(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 981 092**

(51) Int. Cl.:

C12Q 1/6869 (2008.01)
C12Q 1/68 (2008.01)
C12M 1/00 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **10.01.2018 PCT/CN2018/072045**

(87) Fecha y número de publicación internacional: **02.08.2018 WO18137496**

(96) Fecha de presentación y número de la solicitud europea: **10.01.2018 E 18743980 (7)**

(97) Fecha y número de publicación de la concesión europea: **27.03.2024 EP 3575407**

(54) Título: **Procedimiento para determinar la proporción de ácidos nucleicos acelulares procedentes de una fuente predeterminada en una muestra biológica**

(30) Prioridad:

24.01.2017 CN 201710055200

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
07.10.2024

(73) Titular/es:

BGI GENOMICS CO., LTD. (100.0%)
Floors 7-14, Building No. 7, BGI Park, No. 21
Hongan 3rd Street, Yantian District
Shenzhen, Guangdong Province 518083, CN

(72) Inventor/es:

YUAN, YUYING;
CHAI, XIANGHUA;
WANG, SHUYUAN;
CHEN, LINA;
ZHOU, LIJUN;
LIU, QIANG;
ZHANG, HONGYUN;
WANG, WEI;
LIU, NA y
YIN, YE

(74) Agente/Representante:

GONZÁLEZ PECES, Gustavo Adolfo

DESCRIPCIÓN

Procedimiento para determinar la proporción de ácidos nucleicos acelulares procedentes de una fuente predeterminada en una muestra biológica

Campo

La presente divulgación se refiere al campo de la biotecnología, especialmente a pruebas genéticas prenatales no invasivas y pruebas de oncogenes, y más específicamente a un procedimiento para determinar una fracción de ácidos nucleicos acelulares procedentes de una fuente predeterminada en una muestra biológica.

Antecedentes

Desde 1977, los investigadores han descubierto sucesivamente ADN derivado del cáncer en la sangre periférica de pacientes con tumores, y también han confirmado la presencia de ADN fetal libre circulante (cff, "cell-free fetal") en el plasma de una mujer embarazada. La detección o la estimación de una fracción de ADN derivado del cáncer en la sangre periférica de pacientes con tumores o la de ADN fetal libre circulante en el plasma de la mujer embarazada, es decir, la determinación de una fracción de ácidos nucleicos acelulares procedentes de una fuente predeterminada en una muestra biológica, es de gran importancia.

Sin embargo, el procedimiento actual para determinar una fracción de ácidos nucleicos acelulares procedentes de una fuente predeterminada en una muestra biológica sigue siendo mejorable.

Sumario

Las realizaciones de la presente divulgación pretenden resolver, al menos en cierta medida, uno de los problemas existentes en la técnica relacionada. Para ello, un objeto de la presente divulgación es proporcionar un procedimiento capaz de determinar con precisión y eficacia una fracción de ácidos nucleicos acelulares procedentes de una fuente predeterminada en una muestra biológica.

Cabe señalar que las soluciones técnicas de la presente divulgación se logran mediante los siguientes descubrimientos.

En la actualidad, la fracción de ADN fetal libre circulante en la sangre periférica se estima principalmente mediante las siguientes indicaciones: 1) diferente nivel de metilación para un biomarcador concreto entre los fragmentos de ADN materno y los fragmentos de ADN fetal libre circulante en células mononucleares de sangre periférica; 2) selección de una pluralidad de sitios representativos de polimorfismo de un solo nucleótido ("single nucleotide polymorphism", SNP), cada uno de los cuales es diferente entre los fragmentos de ADN materno y los fragmentos de ADN fetal libre circulante; 3) diferencia de tamaño entre los fragmentos de ADN fetal y los fragmentos de ADN materno en la circulación sanguínea materna; 4) estimación de la fracción de ADN fetal de una mujer embarazada con un feto masculino mediante el procedimiento del cromosoma Y. Sin embargo, estas cuatro indicaciones tienen limitaciones individuales. En concreto, la primera requiere una gran cantidad de plasma; la segunda requiere la captura con sondas y una gran profundidad de secuenciación o información genética paterna; la tercera necesita determinar las longitudes tanto de los fragmentos de ADN fetal como de los fragmentos de ADN materno, lo que conlleva un elevado coste de secuenciación y una precisión ordinaria; y la última es aplicable para estimar la fracción de ADN fetal sólo para la mujer embarazada con feto masculino, pero no es válido para una mujer embarazada con feto femenino.

Con el fin de superar las limitaciones mencionadas de estos procedimientos, los inventores han desarrollado un procedimiento para estimar la fracción de ADN fetal que sólo utiliza los datos de secuenciación detectados actualmente por las pruebas prenatales no invasivas ("Non-Invasive Prenatal Testing", NIPT), sin datos de secuenciación adicionales. Es decir, este procedimiento es capaz de cuantificar con precisión la fracción de ADN fetal en la sangre periférica mediante datos de secuenciación de baja cobertura. La investigación y el desarrollo de este procedimiento se basan principalmente en el descubrimiento de que el número de lecturas en cada una de las ventanas, obtenidas por división secuencial de un autosoma por una longitud determinada, está correlacionado con la fracción de ADN fetal. Gracias a este descubrimiento, la fracción de ADN fetal puede estimarse con gran precisión no sólo para el feto masculino, sino también para el feto femenino.

Los inventores han descubierto, a través de investigaciones posteriores, que este procedimiento tiene amplias aplicaciones y puede utilizarse para detectar ADN libre procedente de diferentes fuentes. Por ejemplo, el procedimiento es adecuado para determinar la fracción de ácidos nucleicos acelulares derivados de tumores o de ácidos nucleicos acelulares no derivados de tumores en una muestra de sangre periférica obtenida de un sujeto que padece un tumor, del que se sospecha que padece un tumor o que está sometido a una detección tumoral, y también es posible obtener resultados precisos y fiables mediante datos de secuenciación de baja cobertura.

Por lo tanto, según realizaciones de un aspecto de la presente divulgación, se proporciona un procedimiento para determinar una fracción de ácidos nucleicos acelulares procedentes de una fuente predeterminada en una muestra biológica. En realizaciones de la presente divulgación, el procedimiento incluye: (1) realizar la secuenciación de los ácidos nucleicos acelulares contenidos en la muestra biológica, a fin de obtener un resultado de la secuenciación

consistente en una pluralidad de lecturas de secuenciación; (2) alinear el resultado de la secuenciación con una secuencia de referencia, a fin de determinar el número de lecturas de secuenciación que quedan dentro de una ventana predeterminada en el resultado de la secuenciación; y (3) determinar la fracción de ácidos nucleicos acelulares procedentes de la fuente predeterminada en la muestra biológica en función del número de lecturas de secuenciación que quedan dentro de la ventana predeterminada.

Los inventores han descubierto, sorprendentemente, que la fracción de los ácidos nucleicos acelulares procedentes de la fuente predeterminada en la muestra biológica, especialmente la fracción de los ácidos nucleicos fetales acelulares en una muestra de sangre periférica obtenida de una mujer embarazada, y la fracción de los ácidos nucleicos acelulares derivados de tumores en una muestra de sangre periférica obtenida de un sujeto que padece un tumor, del que se sospecha que padece un tumor o que está sometido a una detección tumoral, puede determinarse con precisión y eficacia mediante el procedimiento de la presente divulgación.

También se divulga, pero no forma parte de la invención, un dispositivo para determinar una fracción de ácidos nucleicos acelulares procedentes de una fuente predeterminada en una muestra biológica. El dispositivo incluye: un aparato de secuenciación, configurado para realizar la secuenciación de los ácidos nucleicos acelulares contenidos en la muestra biológica, con el fin de obtener un resultado de la secuenciación consistente en una pluralidad de lecturas de secuenciación; un aparato de recuento, conectado al aparato de secuenciación y configurado para alinear el resultado de la secuenciación con una secuencia de referencia, con el fin de determinar el número de lecturas de secuenciación que quedan dentro de una ventana predeterminada en el resultado de la secuenciación; y un aparato de determinación de la fracción de ácidos nucleicos acelulares, conectado al aparato de recuento y configurado para determinar la fracción de ácidos nucleicos acelulares procedentes de la fuente predeterminada en la muestra biológica en función del número de lecturas de secuenciación que quedan dentro de la ventana predeterminada.

El dispositivo de la presente divulgación es adecuado para llevar a cabo el procedimiento para determinar una fracción de ácidos nucleicos acelulares procedentes de una fuente predeterminada en una muestra biológica como se describe anteriormente, mediante el cual la fracción de los ácidos nucleicos acelulares procedentes de la fuente predeterminada en la muestra biológica, especialmente la fracción de ácidos nucleicos fetales acelulares en una muestra de sangre periférica obtenida de una mujer embarazada, o la fracción de ácidos nucleicos acelulares derivados de tumores en una muestra de sangre periférica obtenida de un sujeto que padece un tumor, del que se sospecha que padece un tumor o que está sometido a una detección tumoral, puede determinarse de forma precisa y eficiente.

Otros aspectos y ventajas de las realizaciones de la presente divulgación se ofrecerán en parte en las siguientes descripciones, se harán evidentes en parte a partir de las siguientes descripciones, o se aprenderán de la práctica de las realizaciones de la presente divulgación.

