restaron de los valores de muestra, dejando solo los residuos no sesgados. El número de componentes principales utilizados es opcional, siendo el valor predeterminado de ocho.

Después de correcciones, las muestras se puntuaron mediante el uso de una prueba exacta de Fisher. Este ensayo comparó el número de porciones cuyos valores fueron mayores o menores que la mediana entrenada en la región cromosómica de interés. Estos recuentos se evaluaron contra el resto de las porciones en el genoma. La estadística de puntuación se tomó como el valor de $p \log_{10}$ negativo. Se pueden usar otras estadísticas de puntuación en esta etapa, tal como una prueba de rango con signo de Wilcoxon o una prueba F.

Debido a las correlaciones residuales entre porciones, la estadística de prueba se infló tanto en las muestras de entrenamiento como de prueba. Este inflado se estimó a partir de remuestreos de tipo *bootstrapping* del conjunto de entrenamiento (Figura 8).

Las puntuaciones de las muestras de prueba se corrigieron mediante el uso de esta distribución nula como un fondo empírico. Las puntuaciones que son mucho más grandes que las de la distribución empírica se corrigieron mediante el uso de una extrapolación de Pareto de la cola de la distribución nula.

Identificación de género

El sexo se determinó a partir del perfil de componentes principales de una muestra. En un conjunto de datos de entrenamiento, el 2.º componente principal (por ejemplo, PC2) estaba altamente correlacionado con el género. El uso de un coeficiente de regresión de este componente como estadística de prueba fue un ensayo altamente preciso del sexo (Figura 9A-9B).

Eliminación de las dependencias de porciones

Se tomó una etapa adicional durante una ejecución de ChAI para mejorar el poder predictivo del enfoque. Esto implicó reducir la cantidad de estructura de correlación en la matriz porción-muestra, lo que respalda mejor la suposición de prueba de independencia variable y redujo la frecuencia de puntuaciones significativas en las permutaciones nulas. El enfoque involucró reemplazar las porciones con porciones de eigen ortogonales que contienen casi toda la misma información, pero sin estructura de correlación.

La primera etapa fue aprender una matriz de transformación Meig para un conjunto de porciones de entrenamiento;

1. Descomposición SVD: $M = U * D * V^T$

2. Elegir el número de porciones eigen independientes N: (por ejemplo, tal que la fracción acumulativa de los N elementos diagonales de D sea superior al 95 %)

3. Calcular la pseudoinversión: $Meig = pinv(U[1:N] * D[1:N, 1:N])$

5 La multiplicación izquierda de cualquier subconjunto de la matriz de porciones M por su Meig correspondiente dio como resultado una representación libre de correlación reducida en dimensión de ese subconjunto. De esta manera, se derivó Meig en un conjunto de datos de entrenamiento y se aplicó a muestras de prueba sin modificación adicional.

10 También se usó Meig para transformar la variable de prueba. La variable de prueba se representó como un vector que consiste en todos los ceros, con unos en ubicaciones de desviaciones esperadas (por ejemplo, porciones Chr 21). Este vector se transformó con Meig a través de la multiplicación de la izquierda para igualar apropiadamente los datos de la porción transformada.

15 Este enfoque solo puede crear tantas porciones de eigen independientes como muestras en el conjunto de entrenamiento. Para un conjunto de entrenamiento ilustrativo de 50.000 porciones y 1.000 muestras, los datos transformados contendrán, como máximo, 1.000 porciones. Es probable que esto sea una sobrecorrección, reduciendo el número de porciones drásticamente. El enfoque se puede realizar de manera más suelta calculando transformadas de Meig separadas para subconjuntos más pequeños de datos de porción y aplicándolos por separado. Esto fue particularmente útil para eliminar la estructura de correlación local de las porciones vecinas. También se pueden usar otros enfoques para reducir la estructura de correlación de la porción. Por ejemplo, se pueden usar muchos métodos de agrupamiento para agrupar porciones y reemplazarlas con un conjunto más pequeño de porciones de agregado (por ejemplo, en base a promedios de grupo o centroides).

25 Ejemplo 2: Módulo de Generación de Distribución/Perfil

Se escribe una secuencia de comandos en java para generar perfiles de densidad de lectura a partir de datos de lectura de secuencia (por ejemplo, BReads). El código a continuación se diseñó para recoger datos de lectura para cada lectura de secuencia y actualizar un perfil de densidad en las ventanas de densidad de lectura apropiadas (por ejemplo, densidades de lectura individuales para una porción), ponderadas por la distancia de una lectura desde la mediana o punto medio de una porción, y según la corrección de sesgo de GC de una muestra (véase el ejemplo 4). La secuencia de comandos a continuación puede identificar o utilizar los recuentos ponderados y/o normalizados generados a partir de un módulo de relación o módulo de corrección de sesgo (ejemplo 4). En algunas implementaciones, un módulo de distribución puede comprender parte o la totalidad, o una variación de la secuencia de comandos de java que se muestra a continuación. En algunas implementaciones, un módulo de generación de perfil puede comprender parte o la totalidad, o una variación de la secuencia de comandos de java mostrada a continuación:

```
package utilities.genome;
import java.util.Iterator;

import utilities.data.VectorUtil;
import utilities.text.DataFormatter;

public class ChromDensScaleRunnable implements Runnable{

    private GenomeScaleBoolean mask;
    private GenomeScaleFloat density;
    private final String modelPath;
    private final String brPath;
    private final int bandwidth;
    private final GenomeFloat gcdens;
    private final int stepSize;
    private final int sampleSize;
```

40

```

private final int shift;
private final String report;

public ChromDensScaleRunnable(String modelPath, String brPath, int bandwidth,
GenomeFloat gcdens, int stepSize, GenomeScaleBoolean mask, String report, int sampleSize, int
shift)
{
    this.modelPath = modelPath;
    this.brPath = brPath;
    this.bandwidth = bandwidth;
    this.gcdens = gcdens;
    this.stepSize = stepSize;
    this.mask = mask;
    this.report = report;
    this.sampleSize = sampleSize;
    this.shift = shift;
}

public void run()
{
    double[] mdat = (gcdens==null) ? null : VectorUtil.loadDoubleFromFile(modelPath,
6);

    //Build density
    density = new GenomeScaleFloat(stepSize);

    double correction = 0;

    try
    {
        Iterator<GenomicPosition> readIterator = (sampleSize==1) ?
GenomeIO.scanBReads(brPath) : new BReadsSampler(brPath, sampleSize,true);

        while (readIterator.hasNext())
        {
            GenomicPosition gp = readIterator.next().shift(shift);
            int pos = gp.pos;
            int start = Math.max(0, pos-bandwidth);
            int end = Math.min(pos+bandwidth, GenomeUtil.chromosomeSize(gp.chr)-
1);

            int cindex = gp.chr.ordinal();

            double weight;
            if (gcdens!=null)
            {
                float gc = gcdens.values[cindex][pos-1];
                if (gc==0) continue;

```

```

        weight = modelWeight(mdat, gc);
    }else weight = 1;

    int[][] gpoints = density.getScalePoints(cindex, start, end, mask);

    if (gpoints[0].length==0 || Double.isNaN(weight)) continue;
    if (weight>2) weight = 2;
    if (weight<.5) weight = .5;

    correction += weight;

    for (int i=0;i<gpoints[0].length;i++)
        density.values[cindex][gpoints[0][i]] += kernel((gpoints[1][i]-
pos)/(double)bandwidth) * weight;
    }
    }catch (Exception e)
    {
        System.out.println("THROW!");
        e.printStackTrace();
        System.exit(0);
    }

    //System.out.println(GenomeIO.countReadsFromBReads(brPath));
    //System.out.println(correction);

    //Normalize intensity
    for (int i=0;i<density.values.length;i++)
        for (int j=0;j<density.values[i].length;j++)
        {
            float blah = density.values[i][j];

            density.values[i][j] /= correction;

            if (Double.isNaN(density.values[i][j]) ||
Double.isInfinite(density.values[i][j]))
            {
                System.out.println("NA val2: "+modelPath+",
"+density.values[i][j]+", "+blah+", "+correction);
                System.exit(0);
            }
        }

    if (report!=null) System.out.println(report);
}

public GenomeScaleFloat density()

```

```

    {
        return density;
    }

    private static double kernel(double x)
    {
        return .75 * (1.0 - x*x);
    }

    public static double modelWeight(double[] mdat, double gcdens)
    {
        if (mdat[5]==1) gcdens = Math.log(gcdens);

        double x2 = gcdens * gcdens;

        return Math.pow(2, mdat[0] + mdat[1] * gcdens + mdat[2] * x2 + mdat[3] * x2 *
gcdens);
    }
}

```

Ejemplo 3: Módulo de filtrado

- 5 Se escribe una secuencia de comandos en R para filtrar porciones de un perfil de densidad de lectura. Este código examina un perfil de densidad de lectura en todas las muestras e identifica porciones que se retienen y/o porciones que se descartan (por ejemplo, se eliminan del análisis), en base a un rango entre cuartiles. En algunas implementaciones, un módulo de filtrado comprende alguna o todas, o una variación de la secuencia de comandos R que se muestra a continuación:

10

```

rcodepath <- "I:/ghannum/Projects/Binless/RCode"
mdistpath <-
"I:/ghannum/Projects/Binless/Reference/MarkerDistribution_LDTv2_200_50000_50000.txt"
outpath <-
"I:/ghannum/Projects/Binless/Reference/LDTv2_200_50000_50000_MarkerMask.txt"

args <- commandArgs(trailingOnly = TRUE)
rcodepath <- args[1]
mdistpath <- args[2]
outpath <- args[3]

source(paste(rcodepath,"/src/utilities/scanmatrix.R",sep=""))

dat <- scanMatrix(mdistpath,rownames=FALSE,colnames=TRUE)

m <- apply(dat,1,median)
v <- apply(dat,1,mad)

qm <- quantile(m,c(.25,.75))
qv <- quantile(v,c(.25,.75))

scalem <- qm[2]-qm[1]
scalev <- qv[2]-qv[1]

ok <- m > qm[1]-4*scalem & m < qm[2]+4*scalem & v > qv[1]-4*scalev & v < qv[2]+4*scalev

write.table(matrix(as.integer(ok),1),row.names=F,col.names=F,quote=F,file=outpath,sep="")

```

- 15 Ejemplo 4: Módulo de densidad biológica, módulo de relación, Módulo de corrección de bits y Módulo de representación

Se escribe una secuencia de comandos en R para generar densidades de sesgo, generar y comparar una relación y para corregir el sesgo en lecturas de secuencia. Este código generalmente dirige un microprocesador para analizar

una o más muestras y construir un modelo de sesgo (por ejemplo, una relación y/o una comparación de relaciones) en base a estimaciones de sesgo del genoma local (por ejemplo, densidades de GC) para cada muestra y una referencia. La secuencia de comandos a continuación dirige uno o más procesadores, en parte, para generar una relación entre (i) densidades de guanina y citosina (GC) y (ii) frecuencias de densidad de GC para lecturas de secuencia de una muestra de prueba, lo que genera de este modo una relación de densidad de GC de muestra, (b) comparar la relación de densidad de GC de muestra y una relación de densidad de GC de referencia, lo que genera de este modo una comparación, en donde la relación de densidad de GC de referencia es entre (i) densidades de GC y (ii) las frecuencias de densidad de GC para una referencia y, con una modificación adecuada de la secuencia de comandos, (c) normalizar los recuentos de las lecturas de secuencia para la muestra según la comparación determinada en (b), donde el sesgo en las lecturas de secuencia para la muestra se reduce. En algunas implementaciones, un módulo de densidad de sesgo, un módulo de relación, un módulo de corrección de sesgo y/o un módulo de trazado comprende, algunas, todas o una modificación de parte o la totalidad de la secuencia de comandos mostrada a continuación.

```
gcpath <- "I:/ghannum/Projects/Binless/Reference/BiasMaps/DnaseDensity_200_dist.txt"
inpath <- "I:/ghannum/Projects/Binless/Models/LDTv2_DNase-200"
outpath <- "I:/ghannum/Projects/Binless/Models/LDTv2_DNase-200"
makePlots <- TRUE
logTransform <- TRUE
```

```
args <- commandArgs(trailingOnly = TRUE)
```

```

gcpath <- args[1]
inpath <- args[2]
outpath <- args[3]
makePlots <- args[4]
logTransform <- as.logical(args[5])
paths <- dir(inpath)
paths <- paths[grepl("_BiasDistr.txt$",paths)]

gcrcf <- scan(gcpath,0)
gcrcf <- gcrcf[gcrcf!=0]

if (logTransform) gcrcf <- log(gcrcf)

from <- quantile(gcrcf,.005)
to <- quantile(gcrcf,.995)

x <- seq(from,to,length.out=100);

d1y <- predict(smooth.spline(density(gcrcf,from=from,to=to)),x)$y
if (!logTransform) d1y <- sapply(d1y,function(x){max(x,0)})

print(paste("Processing",length(paths),"models."))

for (f in paths)
{
  distr <- scan(paste(inpath,"/",f,sep=""),0)
  distr <- distr[distr!=0]
  if (logTransform) distr <- log(distr)
  d2y <- predict(smooth.spline(density(distr,from=from,to=to)),x)$y
  if (!logTransform) d2y <- sapply(d2y,function(x){max(x,0)})
  pp <- log2(d1y / d2y)
  pp[pp > 2] <- 2; pp[pp < -2] <- -2
  ok <- !is.na(pp)
  mod <- lm(pp[ok]~x+l(x^2)+l(x^3), data=list(x=x[ok], w=d2y[ok])
  w <- 2^predict(mod,list(x=distr))
  fname <- substr(f,1,nchar(f)-14)
  out <- c(mod$coefficients,mean(w))
  out[out==Inf] <- "Infinity"
}

```

```

out[out==-Inf] <- "-Infinity"
write.table(matrix(c(out,as.integer(logTransform))),ncol=1,file=paste(outpath,"/",fname,
e,"_BiasMod.txt",sep=""),row.names=F,col.names=F,quote=F)

if (makePlots)
{

png(units="in",height=4,width=4,res=300,file=paste(outpath,"/",fname,"_BiasMod.png",sep=""))

  if (logTransform)
  {
    plot(x[ok],pp[ok],ylim=c(-4,4),xlab="Bias Density",ylab="Log2 Ratio
(Reference / Sample)")
  }else plot(x[ok],pp[ok],ylim=c(-4,4),xlab="Log-Bias Density",ylab="Log2
Ratio (Reference / Sample)")
  abline(h=0,lty=2)
  lines(x[ok],predict(mod),col=3)
  dev.off()
}
}

##Demo transformation
#load("I:/ghannum/Projects/Binless/2012_11_13_cewi_PERUN_19FCs_AltGCbias_chrFracti
ons.RData")
#
#d <- dir("I:/ghannum/Projects/Binless/GCDistribution/LDTv2/")
#d <- d[grep("_GCDistr.txt",d)]
#v <- as.numeric(df.cewi.GCbiasTable[, "gcBiasRobust"])[1:length(d)]
#
#a <-
scan(paste("I:/ghannum/Projects/Binless/GCDistribution/LDTv2/",d[which.min(v)],sep=""),0);
a <- sort(a)
#b <-
scan(paste("I:/ghannum/Projects/Binless/GCDistribution/LDTv2/",d[which.max(v)],sep=""),0);
d <- sort(d)
#
#r <- scan("I:/ghannum/Projects/Binless/Reference/GCDensity_200_density.txt",0)
#

```

```

#plot(density(r),ylim=c(0,1e10),xlab="GC Density"); lines(density(a),col=3);
lines(density(b),col=2)
#
#a <- a[a!=0]
#b <- b[b!=0]
#r <- r[r!=0]
#
#plot(density(r),ylim=c(0,1e10),xlab="GC Density"); lines(density(a),col=3);
lines(density(b),col=2)
#
#modA <-
as.numeric(scan(paste("l:/ghannum/Projects/Binless/GCDistribution/LDTv2/",substr(d[which.
min(v)],1,nchar(d[which.min(v))]-12),"_GCMod.txt",sep=""),""))
#modB <-
as.numeric(scan(paste("l:/ghannum/Projects/Binless/GCDistribution/LDTv2/",substr(d[which.
max(v)],1,nchar(d[which.max(v))]-12),"_GCMod.txt",sep=""),""))
#
#wa <- sapply(a,function(x){2^sum(c(1,x,x^2,x^3)*modA[1:4])})
#wb <- sapply(b,function(x){2^sum(c(1,x,x^2,x^3)*modB[1:4])})
#
#wa <- wa/(length(wa)*modA[5])
#wb <- wb/(length(wb)*modB[5])
#
#plot(density(r),ylim=c(0,1e10),xlab="GC Density"); lines(density(a,weights=wa),col=3);
lines(density(b,weights=wb),col=2)

```

- La cita de las patentes, solicitudes de patente, publicaciones y documentos anteriores no es una admisión de que ninguno de los anteriores es técnica anterior pertinente, ni constituye ninguna admisión en cuanto al contenido o fecha de estas publicaciones o documentos. El término “un(o)” o “una” puede referirse a uno o a una pluralidad de los elementos que modifica (por ejemplo, “un reactivo” puede significar uno o más reactivos) a menos que se aclare contextualmente que se describe o bien uno de los elementos o bien más de uno de los elementos. El término “aproximadamente”, tal como se usa en el presente documento, se refiere a un valor dentro de 10 % del parámetro subyacente (por ejemplo, más o menos el 10 %), y el uso del término “aproximadamente” al comienzo de una cadena de valores modifica cada uno de los valores (por ejemplo, “aproximadamente 1, 2 y 3” se refiere a aproximadamente 1, aproximadamente 2 y aproximadamente 3). Por ejemplo, un peso de “aproximadamente 100 gramos” puede incluir pesos entre 90 gramos y 110 gramos.