Breve descripción de los dibujos

Estos y otros aspectos y ventajas de las realizaciones de la presente divulgación se harán evidentes y se apreciarán más fácilmente a partir de las siguientes descripciones hechas con referencia a los dibujos, en los que:

la figura 1 es un diagrama de flujo de un procedimiento para determinar una fracción de ácidos nucleicos acelulares procedentes de una fuente predeterminada en una muestra biológica según una realización de la presente divulgación;

la figura 2 es un diagrama de flujo para determinar el número de lecturas de secuenciación que quedan dentro de una ventana predeterminada en el resultado de la secuenciación según una realización de la presente divulgación;

la figura 3 es un diagrama esquemático de un dispositivo para determinar una fracción de ácidos nucleicos acelulares procedentes de una fuente predeterminada en una muestra biológica;

la figura 4 es un diagrama esquemático de un aparato de recuento 200 según una realización de la presente divulgación,

la figura 5 es un gráfico de distribución de pesos obtenido mediante la estimación de pesos utilizando un modelo estadístico de regresión contraída según una realización de la presente divulgación;

la figura 6 es un gráfico que muestra el análisis de correlación entre las fracciones fetales de las muestras de un conjunto de prueba estimadas mediante un modelo estadístico de regresión contraída y las fracciones fetales estimadas utilizando el cromosoma Y según una realización de la presente divulgación, en la que Masculino representa una muestra biológica obtenida de una mujer embarazada con un feto masculino, Femenino representa una muestra biológica obtenida de una mujer embarazada con un feto masculino, Basado en CrY representa una fracción de ácidos nucleicos fetales acelulares estimada basándose en el cromosoma Y, y Regresión contraída representa una fracción de ácidos nucleicos fetales acelulares estimada por el modelo estadístico de regresión contraída según una realización de la presente divulgación;

la figura 7 es un gráfico de distribución de pesos obtenido mediante la estimación de pesos utilizando un modelo de red neuronal según una realización de la presente divulgación;

la figura 8 es un gráfico que muestra el análisis de correlación entre las fracciones fetales de las muestras de un conjunto de prueba estimadas por un modelo de red neuronal y las fracciones fetales estimadas utilizando el cromosoma Y según una realización de la presente divulgación, en la que Masculino representa una muestra biológica obtenida de una mujer embarazada con un feto masculino, Femenino representa una muestra biológica obtenida de una mujer embarazada con un feto femenino, Basado en CrY representa una fracción de ácidos nucleicos fetales acelulares estimada basándose en el cromosoma Y, y FF-QuantSC representa una fracción de ácidos nucleicos fetales acelulares estimada por el modelo de red neuronal según una realización de la presente divulgación; y

la figura 9 es un gráfico que muestra el análisis de correlación entre las fracciones fetales femeninas estimadas por un modelo estadístico de regresión contraída y las estimadas por un modelo de red neuronal según una realización de la presente divulgación, en la que la Regresión contraída representa una fracción de ácidos nucleicos fetales acelulares estimada por el modelo estadístico de regresión contraída según una realización de la presente divulgación, y FF-QuantSC representa una fracción de ácidos nucleicos fetales acelulares estimada por el modelo de red neuronal según una realización de la presente divulgación.

Descripción detallada

Las realizaciones de la presente divulgación se describirán en detalle a continuación, las cuales son explicativas, ilustrativas y se utilizan para explicar de forma general la presente divulgación, por lo que no se interpretará que limitan la presente divulgación.

Procedimiento para determinar una fracción de ácidos nucleicos acelulares en una muestra biológica

Según algunas realizaciones de un primer aspecto de la presente divulgación, se proporciona un procedimiento para determinar una fracción de ácidos nucleicos acelulares procedentes de una fuente predeterminada en una muestra biológica. Los inventores han descubierto, sorprendentemente, que la fracción de ácidos nucleicos acelulares en la muestra biológica, especialmente la fracción de ácidos nucleicos fetales acelulares en una muestra de sangre periférica obtenida de una mujer embarazada, y la fracción de ácidos nucleicos derivados de tumores en una muestra de sangre periférica obtenida de un sujeto que padece un tumor pueden determinarse con precisión y eficacia mediante el procedimiento de la presente divulgación.

Cabe señalar que la expresión "fracción de ácidos nucleicos acelulares procedentes de una fuente predeterminada en una muestra biológica" utilizada en el presente documento se refiere a una fracción del número de ácidos nucleicos acelulares procedentes de una fuente específica con respecto al número total de ácidos nucleicos acelulares en la muestra biológica. Por ejemplo, si la muestra biológica es sangre periférica obtenida de una mujer embarazada, y los ácidos nucleicos acelulares procedentes de la fuente predeterminada son ácidos nucleicos fetales acelulares, "una fracción de ácidos nucleicos acelulares procedentes de una fuente predeterminada en una muestra biológica" es una fracción de ácidos nucleicos fetales acelulares, lo que significa una proporción del número de moléculas de ácidos nucleicos fetales acelulares con respecto al número total de moléculas de ácidos nucleicos acelulares en la sangre periférica obtenida de la mujer embarazada, que a veces también puede denominarse como "una fracción de concentración de ADN fetal libre circulante en la sangre periférica obtenida de la mujer embarazada" o "una proporción de ADN fetal libre circulante" o "un porcentaje de ADN fetal libre circulante". Como otro ejemplo, si la muestra biológica es una muestra de sangre periférica obtenida de un sujeto que padece un tumor, del que se sospecha que padece un tumor o que está sometido a una detección tumoral, y los ácidos nucleicos acelulares procedentes de la fuente predeterminada son ácidos nucleicos acelulares derivados de tumores, "una fracción de ácidos nucleicos acelulares procedentes de la fuente predeterminada en una muestra biológica" es una fracción de ácidos nucleicos acelulares derivados de tumores, entendiéndose por tal una proporción del número de ácidos nucleicos acelulares derivados de tumores con respecto al número total de ácidos nucleicos acelulares en la muestra de sangre periférica obtenida del sujeto que padece un tumor, del que se sospecha que padece un tumor o que está sometido a una detección tumoral. De acuerdo con las realizaciones de la presente divulgación y con referencia a la figura 1, el procedimiento incluye las siguientes etapas.

S100: Secuenciación de ácidos nucleicos

Los ácidos nucleicos acelulares de la muestra biológica se secuencian para obtener un resultado de la secuenciación consistente en una pluralidad de lecturas de secuenciación.

En realizaciones de la presente divulgación, la muestra biológica es una muestra de sangre periférica. En realizaciones de la presente divulgación, el ácido nucleico acelular procedente de la fuente predeterminada es al menos uno seleccionado de entre: ácidos nucleicos fetales acelulares en una muestra de sangre periférica obtenida de una mujer embarazada; ácidos nucleicos maternos acelulares en una muestra de sangre periférica obtenida de una mujer embarazada; y ácidos nucleicos acelulares derivados de tumores o ácidos nucleicos acelulares no derivados de tumores en una muestra de sangre periférica obtenida de un sujeto que padece un tumor, del que se sospecha que

padece un tumor o que está sometido a una detección tumoral. Por lo tanto, se puede determinar fácilmente la fracción de ácidos nucleicos fetales acelulares o de ácidos nucleicos maternos acelulares en la muestra de sangre periférica obtenida de la mujer embarazada, o la fracción de ácidos nucleicos acelulares derivados de tumores en la muestra de sangre periférica obtenida del sujeto que padece un tumor, del que se sospecha que padece un tumor o que está sometido a una detección tumoral. En algunas realizaciones de la presente divulgación, los ácidos nucleicos acelulares de la muestra biológica se secuencian mediante secuenciación de extremos pareados, secuenciación de extremo único o secuenciación de molécula única. En algunas realizaciones específicas de la presente divulgación, los ácidos nucleicos acelulares son ADN. Cabe señalar que la expresión "datos de secuenciación" utilizado en el presente documento se refiere a "lecturas de secuenciación", que corresponden a moléculas de ácido nucleico sometidas a secuenciación.

S200: Determinación del número de lecturas de secuenciación que quedan dentro de una ventana predeterminada

El resultado de la secuenciación se alinea con una secuencia de referencia, para determinar el número de lecturas de secuenciación que quedan dentro de una ventana predeterminada en el resultado de la secuenciación.

En una realización de la presente divulgación, la secuencia de referencia es una secuencia genómica de referencia, preferentemente hg19.

En una realización de la presente divulgación, la ventana predeterminada se obtiene por división secuencial de un cromosoma predeterminado de la secuencia genómica de referencia.

En una realización de la presente divulgación, el cromosoma predeterminado incluye un autosoma. Preferentemente, el autosoma no incluye ninguno de los cromosomas 13, 18 ni 21.

En una realización de la presente divulgación, la ventana predeterminada tiene una longitud de 60 Kpb.

Cabe destacar que la división de las ventanas predeterminadas debe mantener la uniformidad de las lecturas dentro de cada ventana, es decir, asegurar la uniformidad de las lecturas dentro de cada ventana. Debe mencionarse que la "uniformidad de lecturas dentro de cada ventana" significa que el número de lecturas en cada ventana es sustancialmente el mismo, es decir, que la varianza entre ellas es cercana a 0.

Según algunas realizaciones de la presente divulgación, refiriéndose a la figura 2, S200 incluye además lo siguiente.

En S210, el resultado de la secuenciación se alinea con la secuencia genómica de referencia. Específicamente, el resultado de la secuenciación se alinea con la secuencia genómica de referencia con el fin de construir un conjunto de datos consistente en una pluralidad de lecturas de secuenciación cartografiadas inequívocamente, en el que cada lectura de secuenciación cartografiada inequívocamente en el conjunto de datos puede cartografiarse a una única posición en la secuencia genómica de referencia.

En S220, se determina la posición de cada lectura de secuenciación cartografiada inequívocamente en la secuencia genómica de referencia. En concreto, se determinan las posiciones de las lecturas de secuenciación cartografiadas inequívocamente individuales en la secuencia genómica de referencia.

En S230, se determina el número de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de la ventana predeterminada. En concreto, se determina el número de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de ventanas individuales predeterminadas.

Por lo tanto, es fácil determinar el número de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de la ventana predeterminada en el resultado de la secuenciación, y el resultado determinado es preciso, fiable y reproducible.

S300: Determinación de la fracción de ácidos nucleicos acelulares

La fracción de ácidos nucleicos acelulares procedentes de la fuente predeterminada en la muestra biológica se determina en función del número de lecturas de secuenciación que quedan dentro de la ventana predeterminada.

En realizaciones de la presente divulgación, en S300, la fracción de los ácidos nucleicos acelulares procedentes de la fuente predeterminada en la muestra biológica se determina por un peso de cada ventana predeterminada. En algunas realizaciones específicas, en S300, el peso de cada ventana predeterminada se predetermina con muestras de entrenamiento. Por lo tanto, el resultado es preciso, fiable y repetible. En algunas realizaciones, el peso de cada ventana predeterminada se determina mediante al menos uno de un modelo estadístico de regresión contraída y un modelo de red neuronal. En algunas realizaciones, el modelo de red neuronal adopta un sistema de aprendizaje TensorFlow. En algunas realizaciones específicas, el sistema de aprendizaje TensorFlow incluye los siguientes parámetros: los números de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro en ventanas individuales de un autosoma como capa de entrada; una fracción fetal como capa de salida; ReLu como neurona; un optimizador seleccionado de al menos uno de Adam, SGD y Ftrl, preferentemente Ftrl. Preferentemente, el sistema de aprendizaje TensorFlow incluye además los siguientes parámetros: una tasa de aprendizaje fijada en

0,002; 1 capa oculta; y 200 neuronas en la capa oculta. Por lo tanto, el resultado es preciso y fiable. Debe mencionarse que el término "peso" utilizado en el presente documento es un concepto relativo para un determinado índice. Un peso de un determinado índice se refiere a la importancia relativa del índice en la evaluación global. Por ejemplo, "un peso de una ventana predeterminada" se refiere a una importancia relativa de la ventana predeterminada en todas las ventanas predeterminadas. El "peso de conexión" se refiere a la importancia relativa de una determinada conexión de dos capas en todas las conexiones de dos capas.

En realizaciones de la presente divulgación, la muestra de entrenamiento es una muestra de sangre periférica con una fracción conocida de ácidos nucleicos fetales acelulares procedente de una mujer embarazada. Por lo tanto, puede determinarse eficazmente una fracción de ácidos nucleicos fetales acelulares o de ácidos nucleicos maternos acelulares en una muestra de sangre periférica procedente de una mujer embarazada sometida a detección. En algunas realizaciones específicas, la muestra de entrenamiento es una muestra de sangre periférica con una fracción conocida de ácidos nucleicos fetales acelulares procedente de una mujer embarazada con un feto masculino normal. Por lo tanto, el resultado determinado para la fracción de ácidos nucleicos fetales acelulares o de ácidos nucleicos maternos acelulares en una muestra de sangre periférica procedente de una mujer embarazada sometida a detección es más preciso y fiable.

En algunas realizaciones de la presente divulgación, el peso de cada ventana predeterminada se determina mediante un modelo estadístico de regresión contraída que tiene la siguiente fórmula computacional:

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{n} \hat{\beta}_j x_j ,$$

en la que $\hat{y}$ es una fracción fetal predictiva, x_j es el número de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de una ventana, β_j es un peso de una ventana, β_0 es una desviación, y β_j y β_0 se obtienen entrenando el modelo.

En algunas realizaciones de la presente divulgación, el peso de cada ventana predeterminada se determina mediante un modelo de red neuronal que tiene la siguiente fórmula computacional:

$$z_j^l = f\left(\sum_k w_{jk}^l z_k^{l-1} + b_j^l\right) ,$$

en la que *l* es un número de serie de una capa en el modelo de red neuronal, la primera capa es una capa de entrada, la última capa es una capa de salida (que tiene una sola neurona), y una capa intermedia es una capa oculta,

z_j^l es un valor para una *j*-ésima neurona en una *l*-ésima capa, z_k^{l-1} es un valor para una *k*-ésima neurona en una *(l-1)*-ésima capa, w_{jk}^l es un peso de conexión desde la *k*-ésima neurona en la *(l-1)*-ésima capa a la *j*-ésima neurona en la *l*-ésima capa, b_j^l es una desviación de entrada para la *j*-ésima neurona en la *l*-ésima capa, y w y b se obtienen entrenando el modelo. Una forma muy común de la función *f* es una unidad lineal rectificada, es decir, $f(x) = \text{máx}(0,x)$.

En algunas realizaciones, cuando se aplica el modelo de red neuronal, los valores para las neuronas individuales se calculan capa por capa basándose en la fórmula computacional anterior del modelo de red neuronal, y un valor para la neurona en la última capa es la fracción fetal predictiva.

Es decir, según realizaciones de la presente divulgación, determinar el peso de cada ventana predeterminada por el modelo de red neuronal incluye: calcular valores para neuronas individuales capa por capa según la fórmula computacional del modelo de red neuronal, en la que el valor para la neurona en la última capa es la fracción fetal predictiva.

En algunas realizaciones, antes de S300, se realiza una corrección de GC en el número de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de la ventana predeterminada por adelantado, para obtener el número de lecturas de secuenciación cartografiadas inequívocamente corregidas para GC que quedan dentro de la ventana predeterminada. Por lo tanto, el número determinado de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de la ventana predeterminada es preciso y fiable.

Preferentemente, en algunas realizaciones, la corrección de GC se lleva a cabo mediante:

ajustar los números de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de ventanas predeterminadas individuales del cromosoma predeterminado a los contenidos de GC correspondientes para determinar $ER = f(gc)$;

realizar una corrección del número de lecturas de secuenciación cartografiadas inequívocamente para cada ventana predeterminada de todos los cromosomas: $ERA_i = ER_i * (ER/f(GC_i))$, $i = 1,2,3,\ldots ,N$,

en la que para una muestra, ER_i representa el número de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de una *i*-ésima ventana predeterminada, GC_i representa un contenido de GC de una secuencia de referencia para la *i*-ésima ventana predeterminada, *ER* representa un valor medio para los números de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de ventanas individuales del cromosoma predeterminado; y ERA_i representa el número de lecturas de secuenciación cartografiadas inequívocamente corregidas para GC en la *i*-ésima ventana predeterminada después de la corrección.

En realizaciones de la presente divulgación, antes de S300, se predetermina el sexo del feto. Preferentemente, el sexo del feto se determina mediante una proporción entre el número de lecturas de secuenciación cartografiadas inequívocamente en el cromosoma Y y el número total de lecturas de secuenciación cartografiadas inequívocamente en todos los cromosomas. Por lo tanto, el resultado determinado para la fracción de ácidos nucleicos fetales acelulares o de ácidos nucleicos maternos acelulares en una muestra de sangre periférica procedente de una mujer embarazada sometida a detección es más preciso y fiable.

Además, según algunas realizaciones específicas de la presente divulgación, la estimación de una fracción fetal de una muestra secuenciada con el procedimiento de la presente divulgación utilizando el modelo estadístico de regresión contraída o el modelo de red neuronal en un banco TensorFlow incluye las siguientes etapas específicas.

1. Estimación de la fracción fetal de la muestra secuenciada utilizando el modelo estadístico de regresión contraída:

1) Dividir secuencialmente una secuencia de referencia (hg19) en ventanas adyacentes de una longitud fija (tal como 60 Kpb en esta realización), filtrar las ventanas en una región N y determinar el contenido de GC de cada ventana, para obtener un archivo de ventana de referencia hg19.gc;

2) Alineación: alinear las lecturas de secuenciación (cada una de 28 pb) secuenciadas por secuenciación de extremo único ("single-end", SE) en la plataforma CG con la secuencia de referencia (hg19) utilizando, por ejemplo, BWA V0.7.7-r441;

3) Filtración y estadísticas preliminares: seleccionar las lecturas de secuenciación cartografiadas inequívocamente a partir de los resultados de la alineación, filtrar las lecturas repetidas y las lecturas con bases no coincidentes para obtener lecturas de secuenciación eficaces, y determinar el número de lecturas de secuenciación eficaces que quedan dentro de cada ventana del archivo de ventana de referencia hgl9.gc y los contenidos de GC correspondientes;

4) Corrección de GC, que incluye las siguientes etapas específicas:

ajustar los números de lecturas de secuenciación eficaces que quedan dentro de ventanas individuales de un autosoma a los correspondientes contenidos de GC (tal como usando un ajuste de interpolación cúbica en esta realización) para determinar ER = f(gc);

realizar la corrección en el número de lecturas de secuenciación eficaces para cada ventana de todos los cromosomas: $ERA_t = ER_i * (ER/f(GC_i))$, $i = 1,2,3,\ldots , N$,

en la que, para una muestra, ER_i representa el número de lecturas de secuenciación eficaces que quedan dentro de una *i*-ésima ventana, GC_i representa un contenido de GC de la secuencia de referencia para la *i*-ésima ventana (registros en el archivo de ventana de referencia hg19.gc), *ER* representa un valor medio para los números de lecturas de secuenciación eficaces que quedan dentro de ventanas individuales de un autosoma (tales como los cromosomas 1-22); y ERA_i representa el número de lecturas de secuenciación eficaces corregidas para GC en la *i*-ésima ventana después de la corrección de GC;

5) Determinar el sexo de un feto: comparar una proporción (%ER) del número de lecturas de secuenciación eficaces corregidas para GC en el cromosoma Y al número total de lecturas de secuenciación eficaces corregidas para GC en todos los cromosomas con un umbral especificado *a* dentro de un intervalo de [0,001, 0,003], determinar que el feto es un feto masculino si la %ER es mayor o igual que *a*, determinar en caso contrario que el feto es un feto femenino si la %ER es menor que *a*;

6) Estimar la fracción fetal mediante el modelo estadístico de regresión contraída, que incluye las siguientes etapas:

a) seleccionar muestras obtenidas cada una de ellas independientemente de una mujer embarazada con un feto masculino como conjunto de entrenamiento; y seleccionar un lote de muestras como conjunto de prueba (en el que el número de muestras obtenidas cada una de ellas

independientemente de una mujer embarazada con un feto masculino puede ser el mismo que el de una mujer embarazada con un feto femenino);

b) estimar el peso de cada ventana en el autosoma (excluidos los cromosomas 13, 18 y 21) utilizando el modelo de regresión contraída (el peso equivale a un coeficiente de regresión β en el modelo de regresión contraída) para el conjunto de entrenamiento, a fin de obtener una distribución de pesos estimada; y

c) estimar la fracción fetal de la muestra en el conjunto de prueba con los pesos conocidos.