REIVINDICACIONES

1. Un método para determinar la presencia o ausencia de una aneuploidía cromosómica para una muestra que comprende:

(a) filtrar, según una distribución de densidad de lectura, porciones de un genoma de referencia, lo que proporciona de este modo un perfil de densidad de lectura para una muestra de prueba que comprende densidades de lectura de porciones filtradas, en donde:

(i) las densidades de lectura comprenden mediciones cuantitativas de recuentos de lecturas de secuencia mapeadas en porciones de un genoma de referencia, en donde las lecturas de secuencia son lecturas de ácido nucleico circulante libre de células de una muestra de prueba de una mujer embarazada, y

(ii) la distribución de densidad de lectura es una distribución de densidades de lectura promedio, media o en mediana, y se determina para densidades de lectura de porciones para múltiples muestras;

(b) ajustar el perfil de densidad de lectura para la muestra de prueba al 1) restar un valor de mediana entrenado de las densidades de lectura de la muestra de prueba y 2) eliminar los componentes de las densidades de lectura de la muestra de prueba que se correlacionan con uno o más componentes principales del perfil, cuyos componentes principales (i) se obtienen a partir de un conjunto de entrenamiento de muestras euploides conocidas mediante un análisis de componentes principales, y (ii) representan uno o más sesgos en un perfil de densidad de lectura, lo que proporciona de este modo un perfil de densidad de lectura ajustado que comprende densidades de lectura ajustadas, en donde una pluralidad de sesgos se elimina del perfil de densidad de lectura ajustada;

(c) comparar el perfil de densidad de lectura ajustado con un perfil de referencia que comprende densidades de lectura obtenidas de una o más muestras de referencia, lo que proporciona de este modo una comparación; y

(d) determinar la presencia o ausencia de una aneuploidía cromosómica para la muestra de prueba según la comparación, en donde uno o más o todos los métodos de procesamiento se realizan por un procesador, un microprocesador, un ordenador, junto con memoria y/o por un aparato controlado por microprocesador.

2. El método de la reivindicación 1, en donde el perfil de densidad de lectura se ajusta en (b) por 2 a 10 componentes principales.

3. El método de la reivindicación 1 o 2, en donde la comparación en (c) comprende determinar un nivel de significancia, en donde el nivel de significancia indica una diferencia estadísticamente significativa entre el perfil de densidad de lectura ajustado y el perfil de referencia, y se determina la presencia de una aneuploidía cromosómica.

4. El método de la reivindicación 3, en donde determinar el nivel de significancia comprende determinar un valor p.

5. El método de una cualquiera de las reivindicaciones 1 a 4, en donde el perfil de referencia en (c) comprende:

(i) densidades de lectura determinadas para una o más muestras euploides conocidas;

(ii) densidades de lectura de porciones filtradas; y/o

(iii) densidades de lectura ajustadas según el uno o más componentes principales.

6. El método según una cualquiera de las reivindicaciones 1 a 5, en donde:

(i) las densidades de lectura de las porciones para las múltiples muestras en (a)(ii) son las densidades de lectura de la mediana;

(ii) las densidades de lectura de las porciones filtradas para la muestra de prueba son las densidades de lectura de la mediana; y/o

(iii) las densidades de lectura para el perfil de referencia en (c) son medianas de las densidades de lectura.

7. El método de la reivindicación 6, en donde las densidades de lectura para el perfil de densidad de lectura para la muestra de prueba, las densidades de lectura de porciones para las múltiples muestras en (a)(ii), y las densidades de lectura para el perfil de referencia en (c) se determinan según un proceso que comprende el uso de una estimación de densidad de núcleo.

8. El método de la reivindicación 6 o 7, en donde el perfil de densidad de lectura para la muestra de prueba se determina según las densidades de lectura de la mediana para la muestra de prueba; y el perfil de referencia en (c) se determina según las densidades de lectura de la mediana para las una o más muestras de referencia.
9. El método de una cualquiera de las reivindicaciones 1 a 8, en donde las porciones de un genoma de referencia se filtran en (a) según una medida de incertidumbre para la distribución de densidad de lectura.
10. El método de la reivindicación 9, en donde la medida de incertidumbre es una MAD.
11. El método de una cualquiera de las reivindicaciones 1 a 10, en donde el perfil de densidad de lectura para la muestra de prueba y el perfil de referencia en (c) comprenden puntuaciones z de densidades de lectura.
12. El método de una cualquiera de las reivindicaciones 1 a 11, en donde el perfil de densidad de lectura para la muestra de prueba es representativo de la dosificación cromosómica para la muestra de prueba, y el método comprende comparar la dosificación cromosómica para el perfil de densidad de lectura para la muestra de prueba con la dosificación cromosómica para el perfil de referencia en (c), lo que genera de este modo una comparación de dosificación cromosómica, en donde determinar la presencia o ausencia de una aneuploidía cromosómica para la muestra de prueba es según la comparación de dosificación cromosómica.
13. El método de una cualquiera de las reivindicaciones 1 a 12, en donde determinar la presencia o ausencia de una aneuploidía cromosómica para la muestra de prueba comprende identificar la presencia o ausencia de una copia de un cromosoma, dos copias de un cromosoma, tres copias de un cromosoma, cuatro copias de un cromosoma, cinco copias de un cromosoma, una delección de uno o más segmentos de un cromosoma o una inserción de uno o más segmentos de cromosoma.
14. El método de una cualquiera de las reivindicaciones 1 a 13, que comprende, antes de (a), obtener las lecturas de secuencias.
15. El método de una cualquiera de las reivindicaciones 1 a 14, que comprende, antes de (b), ajustar el perfil de densidad de lectura para la muestra de prueba mediante una normalización de LOESS de GC.