2. Estimación de la fracción fetal de la muestra secuenciada utilizando el modelo de red neuronal del banco TensorFlow:

Las cinco primeras etapas 1)-5) son las mismas que las descritos anteriormente en el modelo estadístico de regresión contraída;

6) Estimar la fracción fetal utilizando el modelo de red neuronal del banco TensorFlow, que incluye las siguientes etapas específicas:

a) seleccionar muestras, cada una de ellas obtenida independientemente de una mujer embarazada con un feto de sexo masculino, como conjunto de entrenamiento; y seleccionar un lote de muestras como conjunto de prueba (en el que el número de muestras, cada una de ellas obtenida independientemente de una mujer embarazada con un feto de sexo masculino, puede ser el mismo que el de una mujer embarazada con un feto de sexo femenino), normalizar todos los datos (todos los números de lecturas de secuenciación eficaces corregidas para GC que quedan dentro de ventanas individuales), es decir, transformar linealmente cada variable (el número de lecturas de secuenciación eficaces corregidas para GC que quedan dentro de cada ventana), de modo que el valor medio de las variables en todas las muestras sea 0 y la desviación estándar sea 1;

b) construir una red neuronal en la que los números normalizados de lecturas de secuenciación eficaces que quedan dentro de ventanas individuales de un autosoma relativamente estable se toman como capa de entrada, una sola neurona se toma como capa de salida (correspondiente a la fracción fetal), no hay capa oculta, se selecciona ReLU como tipo de neurona y Adam como optimizador;

c) aplicar la red neuronal (aprendizaje) en el conjunto de entrenamiento para predecir la fracción fetal, y ajustar una tasa de aprendizaje de acuerdo con el cambio del efecto de aprendizaje en cada ronda, con el fin de maximizar la tasa de aprendizaje al tiempo que se garantiza el efecto de aprendizaje para el conjunto de entrenamiento sin repetir la fluctuación;

d) entrenar tantas rondas como permita la capacidad de cálculo, hasta que se sature el efecto de aprendizaje;

e) cambiar a otros optimizadores (tales como SGD, Ftrl, etc.), repitiendo las etapas b)-d), y seleccionar un optimizador óptimo basándose en los efectos de aprendizaje;

f) intentar añadir un término de regularización de segundo orden al modelo de red neuronal y ajustar su tamaño para observar los efectos de aprendizaje antes y después de añadir y ajustar el tamaño del término de regularización de segundo orden;

g) añadir una capa oculta, ajustar el número de neuronas en la capa oculta, repetir las etapas b)-f) y seleccionar una arquitectura de capa oculta óptima basándose en los efectos de aprendizaje;

h) entrenar el modelo de red neuronal optimizado en el conjunto de entrenamiento para obtener parámetros óptimos y una distribución de pesos (pesos promedio de las neuronas en la capa de entrada a la capa oculta) de ventanas individuales; y

i) estimar la fracción fetal de la muestra en el conjunto de prueba con el modelo de red neuronal entrenado.

Para facilitar la comprensión, a continuación se presentan brevemente los principios básicos de los modelos descritos en la presente divulgación.

(1) Modelo estadístico de regresión contraída

La regresión contraída es un procedimiento de mínimos cuadrados modificado, que reduce el sobreajuste del modelo añadiendo el término de regularización de segundo orden.

En forma matemática, el procedimiento de mínimos cuadrados consiste en resolver $\beta_0, \beta_1, \beta_2, \ldots$, que minimizan una suma de cuadrados de residuales:

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

en la que RSS es la suma de cuadrados de los residuales, y_i es una variable dependiente y x_{ij} es una variable independiente.

Si se modifica la función anterior añadiendo el término de regularización de segundo orden, se obtiene la regresión contraída, que consiste en resolver β_0, β_1, β_2, ... para minimizar la siguiente función:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2 ,$$

en la que es necesario especificar λ , y una práctica general para especificar λ es tomar varios valores y determinar qué λ minimiza la función objetivo del conjunto de validación mediante validación cruzada.

(2) Red neuronal artificial

La red neuronal artificial (es decir, el modelo de red neuronal) es un procedimiento de aprendizaje automático no lineal. Sus elementos básicos son neuronas, y cada neurona realiza una media ponderada con desviación sobre varias entradas x_j:

$$z = \sum_j (w_j x_j + w_0) ,$$

en la que w_j es un peso, y w_0 es la desviación,

$f(z)$ es una salida que depende del resultado de la media ponderada. En la actualidad, la forma de función más utilizada es la unidad lineal rectificada, es decir, $f(z) = máx(0,z)$.

Varias neuronas constituyen una red multicapa, en la que la primera capa (es decir, la capa de entrada) toma los números normalizados de lecturas de secuenciación eficaces que quedan dentro de ventanas predeterminadas individuales como variable independiente para la entrada, mientras que una salida de una capa anterior se toma como entrada para una capa posterior, actuando así hasta la última capa (es decir, la capa de salida) que tiene una sola neurona y una salida (es decir, un valor predicho por el modelo). Una capa distinta de las capas de entrada y salida se denomina capa oculta.

Los parámetros básicos del modelo de red neuronal incluyen el peso de cada capa y la desviación, que generalmente se entrenan por retropropagación. Además, hay otros parámetros, como la tasa de aprendizaje, el tipo de neurona, el algoritmo de optimización (optimizador), el número de capas ocultas, el número de neuronas en la capa oculta, el coeficiente de regularización, etc., que suelen preestablecerse según la experiencia y ajustarse repetidamente en función de los efectos del entrenamiento.

Dispositivo para determinar una fracción de ácidos nucleicos acelulares en una muestra biológica

Según algunas realizaciones de un segundo aspecto de la presente divulgación, también se proporciona y divulga, aunque no forma parte de la invención, un dispositivo para determinar una fracción de ácidos nucleicos acelulares procedentes de una fuente predeterminada en una muestra biológica. Los inventores han descubierto, sorprendentemente, que el dispositivo de la presente divulgación es adecuado para llevar a cabo el procedimiento para determinar una fracción de ácidos nucleicos acelulares procedentes de una fuente predeterminada en una muestra biológica como se describe anteriormente, por el cual la fracción de los ácidos nucleicos acelulares procedentes de la fuente predeterminada en la muestra biológica, especialmente la fracción de ácidos nucleicos fetales acelulares o de ácidos nucleicos maternos acelulares en una muestra de sangre periférica obtenida de una mujer embarazada, o la fracción de ácidos nucleicos acelulares derivados de tumores en una muestra de sangre periférica obtenida de un sujeto que padece un tumor, del que se sospecha que padece un tumor o que está sometido a una detección tumoral, puede determinarse con precisión y eficacia.

En referencia a la figura 3, el dispositivo incluye: un aparato de secuenciación 100, un aparato de recuento 200 y un aparato de determinación de la fracción de ácidos nucleicos acelulares 300.

Específicamente, el aparato de secuenciación 100 está configurado para realizar la secuenciación de los ácidos nucleicos acelulares contenidos en la muestra biológica, con el fin de obtener un resultado de la secuenciación consistente en una pluralidad de lecturas de secuenciación. El aparato de recuento 200 está conectado al aparato de secuenciación 100 y está configurado para alinear el resultado de la secuenciación con una secuencia de referencia,

a fin de determinar el número de lecturas de secuenciación que quedan dentro de una ventana predeterminada en el resultado de la secuenciación. El aparato de determinación de la fracción de ácidos nucleicos acelulares 300 está conectado al aparato de recuento 200 y está configurado para determinar la fracción de ácidos nucleicos acelulares procedentes de la fuente predeterminada en la muestra biológica basándose en el número de lecturas de secuenciación que quedan dentro de la ventana predeterminada.

Según algunas realizaciones de la presente divulgación, la muestra biológica no se limita a un tipo específico. En una realización específica, la muestra biológica es una muestra de sangre periférica. En una realización, el ácido nucleico acelular procedente de la fuente predeterminada es al menos uno seleccionado de entre los siguientes: ácidos nucleicos fetales acelulares en una muestra de sangre periférica obtenida de una mujer embarazada; ácidos nucleicos maternos acelulares en una muestra de sangre periférica obtenida de una mujer embarazada; y ácidos nucleicos acelulares derivados de tumores o ácidos nucleicos acelulares no derivados de tumores en una muestra de sangre periférica obtenida de un sujeto que padece un tumor, del que se sospecha que padece un tumor o que está sometido a una detección tumoral. Por lo tanto, se puede determinar fácilmente la fracción de ácidos nucleicos fetales acelulares o de ácidos nucleicos maternos acelulares en la muestra de sangre periférica obtenida de la mujer embarazada, o la fracción de ácidos nucleicos acelulares derivados de tumores en la muestra de sangre periférica obtenida del sujeto que padece un tumor, del que se sospecha que padece un tumor o que está sometido a una detección tumoral.

En realizaciones de la presente divulgación, los ácidos nucleicos acelulares son ADN.

En realizaciones de la presente divulgación, los ácidos nucleicos acelulares de la muestra biológica se secuencian mediante secuenciación de extremos pareados, secuenciación de extremo único o secuenciación de molécula única.

En realizaciones de la presente divulgación, la secuencia de referencia es una secuencia genómica de referencia, preferentemente hg19.

En realizaciones de la presente divulgación, la ventana predeterminada se obtiene por división secuencial de un cromosoma predeterminado de la secuencia genómica de referencia. Según algunas realizaciones de la presente divulgación, el cromosoma predeterminado incluye un autosoma. Preferentemente, el autosoma no incluye ninguno de los cromosomas 13, 18 ni 21. En una realización preferida de la presente divulgación, la ventana predeterminada tiene una longitud de 60 Kpb.

En realizaciones de la presente divulgación, refiriéndose a la figura 4, el aparato de recuento 200 incluye además: una unidad de alineación 210, una unidad de determinación de la posición 220 y una unidad de determinación del número 230. Específicamente, la unidad de alineación 210 está configurada para alinear el resultado de la secuenciación con la secuencia genómica de referencia, a fin de construir un conjunto de datos consistente en una pluralidad de lecturas de secuenciación cartografiadas inequívocamente, en la que cada lectura de secuenciación cartografiada inequívocamente en el conjunto de datos puede cartografiarse a una única posición en la secuencia genómica de referencia. La unidad de determinación de la posición 220 está conectada a la unidad de alineación 210 y está configurada para determinar la posición de cada lectura de secuenciación cartografiada inequívocamente en la secuencia genómica de referencia. La unidad de determinación del número 230 está conectada a la unidad de determinación de la posición 220 y está configurada para determinar el número de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de la ventana predeterminada. Por lo tanto, es fácil determinar el número de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de la ventana predeterminada, y el resultado determinado es preciso, fiable y reproducible.

Según algunas realizaciones de la presente divulgación, el aparato de determinación de la fracción de ácidos nucleicos acelulares 300 es adecuado para determinar la fracción de ácidos nucleicos acelulares procedentes de la fuente predeterminada en la muestra biológica por un peso de cada ventana predeterminada. En algunas realizaciones específicas, el peso de cada ventana predeterminada se predetermina con muestras de entrenamiento. Por lo tanto, el resultado es preciso, fiable y repetible. En una realización, el peso de cada ventana predeterminada se determina mediante al menos uno de un modelo estadístico de regresión contraída y un modelo de red neuronal. En algunas realizaciones, el modelo de red neuronal adopta un sistema de aprendizaje TensorFlow. En algunas realizaciones específicas, el sistema de aprendizaje TensorFlow incluye los siguientes parámetros: el número de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de ventanas individuales de un autosoma como capa de entrada; una fracción fetal como capa de salida; ReLu como neurona; un optimizador seleccionado entre al menos uno de Adam, SGD y Ftrl, preferentemente Ftrl. Preferentemente, el sistema de aprendizaje TensorFlow incluye además los siguientes parámetros: una tasa de aprendizaje fijada en 0,002; 1 capa oculta; y 200 neuronas en la capa oculta. Por lo tanto, el resultado es preciso y fiable.

En realizaciones de la presente divulgación, la muestra de entrenamiento es una muestra de sangre periférica con una fracción conocida de ácidos nucleicos fetales acelulares procedente de una mujer embarazada. Por lo tanto, puede determinarse eficazmente una fracción de ácidos nucleicos fetales acelulares o de ácidos nucleicos maternos acelulares en una muestra de sangre periférica procedente de una mujer embarazada sometida a detección. En algunas realizaciones específicas, la muestra de entrenamiento es una muestra de sangre periférica con una fracción conocida de ácidos nucleicos fetales acelulares procedente de una mujer embarazada con un feto masculino normal. Por lo tanto, el resultado determinado para la fracción de ácidos nucleicos fetales acelulares o de ácidos nucleicos

maternos acelulares en una muestra de sangre periférica procedente de una mujer embarazada sometida a detección es más preciso y fiable.

En algunas realizaciones de la presente divulgación, el peso de cada ventana predeterminada se determina mediante un modelo estadístico de regresión contraída que tiene la siguiente fórmula computacional:

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j ,$$

en la que $\hat{y}$ es una fracción fetal predictiva, x_j es el número de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de una ventana, β_j es un peso de una ventana, β_0 es una desviación, y β_j y β_0 se obtienen entrenando el modelo.

En algunas realizaciones de la presente divulgación, el peso de cada ventana predeterminada se determina mediante un modelo de red neuronal que tiene la siguiente fórmula computacional:

$$z_j^l = f\left(\sum_k w_{jk}^l z_k^{l-1} + b_j^l\right) ,$$

en la que *l* es un número de serie de una capa en el modelo de red neuronal, la primera capa es una capa de entrada, la última capa es una capa de salida (que tiene una sola neurona), y una capa intermedia es una capa oculta,

z_j^l es un valor para una *j*-ésima neurona en una *l*-ésima capa, z_k^{l-1} es un valor para una *k*-ésima neurona en una *(l-1)*-ésima capa, w_{jk}^l es un peso de conexión desde la *k*-ésima neurona en la *(l-1)*-ésima capa a la *j*-ésima neurona en la *l*-ésima capa, b_j^l es una desviación de entrada para la *j*-ésima neurona en la *l*-ésima capa, y w y b se obtienen entrenando el modelo. Una forma muy común de la función *f* es una unidad lineal rectificada, es decir, $f(x) = \max(0,x)$.

En algunas realizaciones, la determinación del peso por el modelo de red neuronal incluye: calcular valores para neuronas individuales capa por capa según la fórmula computacional del modelo de red neuronal, en la que el valor para la neurona en la última capa es la fracción fetal predictiva.

El dispositivo divulgado incluye además un aparato de corrección de GC (no mostrado en las figuras) conectado a cada uno de los aparatos de recuento 200 y al aparato de determinación de la fracción de ácidos nucleicos acelulares 300 y configurado para realizar la corrección de GC en el número de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de la ventana predeterminada por adelantado, con el fin de obtener el número de lecturas de secuenciación cartografiadas inequívocamente corregidas para GC que quedan dentro de la ventana predeterminada, antes de determinar la fracción de ácido nucleico acelular procedente de la fuente predeterminada en la muestra biológica. Por lo tanto, el número determinado de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de la ventana predeterminada es preciso y fiable.

Preferentemente, en alguna realización de la presente divulgación, el aparato de corrección de GC es adecuado para realizar la corrección de GC de acuerdo con las siguientes operaciones:

ajustar los números de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de ventanas predeterminadas individuales del cromosoma predeterminado a los contenidos de GC correspondientes para determinar $ER = f(gc)$;

realizar una corrección del número de lecturas de secuenciación cartografiadas inequívocamente para cada ventana predeterminada de todos los cromosomas: $ERA_i = ER_i * (ER/f(GC_i))$, $i = 1,2,3, \ldots, N$,

en la que, para una muestra, ER_i representa el número de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de una *i*-ésima ventana predeterminada, GC_i representa un contenido de GC de una secuencia de referencia para la *i*-ésima ventana predeterminada, *ER* representa un valor medio para los números de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de ventanas individuales del cromosoma predeterminado; y ERA_i representa el número de lecturas de secuenciación cartografiadas inequívocamente corregidas para GC en la *i*-ésima ventana predeterminada después de la corrección.

El dispositivo divulgado puede incluir además un aparato de determinación del sexo (no mostrado en las figuras) conectado al aparato de determinación de la fracción de ácidos nucleicos acelulares 300 y configurado para predeterminar el sexo del feto. Preferentemente, el sexo del feto se determina mediante una proporción entre el número de lecturas de secuenciación cartografiadas inequívocamente en el cromosoma Y y el número total de lecturas de secuenciación cartografiadas inequívocamente en todos los cromosomas. Por lo tanto, el resultado determinado

para la fracción de ácidos nucleicos fetales acelulares o de ácidos nucleicos maternos acelulares en una muestra de sangre periférica procedente de una mujer embarazada sometida a detección es más preciso y fiable.

Cabe señalar que la expresión "un feto/feto femenino/feto masculino normal" significa que el feto tiene cromosomas normales, por ejemplo, "un feto masculino normal" se refiere a un feto masculino con cromosomas normales. Además, la expresión "un feto/feto masculino/feto femenino normal" puede referirse a un feto único o a gemelos, por ejemplo, "un feto masculino normal" puede ser un feto único normal o gemelos normales; y "un feto normal" no limita el sexo del feto ni que sea feto único o gemelos.

Las realizaciones de la presente divulgación se describirán en detalle a continuación con referencia a ejemplos. Los expertos en la materia podrán apreciar que los siguientes ejemplos se utilizan únicamente para ilustrar la presente divulgación, por lo que no se interpretará que limiten el alcance de la misma. Un ejemplo sin condiciones especificadas se realizará en condiciones normales o como recomiende el fabricante. Los reactivos o instrumentos en los que no se especifica el fabricante son productos convencionales disponibles en el mercado.

Ejemplo 1

Se estimó la fracción fetal para 1400 muestras secuenciadas utilizando el modelo estadístico de regresión contraída de acuerdo con las siguientes etapas:

1) Dividir secuencialmente una secuencia de referencia (hg19) en ventanas adyacentes de una longitud fija (tal como 60 Kpb en este ejemplo), filtrar las ventanas en una región N y determinar el contenido de GC de cada ventana, para obtener un archivo de ventana de referencia hg19.gc;

2) Alineación: alinear las lecturas de secuenciación (cada una de 28 pb) secuenciadas mediante secuenciación SE en la plataforma CG con la secuencia de referencia (hg19) utilizando, por ejemplo, BWA V0.7.7-r441;

3) Filtración y estadísticas preliminares: seleccionar las lecturas de secuenciación cartografiadas inequívocamente a partir de los resultados de la alineación, filtrar las lecturas repetidas y las lecturas con bases no coincidentes para obtener lecturas de secuenciación eficaces, y determinar el número de lecturas de secuenciación eficaces que quedan dentro de cada ventana del archivo de ventana de referencia hgl9.gc y los contenidos de GC correspondientes;

4) Corrección de GC, que incluye las siguientes etapas específicas:

ajustar los números de lecturas de secuenciación eficaces que quedan dentro de ventanas individuales de un autosoma a los contenidos de GC correspondientes (tal como usando el ajuste de interpolación cúbica en este ejemplo) para determinar $ER = f(gc)$;

realizar la corrección en el número de lecturas de secuenciación eficaces para cada ventana de todos los cromosomas: $ERA_i = ER_i * (ER/f(GC_i))$, $i = 1,2,3, \ldots, N$,

en la que, para una muestra, ER_i representa el número de lecturas de secuenciación eficaces que quedan dentro de una *i*-ésima ventana, GC_i representa un contenido de GC de la secuencia de referencia en la *i*-ésima ventana (registros en el archivo de ventana de referencia hg19.gc), *ER* representa un valor medio para los números de lecturas de secuenciación eficaces que quedan dentro de ventanas individuales de un autosoma (tales como los cromosomas 1-22); y ERA_i representa el número de lecturas de secuenciación eficaces corregidas para GC en la *i*-ésima ventana después de la corrección de GC;

5) Determinar el sexo de un feto: comparar una proporción (%ER) del número de lecturas de secuenciación eficaces corregidas por GC en el cromosoma Y al número total de lecturas de secuenciación eficaces corregidas para GC en todos los cromosomas con un umbral especificado de [0,001, 0,003], determinar que el feto es un feto masculino si la %ER es mayor o igual a 0,003, y determinar que el feto es un feto femenino si la %ER es inferior a 0,001;