Figura 1

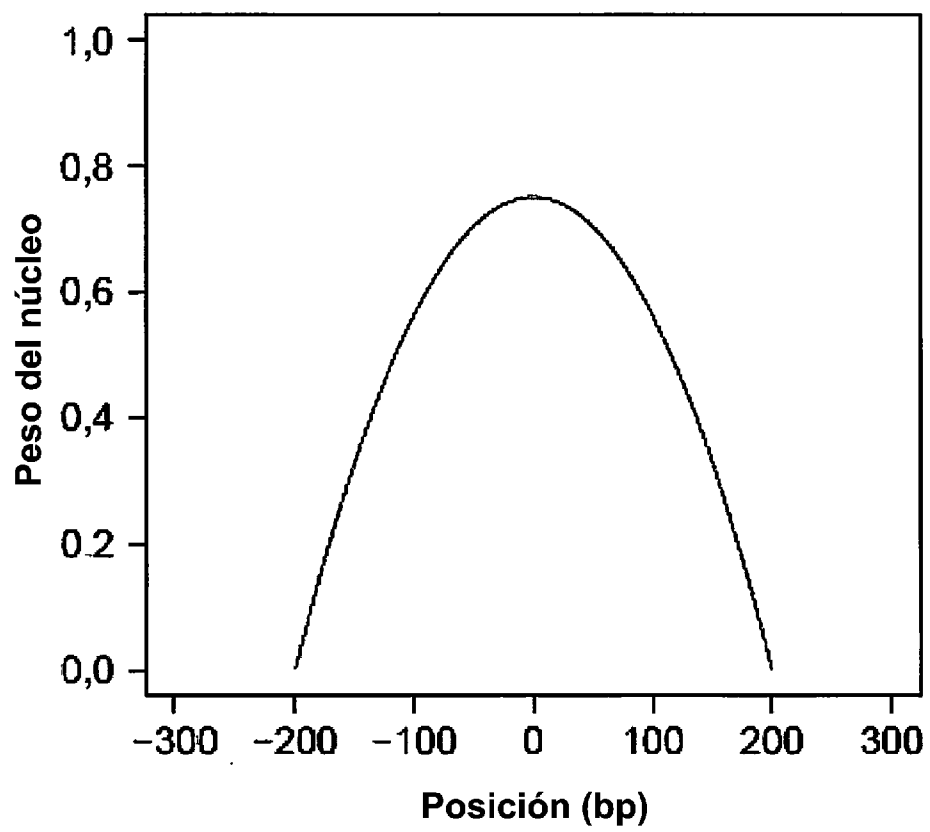


Figura 2

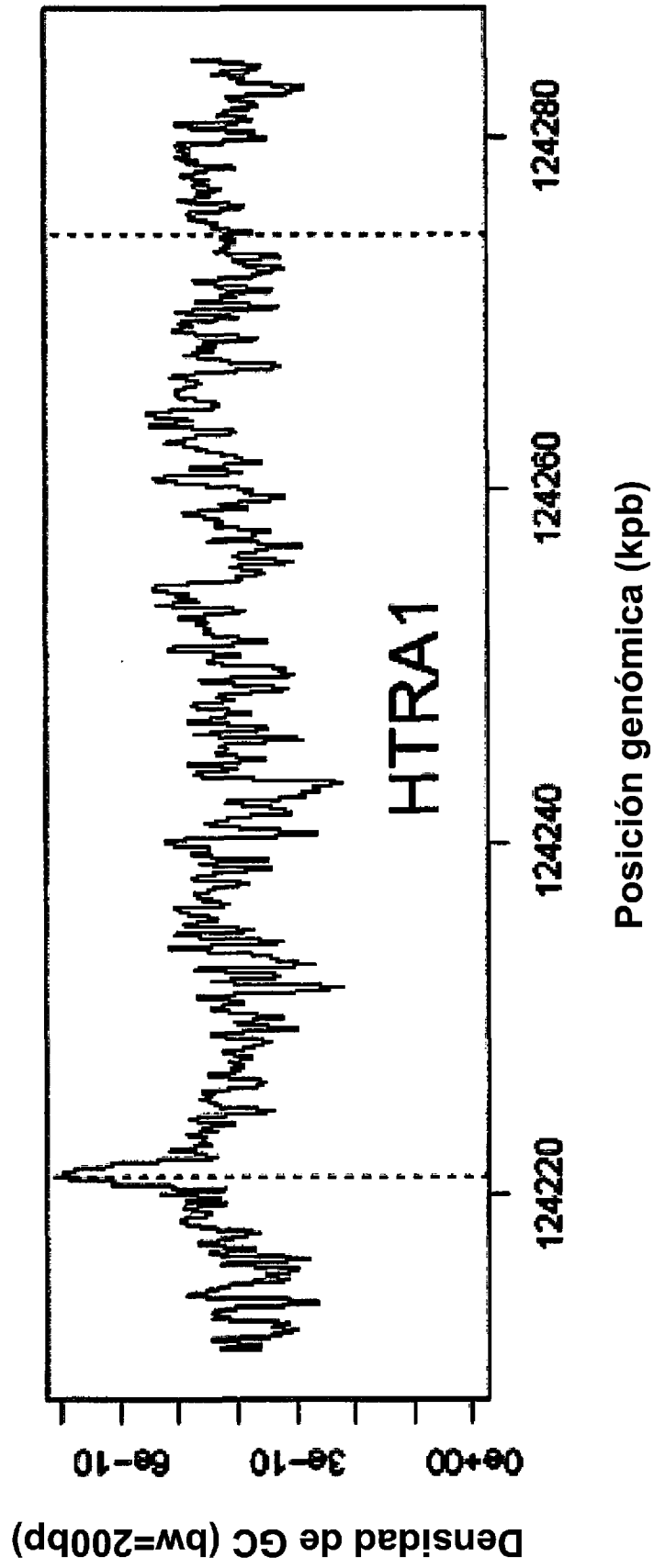


Figura 3

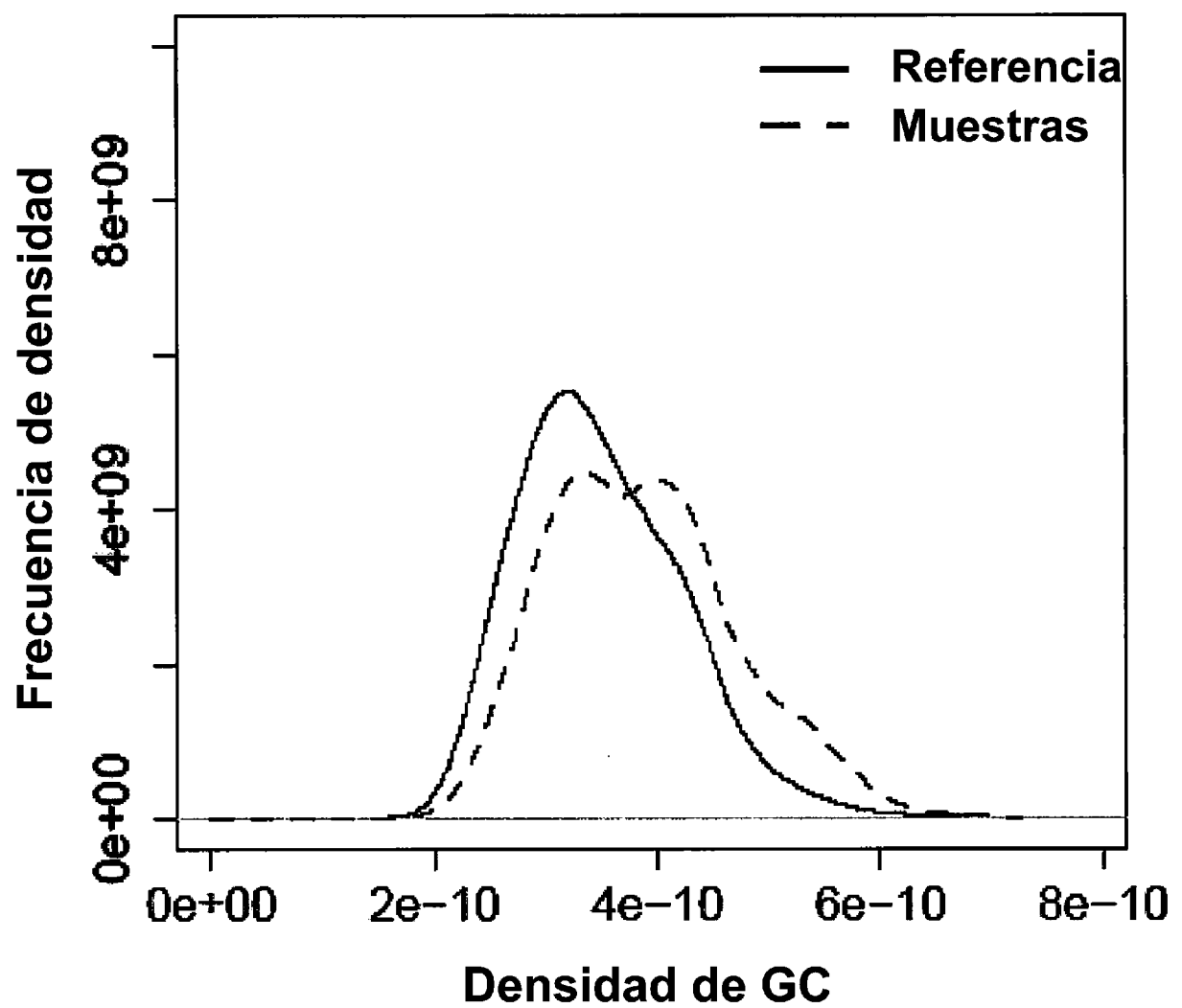


Figura 4

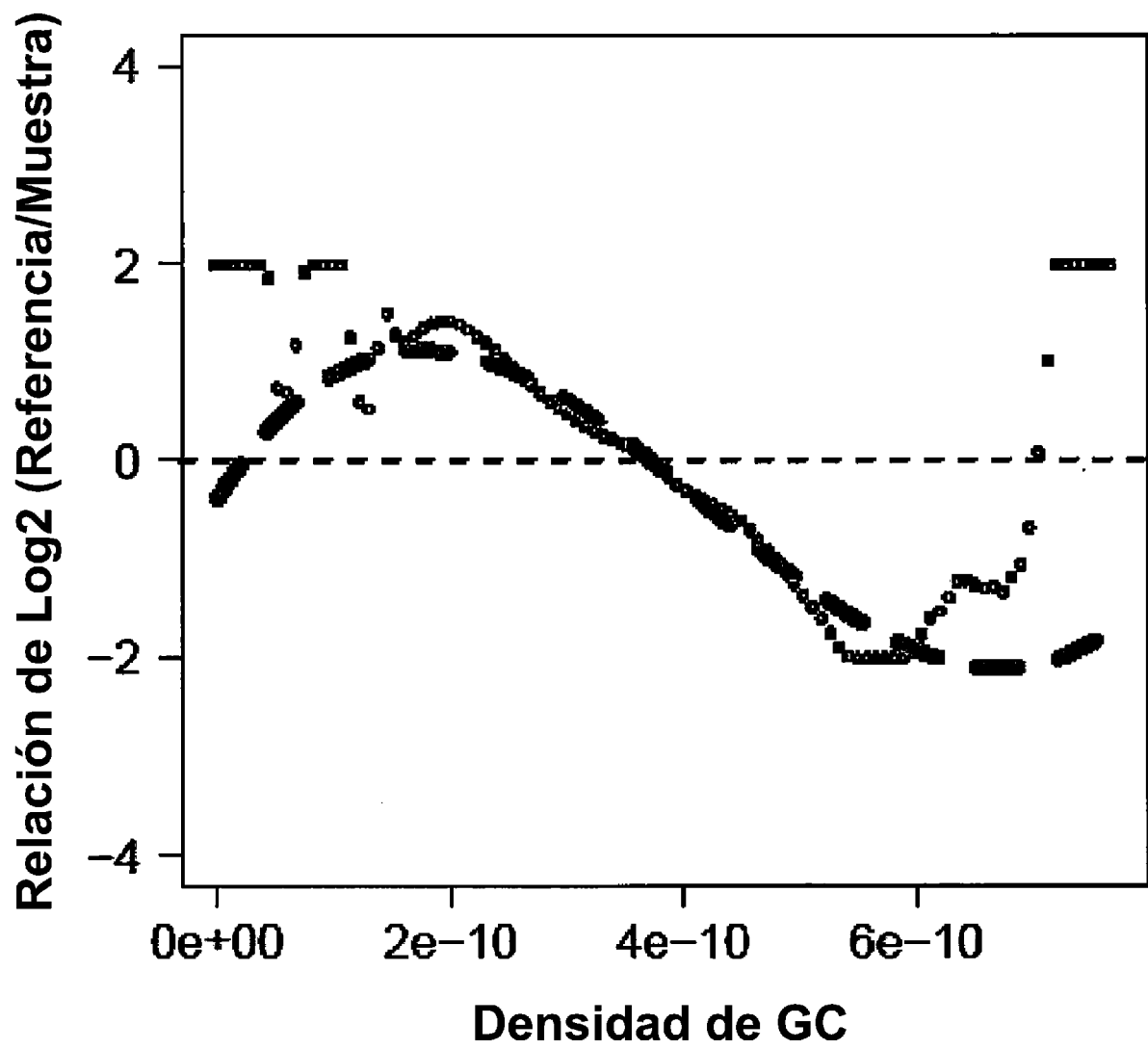


Figura 5A

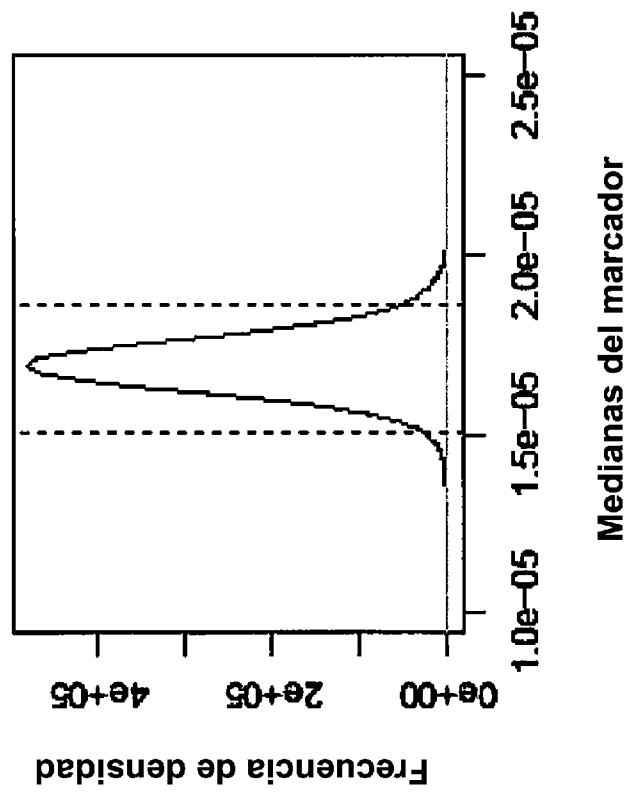


Figura 5B

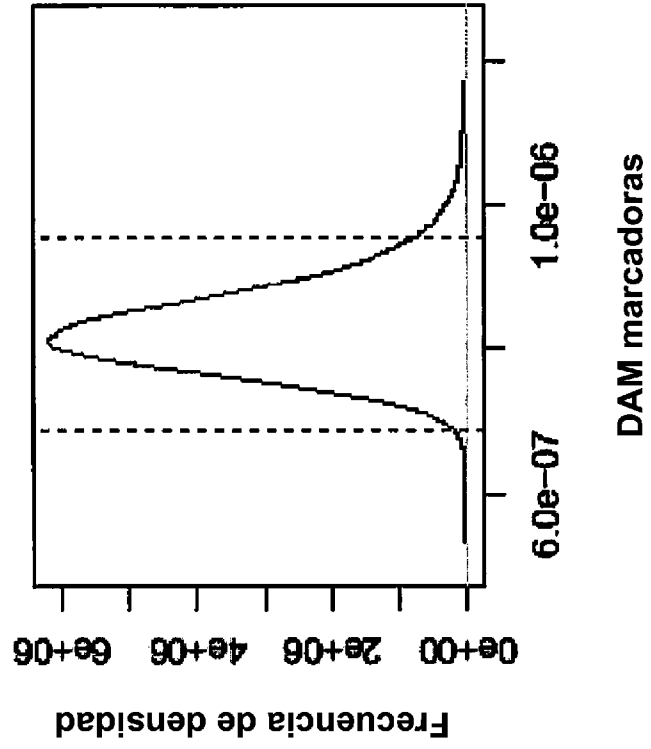


Figura 6A

Mediana

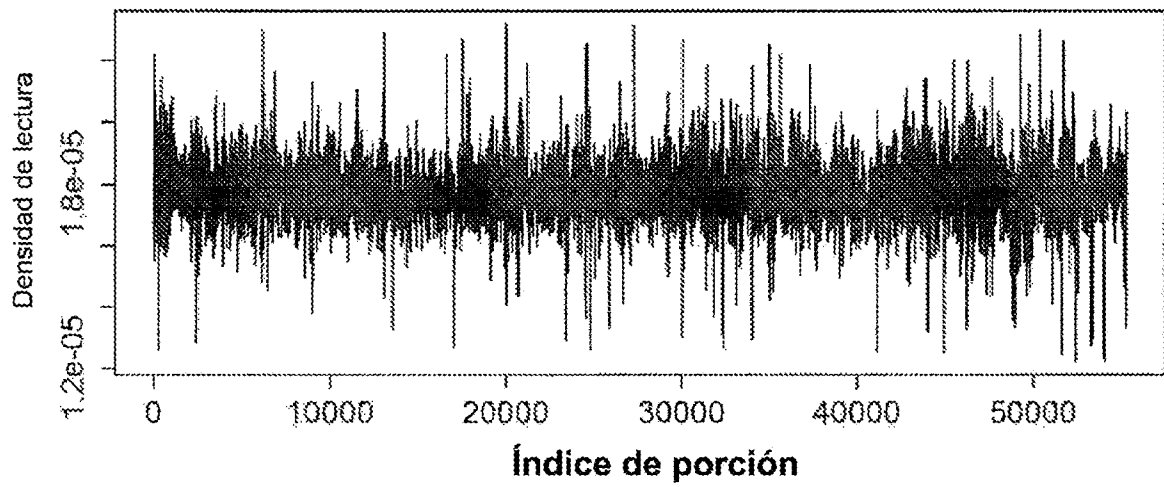


Figura 6B

PC1

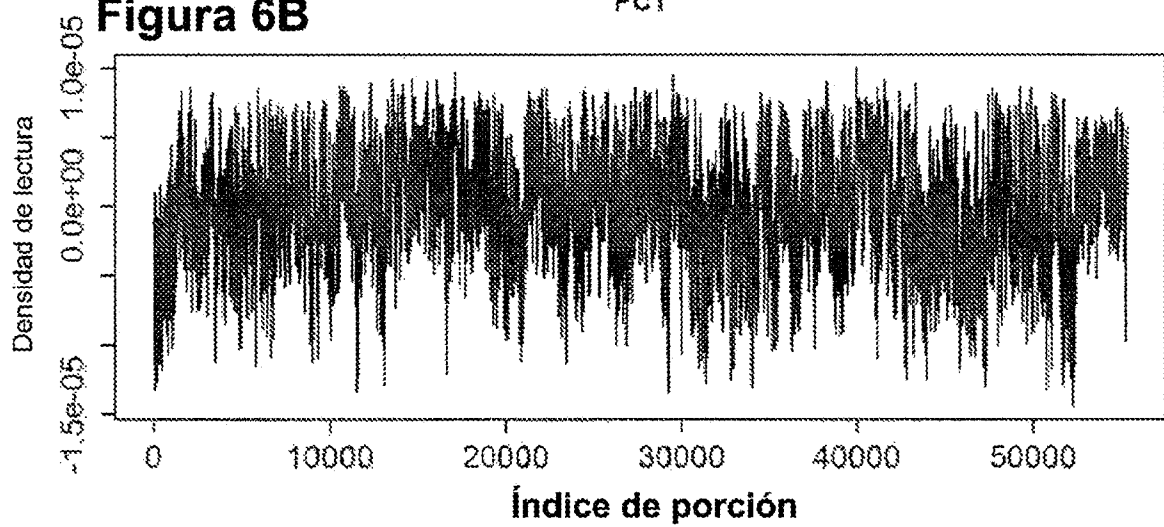
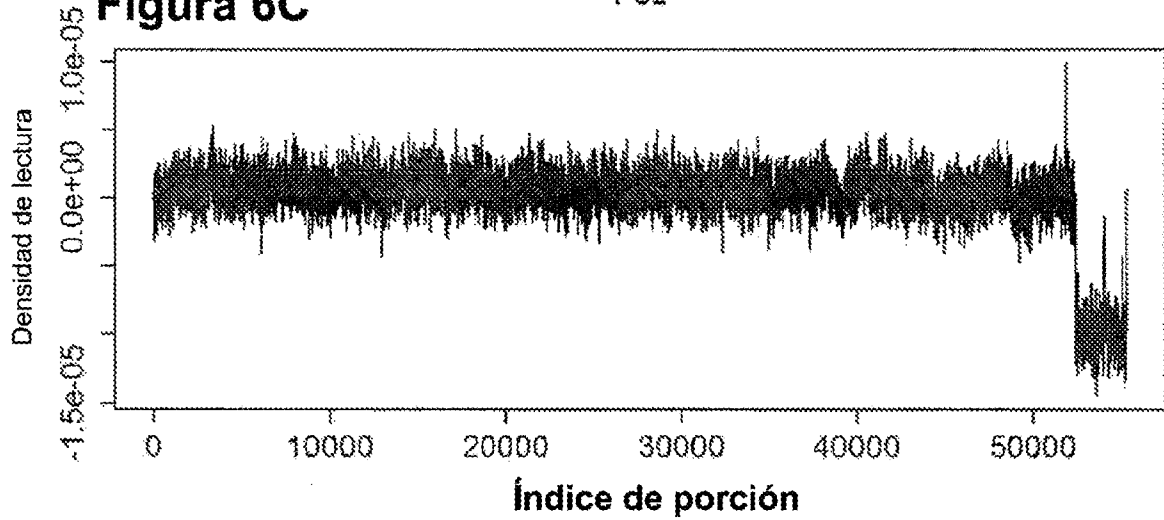