6) Estimar la fracción fetal mediante el modelo estadístico de regresión contraída, que incluye las siguientes etapas:

a) seleccionar 1000 muestras, cada una de ellas obtenida independientemente de una mujer embarazada con un feto masculino, como conjunto de entrenamiento; y seleccionar 400 muestras como conjunto de prueba, de las cuales 200 muestras han sido obtenidas cada una independientemente de una mujer embarazada con un feto masculino y 200 muestras han sido obtenidas cada una independientemente de una mujer embarazada con un feto femenino;

b) estimar un peso de cada ventana en el autosoma (excluyendo los cromosomas 13, 18 y 21) utilizando el modo de regresión contraída para el conjunto de entrenamiento, es decir, determinando una desviación β_0 y un peso β_j que minimicen la siguiente fórmula:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2,$$

$$\mathrm{RSS}=\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

en la que , y_i es una fracción fetal estimada con el cromosoma Y, x_{ij} es el número de lecturas de secuenciación eficaces corregidas para GC que quedan dentro de una ventana, y λ es un coeficiente de un término de regularización de segundo orden, que se determina para minimizar la función objetivo del conjunto de validación mediante validación cruzada, mostrándose la distribución de pesos estimados en la figura 5; y

c) estimar la fracción fetal de la muestra en el conjunto de prueba con los pesos conocidos de acuerdo con la siguiente fórmula:

$$\hat{y}=\hat{\beta}_0+\sum_{j=1}^{p}\hat{\beta}_j x_j$$

en la que $\hat{y}$ es una fracción fetal predictiva, x_j es el número de lecturas de secuenciación eficaces corregidas para GC que quedan dentro de una ventana, β_j es un peso de una ventana, β_0 es una desviación, y β_j y β_0 se obtienen entrenando el modelo.

En la figura 6 se muestra una correlación entre las fracciones fetales estimadas por el modelo estadístico de regresión contraída y la estimada por el cromosoma Y, a partir de la cual puede verse que existe una fuerte correlación entre ellas (r = 0,92; valor de p < 1e-10), lo que indica que la fracción de ácidos nucleicos fetales acelulares estimada por el procedimiento según las realizaciones de la presente divulgación es precisa y fiable.

Ejemplo 2

Se estimó la fracción fetal para 1400 muestras secuenciadas utilizando el modelo de red neuronal en el banco TensorFlow de acuerdo con las siguientes etapas: las etapas 1)-5) son las mismas que las descritas anteriormente en el ejemplo 1;

6) Estimar la fracción fetal utilizando el modelo de red neuronal en el banco TensorFlow, que incluye las siguientes etapas específicas:

a) seleccionar 1000 muestras, cada una obtenida independientemente de una mujer embarazada con un feto masculino como conjunto de entrenamiento; y seleccionar 400 muestras como conjunto de prueba, de las cuales 200 muestras han sido obtenidas cada una independientemente de una mujer embarazada con un feto masculino y 200 muestras han sido obtenidas cada una independientemente de una mujer embarazada con un feto femenino, normalizar todos los datos (todos los números de lecturas de secuenciación eficaces corregidas para GC que quedan dentro de cada ventana), de modo que el valor medio de las variables en todas las muestras sea 0, y la desviación estándar sea 1;

b) construir un modelo de red neuronal en el que el número normalizado de lecturas de secuenciación eficaces que quedan dentro de cada ventana de un autosoma relativamente estable se toma como capa de entrada, una sola neurona se toma como capa de salida (correspondiente a la fracción fetal), no hay capa oculta, se selecciona ReLU como tipo de neurona y Adam como optimizador;

c) aplicar el modelo de red neuronal (aprendizaje) en el conjunto de entrenamiento para predecir la fracción fetal, y ajustar una tasa de aprendizaje de acuerdo con el cambio del efecto de aprendizaje en cada ronda, a fin de maximizar la tasa de aprendizaje al tiempo que se garantiza el efecto de aprendizaje para el conjunto de entrenamiento sin fluctuaciones repetidas, en el que el procedimiento de aprendizaje consiste en calcular un valor z_k^l para cada neurona capa por capa de acuerdo con la siguiente fórmula basándose en el número normalizado z_k^1 de lecturas de secuenciación eficaces que quedan dentro de cada ventana:

$$z_j^l = f(\sum_k w_{jk}^l z_k^{l-1} + b_j^l)$$

en la que *l* es un número de serie de una capa en el modelo de red neuronal, la primera capa es una capa de entrada, la última capa es una capa de salida (que tiene una sola neurona), y una capa intermedia es una capa oculta, z_j^l es un valor para una *j*-ésima neurona en una *l*-ésima capa, z_k^{l-1} es un valor para una *k*-ésima neurona en una *(l-1)*-ésima capa, w_{jk}^l un peso de conexión de la *k*-ésima neurona en la *(l-1)*-ésima capa a la *j*-ésima neurona en la *l*-ésima capa, b_j^l es una desviación de entrada para la *j*-ésima neurona en la *l*-ésima capa, y una forma más común de la función *f* es una unidad lineal rectificada, es decir, *f(x)* = máx(0,*x*);

para cada muestra *{s}*, comparar un valor z_s^L para una neurona en la capa de salida con una fracción fetal y_s estimada por el cromosoma Y, y ajustar el peso de conexión w_{jk}^l y la desviación b_j^l para cada capa para minimizar $\sum_s (z_s^L - y_s)^2$;

d) entrenar tantas rondas como permita la capacidad computacional, hasta que se sature el efecto de aprendizaje;

e) cambiar a otros optimizadores (tales como SGD, Ftrl, etc.), repetir las etapas b)-d), y seleccionar un optimizador óptimo basándose en los efectos de aprendizaje;

f) intentar añadir un término de regularización de segundo orden λ al modelo de red neuronal y ajustar su tamaño para observar los efectos de aprendizaje antes y después de añadir y ajustar el tamaño del término de regularización de segundo orden, en el que la importancia del término de regularización de segundo orden es que ya no se trata de buscar el mínimo de $\sum_s (z_s^L - y_s)^2$; sino buscar el mínimo de $\sum_s (z_s^L - y_s)^2 + \lambda \sum w^2$ en el proceso de aprendizaje;

g) añadir una capa oculta, ajustar el número de neuronas en la capa oculta, repetir las etapas b)-f), y seleccionar una arquitectura de capa oculta óptima basándose en los efectos de aprendizaje;

h) entrenar el modelo de red neuronal optimizado en el conjunto de entrenamiento para obtener los parámetros óptimos como se muestra en la tabla 1 y una distribución de pesos (pesos promedio de las neuronas en la capa de entrada a la capa oculta) de ventanas individuales (refiriéndose a la figura 7); y

i) estimar la fracción fetal de la muestra en el conjunto de prueba con el modelo de red neuronal entrenado calculando un valor zk para cada neurona capa por capa de acuerdo con la siguiente fórmula basándose en el número normalizado z_k^1 de lecturas de secuenciación eficaces que quedan dentro de cada ventana, y un valor para una neurona en la última capa que es la fracción fetal predictiva:

$$z_j^l = f(\sum_k w_{jk}^l z_k^{l-1} + b_j^l)$$

Tabla 1: Parámetros óptimos obtenidos por el modelo de red neuronal en el banco TensorFlow

Tasa de aprendizaje	Optimizador	Fuerza de regularización de segundo orden	Número de capas ocultas	Número de neuronas en la capa oculta
0,002	Ftrl	0	1	200

En la figura 8 se muestra una correlación entre las fracciones fetales estimadas por el modelo de red neuronal y la estimada por el cromosoma Y, a partir de la cual puede verse que existe una fuerte correlación entre ellas (r = 0,982; valor de *p* < 1e-10), lo que indica que la fracción de ácidos nucleicos fetales acelulares estimada por el procedimiento según las realizaciones de la presente divulgación es precisa y fiable.

Por último, los inventores analizaron una correlación entre el modelo estadístico de regresión contraída y el modelo de red neuronal en términos de estimación para la fracción fetal femenina, cuyo resultado se muestra en la figura 9, a partir de la cual se puede ver claramente que las fracciones fetales obtenidas por los dos modelos están altamente correlacionadas (r = 0,935; valor de p < 1e-10).

La referencia a lo largo de la presente memoria descriptiva a "una realización", "algunas realizaciones", "otro ejemplo", "un ejemplo", "un ejemplo específico" o "algunos ejemplos" significa que un rasgo, una estructura, un material o una característica concretos descritos en relación con la realización o ejemplo está incluido en al menos una realización o ejemplo de la presente divulgación. Por lo tanto, las apariciones de expresiones tales como "en algunas realizaciones", "en una realización", "en otro ejemplo", "en un ejemplo", "en un ejemplo específico" o "en algunos ejemplos" en diversos lugares a lo largo de la presente memoria descriptiva no se refieren necesariamente a la misma realización o ejemplo de la presente divulgación. Además, los rasgos, las estructuras, los materiales o las características concretos pueden combinarse de cualquier manera adecuada en una o más realizaciones o ejemplos.

REIVINDICACIONES

1. Un procedimiento para determinar una fracción de ácidos nucleicos acelulares procedentes de una fuente predeterminada en una muestra biológica, que comprende:

(1) realizar la secuenciación de los ácidos nucleicos acelulares contenidos en la muestra biológica, a fin de obtener un resultado de la secuenciación consistente en una pluralidad de lecturas de secuenciación;

(2) alinear el resultado de la secuenciación con una secuencia de referencia, a fin de determinar el número de lecturas de secuenciación que quedan dentro de una ventana predeterminada en el resultado de la secuenciación,

en el que la secuencia de referencia es una secuencia genómica de referencia,

en el que la ventana predeterminada se obtiene por división secuencial de un cromosoma predeterminado de la secuencia genómica de referencia; y

(3) determinar la fracción de ácidos nucleicos acelulares procedentes de la fuente predeterminada en la muestra biológica basándose en el número de lecturas de secuenciación que quedan dentro de la ventana predeterminada, en el que la fracción de ácidos nucleicos acelulares procedentes de la fuente predeterminada en la muestra biológica se determina por un peso de cada ventana predeterminada,

en el que el peso de cada ventana predeterminada se predetermina con muestras de entrenamiento y se determina mediante al menos uno de un modelo estadístico de regresión contraída y un modelo de red neuronal,

en el que la muestra de entrenamiento es una muestra de sangre periférica con una fracción conocida de ácidos nucleicos fetales acelulares procedentes de una mujer embarazada,

en el que la muestra de entrenamiento es una muestra de sangre periférica con una fracción conocida de ácidos nucleicos fetales acelulares procedentes de una mujer embarazada con un feto masculino normal, en el que el feto masculino normal se refiere a un feto masculino con cromosomas normales,

en el que la muestra biológica es una muestra de sangre periférica,

en el que el ácido nucleico acelular procedente de la fuente predeterminada es al menos uno seleccionado de entre los siguientes:

ácidos nucleicos fetales acelulares en una muestra de sangre periférica obtenida de una mujer embarazada; y

ácidos nucleicos maternos acelulares en una muestra de sangre periférica obtenida de una mujer embarazada.