Figura 6C

PC2



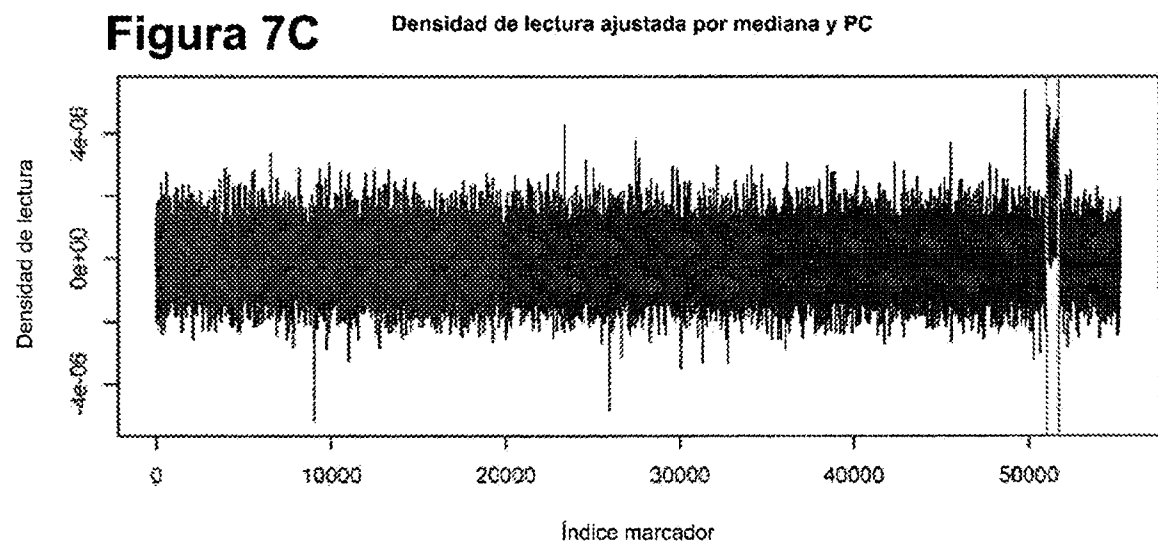
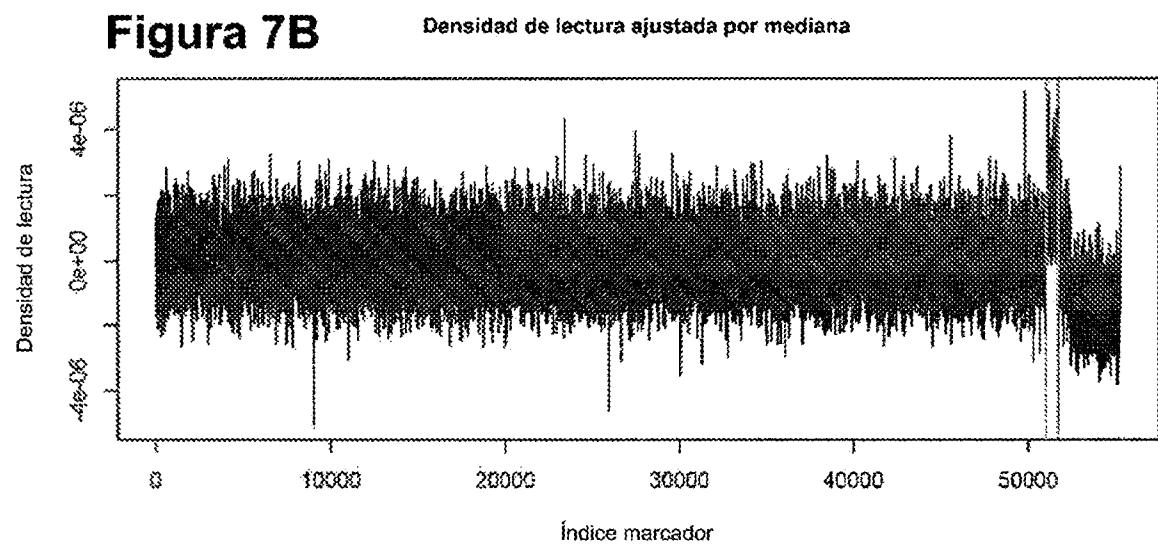
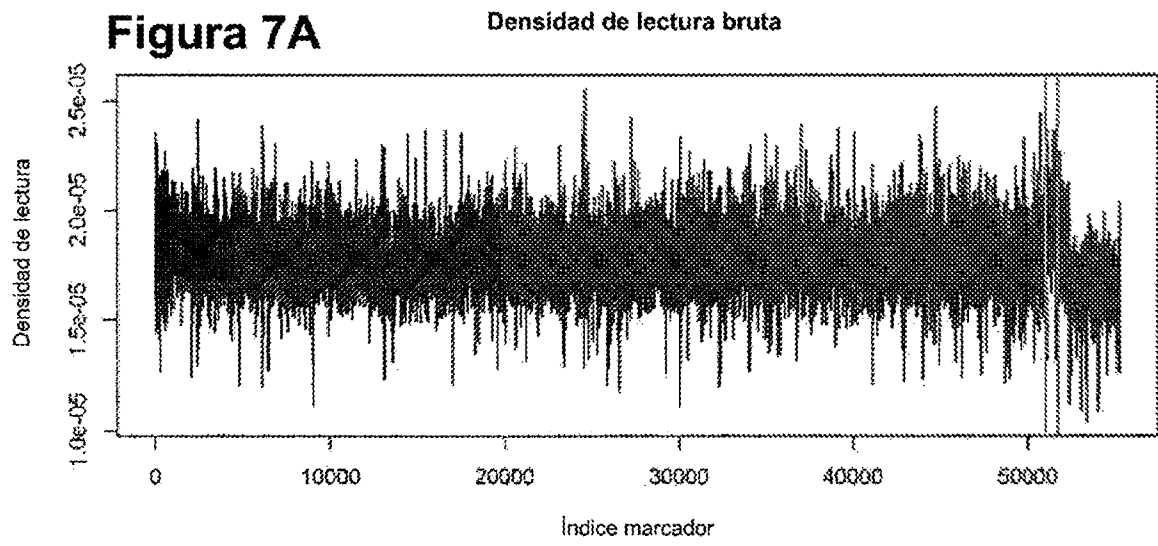


Figura 8

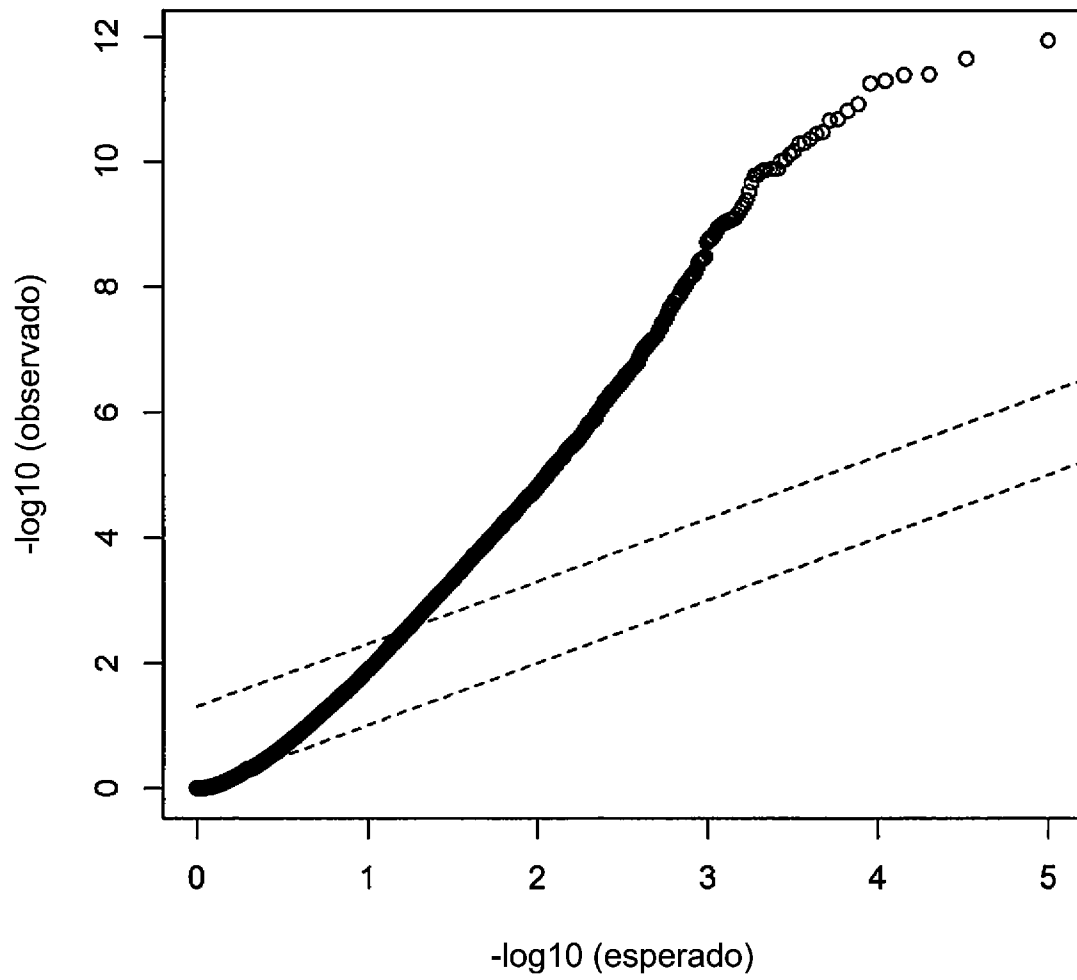


Figura 9A

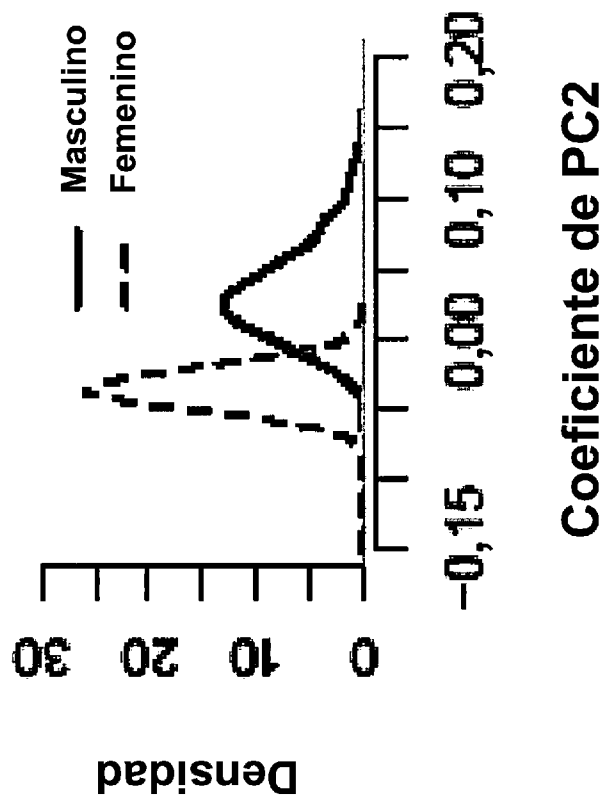


Figura 9B

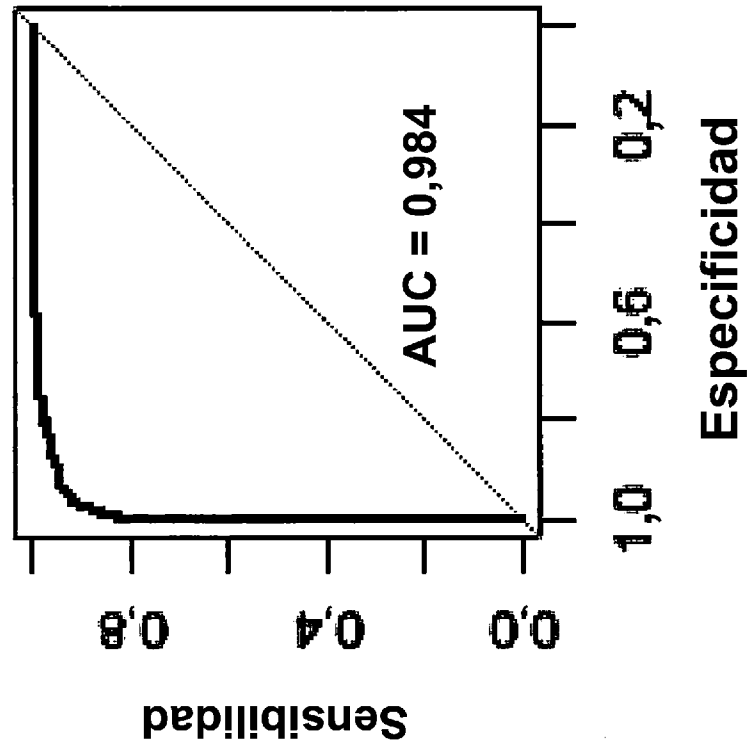