2. El procedimiento según la reivindicación 1, en el que el cromosoma predeterminado comprende un autosoma, preferentemente, en el que el autosoma no comprende ninguno de los cromosomas 13, 18 ni 21.

3. El procedimiento según una cualquiera de las reivindicaciones 1 o 2, en el que la etapa (2) comprende además:

(2-1) alinear el resultado de la secuenciación con la secuencia genómica de referencia, a fin de construir un conjunto de datos formado por una pluralidad de lecturas de secuenciación cartografiadas inequívocamente, en el que cada lectura de secuenciación cartografiada inequívocamente del conjunto de datos corresponde a una única posición en la secuencia genómica de referencia;

(2-2) determinar la posición de cada lectura de secuenciación cartografiada inequívocamente en la secuencia genómica de referencia; y

(2-3) determinar el número de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de la ventana predeterminada.

4. El procedimiento según una cualquiera de las reivindicaciones 1 a 3, en el que, en la etapa (3),

el modelo de red neuronal adopta un sistema de aprendizaje TensorFlow,

preferentemente, en el que el sistema de aprendizaje TensorFlow comprende los siguientes parámetros:

el número de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de ventanas individuales de un autosoma como capa de entrada;

una fracción fetal como capa de salida;

ReLu como neurona;

un optimizador seleccionado entre al menos uno de Adam, SGD y Ftrl,

preferentemente, en el que el optimizador es Ftrl,

preferentemente, en el que el sistema de aprendizaje TensorFlow comprende además los siguientes parámetros:

una tasa de aprendizaje fijada en 0,002;

1 capa oculta; y

200 neuronas en la capa oculta.

5. El procedimiento según la reivindicación 4, en el que el peso de cada ventana predeterminada se determina mediante un modelo estadístico de regresión contraída que tiene la siguiente fórmula computacional:

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j ,$$

en la que $\hat{y}$ es una fracción fetal predictiva, x_j es el número de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de una ventana, β_j es un peso de una ventana, β_0 es una desviación, β_j y β_0 se obtienen entrenando el modelo.

6. El procedimiento según la reivindicación 4, en el que el peso de cada ventana predeterminada se determina mediante un modelo de red neuronal que tiene la siguiente fórmula computacional:

$$z_j^l = f\left(\sum_k w_{jk}^l z_k^{l-1} + b_j^l\right) ,$$

en la que l es un número de serie de una capa en el modelo de red neuronal, zj es un valor para una *j*-ésima neurona en una *l*-ésima capa, z_k^{l-1} es un valor para una *k*-ésima neurona en una *(l-1)*-ésima capa, w_{jk}^l un peso de conexión desde la *k*-ésima neurona en la *(l-1)*-ésima capa a la *j*-ésima neurona en la *l*-ésima capa, b_j^l es una desviación de entrada para la *j*-ésima neurona en la *l*-ésima capa, y w y b se obtienen entrenando el modelo,

preferentemente, en el que la determinación del peso de cada ventana predeterminada determinada por el modelo de red neuronal comprende:

calcular un valor para cada neurona capa por capa según la fórmula computacional del modelo de red neuronal, en el que un valor para una neurona en la última capa es la fracción fetal predictiva.

7. El procedimiento según una cualquiera de las reivindicaciones 1 a 6, en el que, antes de la etapa (3), se realiza la corrección de GC en el número de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de la ventana predeterminada por adelantado, para obtener el número de lecturas de secuenciación cartografiadas inequívocamente corregidas para GC que quedan dentro de la ventana predeterminada, preferentemente, en el que la corrección de GC se realiza mediante:

ajustar los números de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de ventanas predeterminadas individuales del cromosoma predeterminado a los contenidos de GC correspondientes para determinar *ER = f(gc)*;

realizar una corrección en el número de lecturas de secuenciación cartografiadas inequívocamente para cada ventana predeterminada del cromosoma predeterminado:

$$ERA_i = ER_i * (\overline{ER}/f(GC_i)), \ i = 1,2,3,\ldots,N ,$$

en la que, para una muestra, ER_i representa el número de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de una *i*-ésima ventana predeterminada, GC_i representa un contenido de GC de una secuencia de referencia para la *i*-ésima ventana predeterminada, *ER* representa un valor medio para los números de lecturas de secuenciación cartografiadas inequívocamente que quedan dentro de ventanas predeterminadas individuales del cromosoma predeterminado; y ERA_i representa el número de

lecturas de secuenciación cartografiadas inequívocamente corregidas para GC en la *i*-ésima ventana predeterminada después de la corrección.

8. El procedimiento según una cualquiera de las reivindicaciones 1 a 7, en el que, antes de la etapa (3), se predetermina el sexo del feto,

preferentemente, en el que el sexo del feto se determina por una proporción entre el número de lecturas de secuenciación cartografiadas inequívocamente en el cromosoma Y y el número total de lecturas de secuenciación cartografiadas inequívocamente en todos los cromosomas.

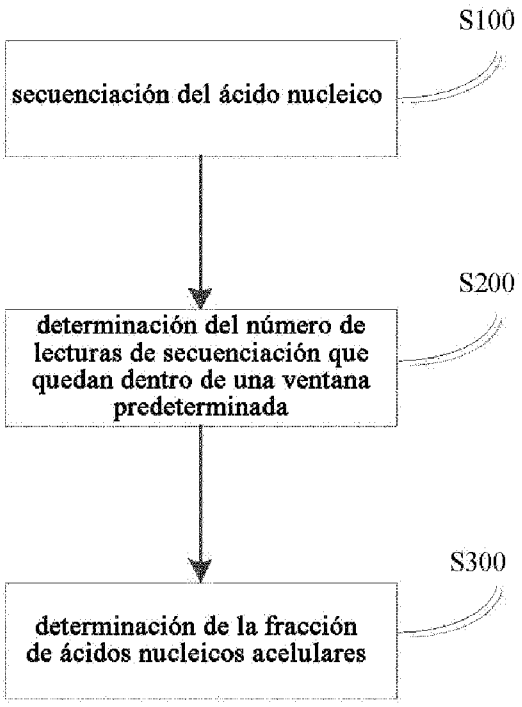
S100
secuenciación del ácido nucleico
S200
determinación del número de lecturas de secuenciación que quedan dentro de una ventana predeterminada
S300
determinación de la fracción de ácidos nucleicos acelulares

Fig. 1

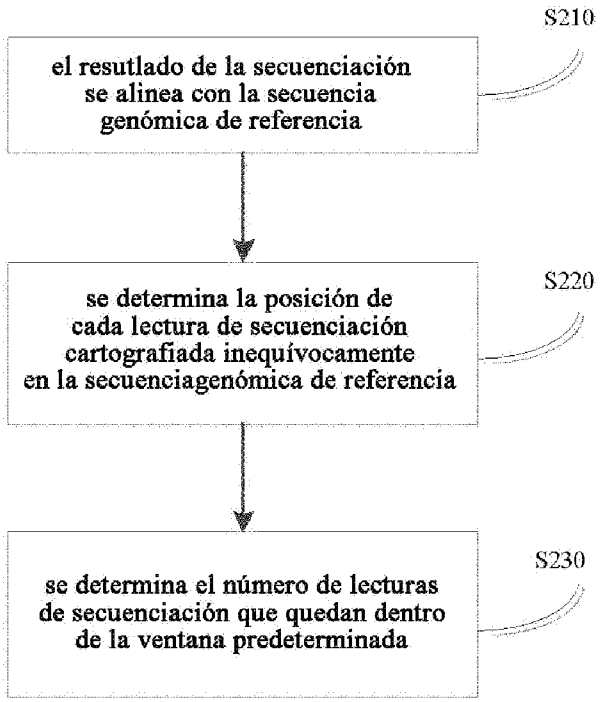
S210
el resutlado de la secuenciación se alinea con la secuencia genómica de referencia
S220
se determina la posición de cada lectura de secuenciación cartografiada inequívocamente en la secuenciagenómica de referencia
S230
se determina el número de lecturas de secuenciación que quedan dentro de la ventana predeterminada

Fig. 2

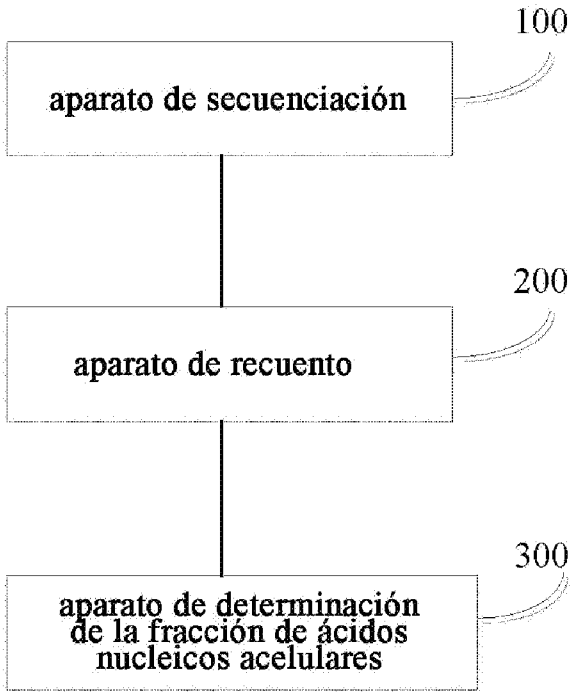
100
aparato de secuenciación
200
aparato de recuento
300
aparato de determinación de la fracción de ácidos nucleicos acelulares