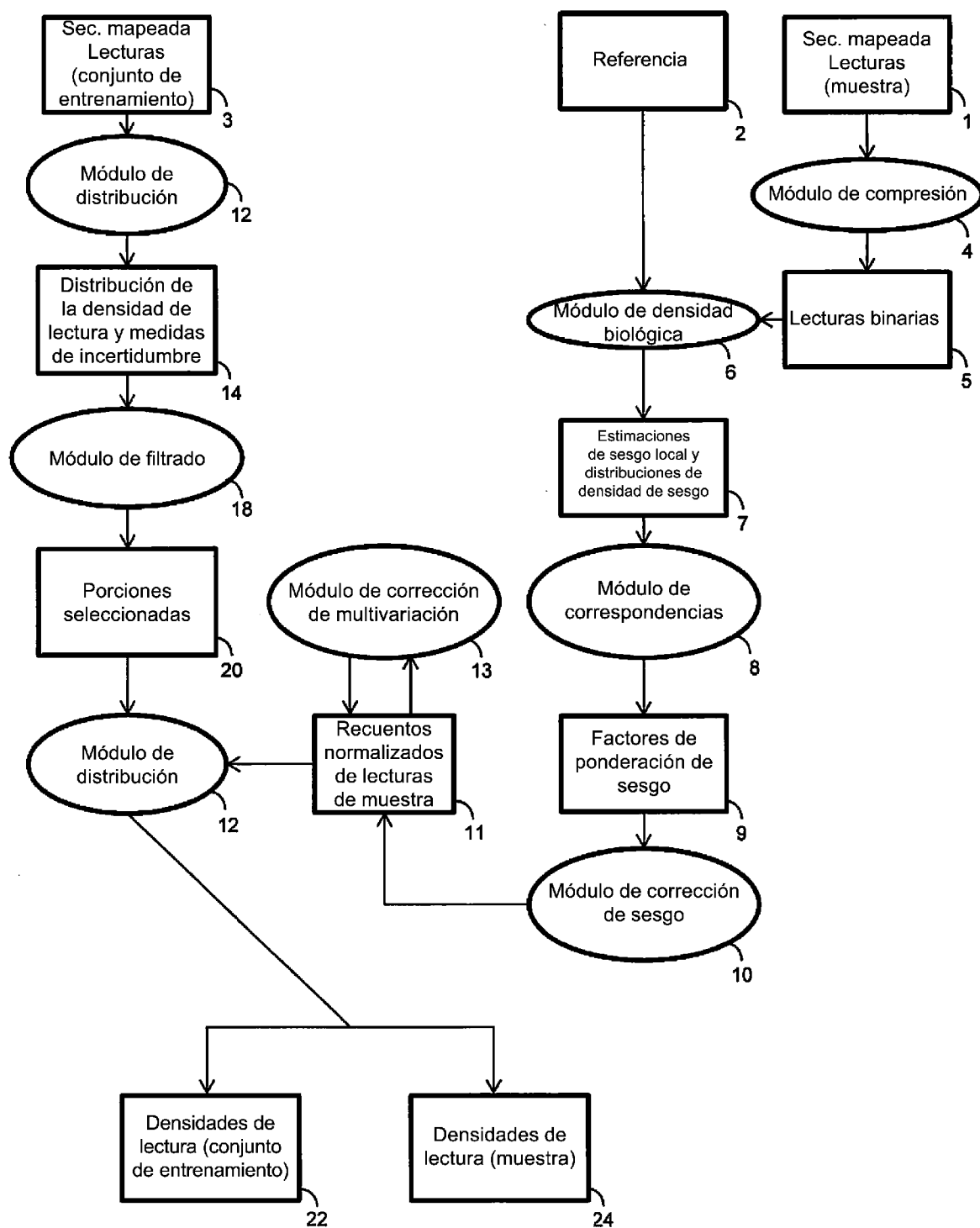
Figura 10A

Figura 10B

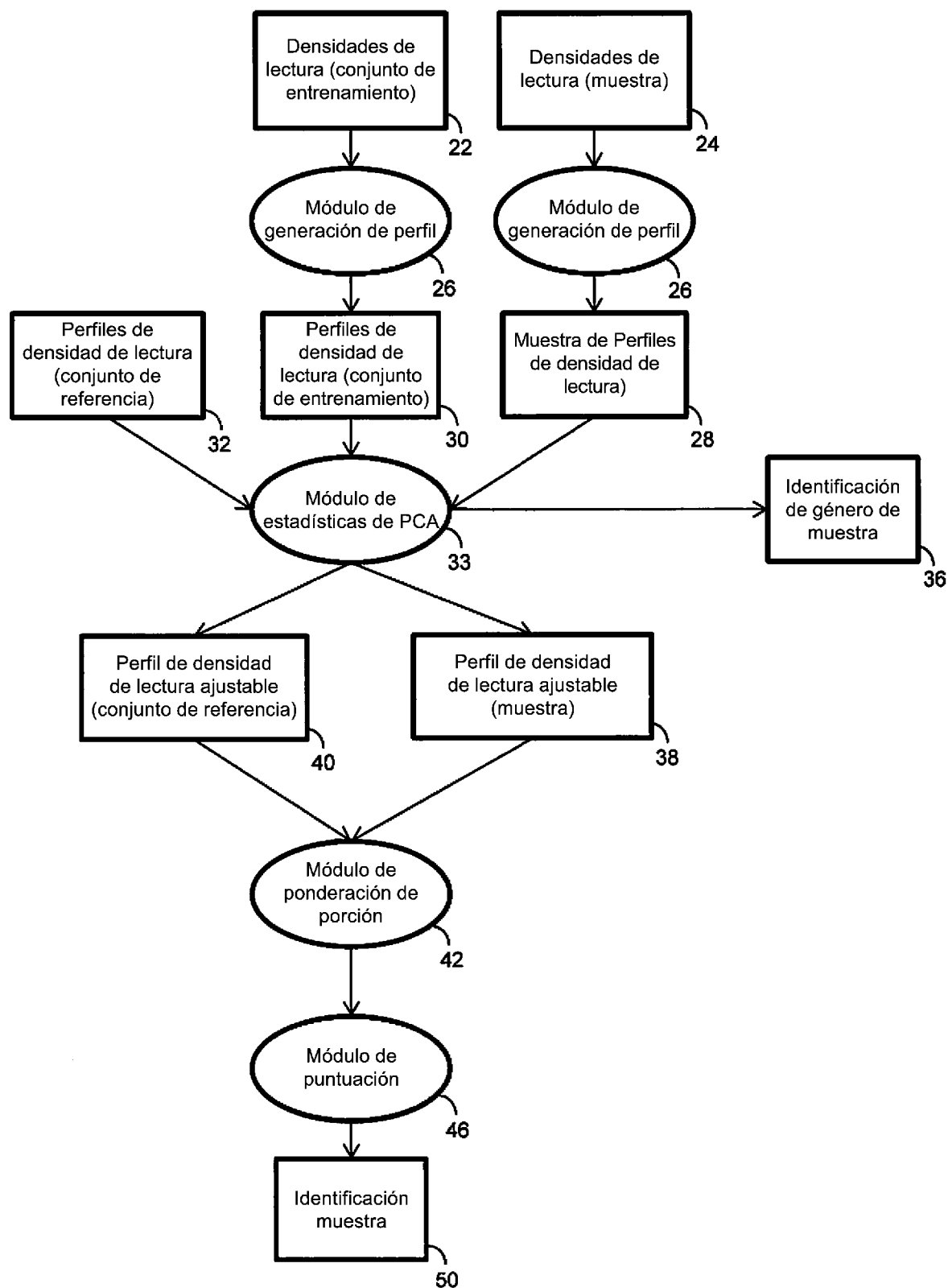


Figura 11

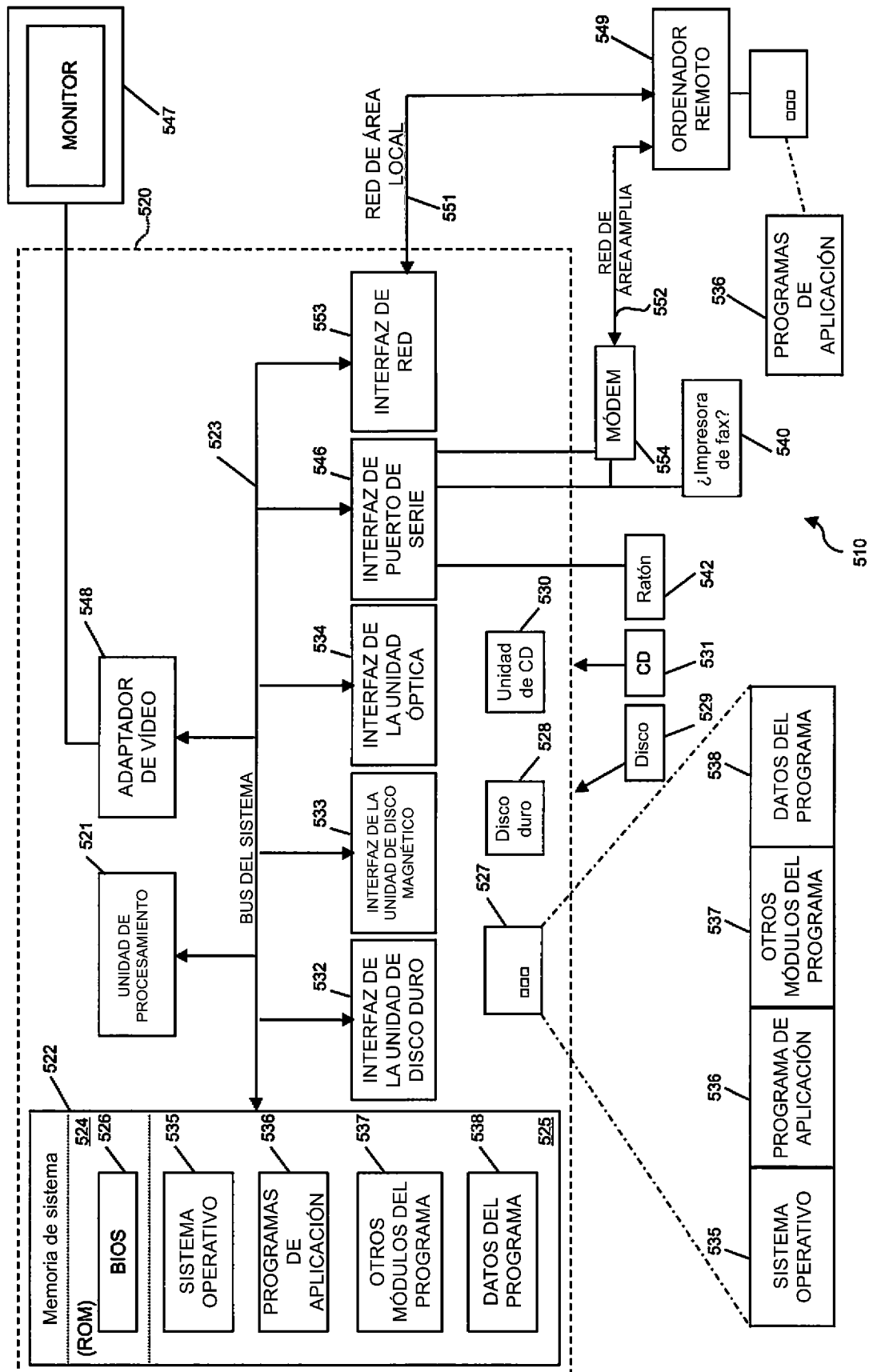


Figura 12