Fig. 3

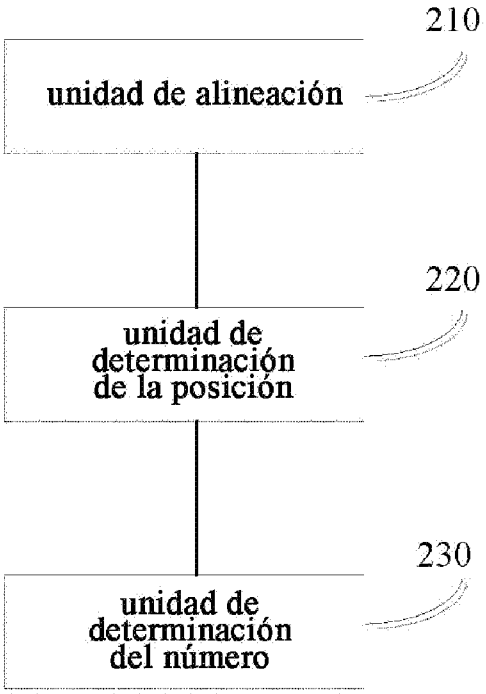
210
unidad de alineación
220
unidad de determinación de la posición
230
unidad de determinación del número

Fig. 4

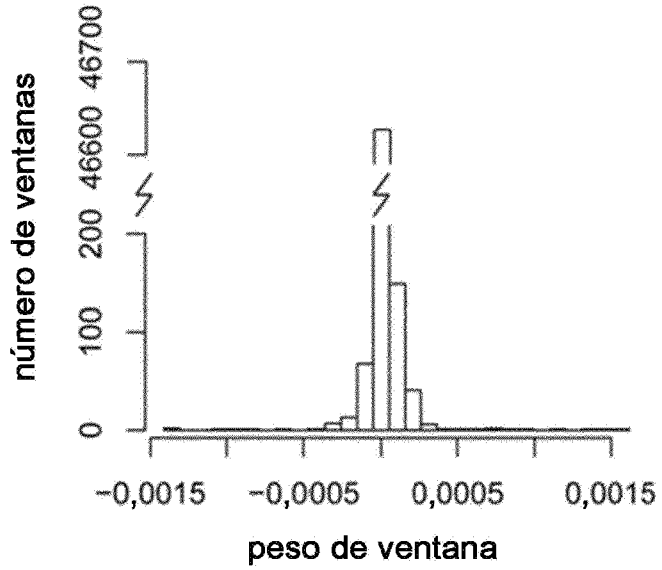
número de ventanas
0
100
200
46600
46700
−0,0015
−0,0005
0,0005
0,0015
peso de ventana

Fig. 5

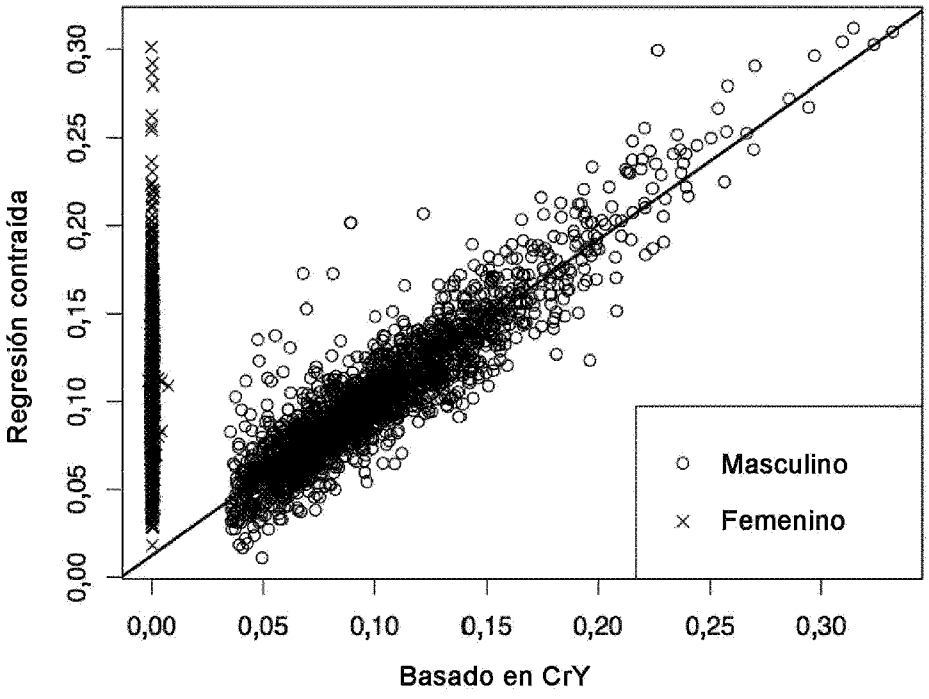
Regresión contraída
0,00
0,05
0,10
0,15
0,20
0,25
0,30
Basado en CrY
Masculino
Femenino

Fig. 6

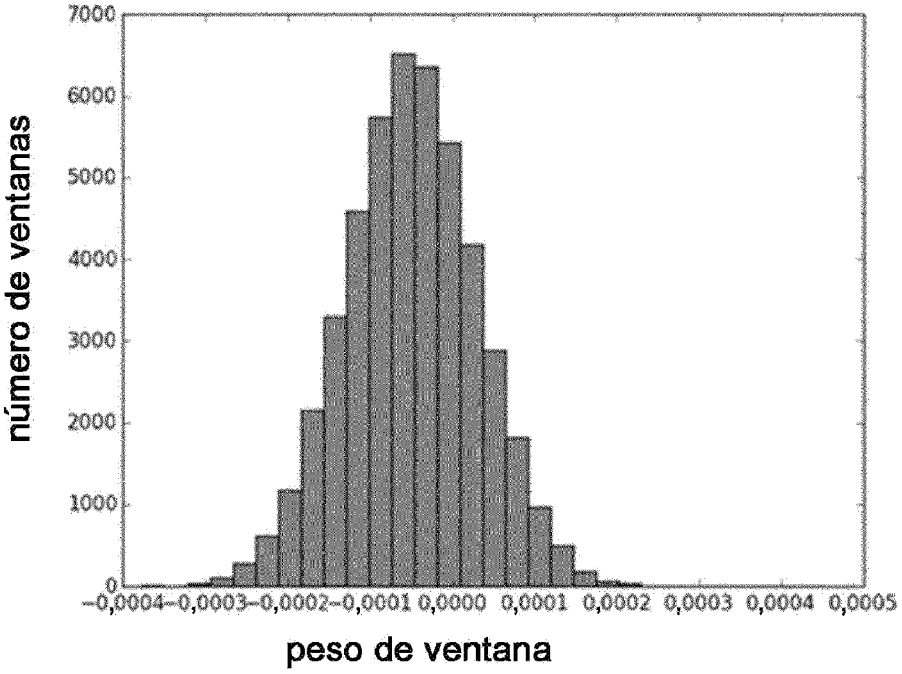
número de ventanas
7000
6000
5000
4000
3000
2000
1000
0
−0,0004 −0,0003 −0,0002 −0,0001 0,0000 0,0001 0,0002 0,0003 0,0004 0,0005
peso de ventana

Fig. 7

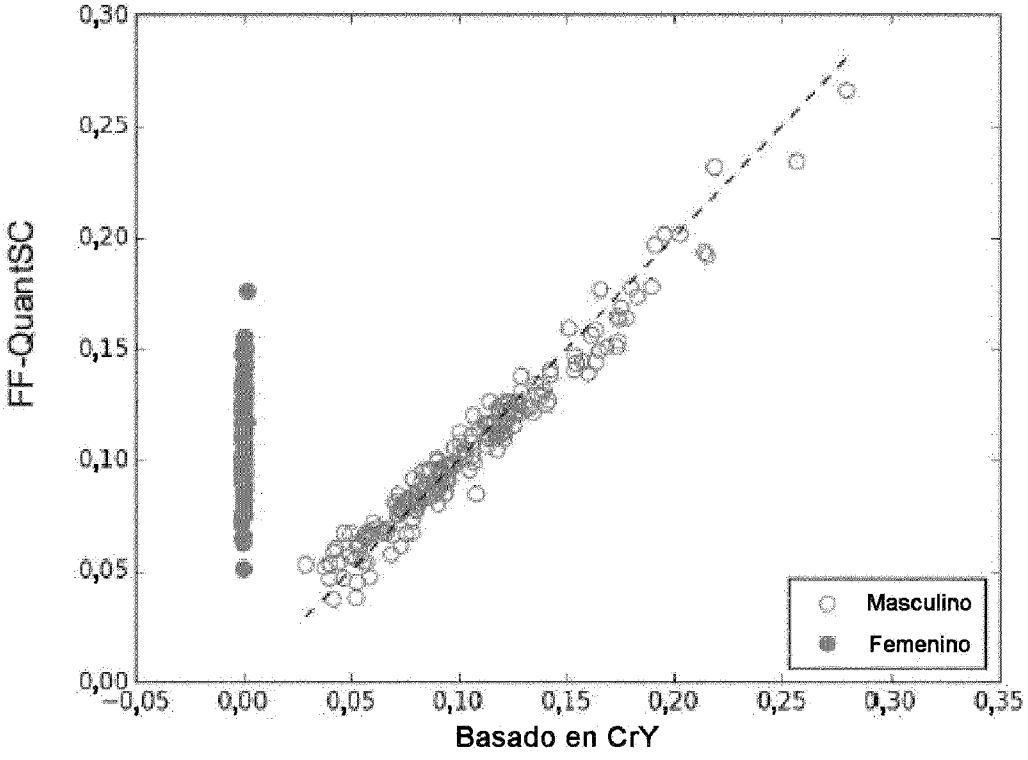
FF-QuantSC
0,30
0,25
0,20
0,15
0,10
0,05
0,00
−0,05 0,00 0,05 0,10 0,15 0,20 0,25 0,30 0,35
Basado en CrY
Masculino
Femenino

Fig. 8

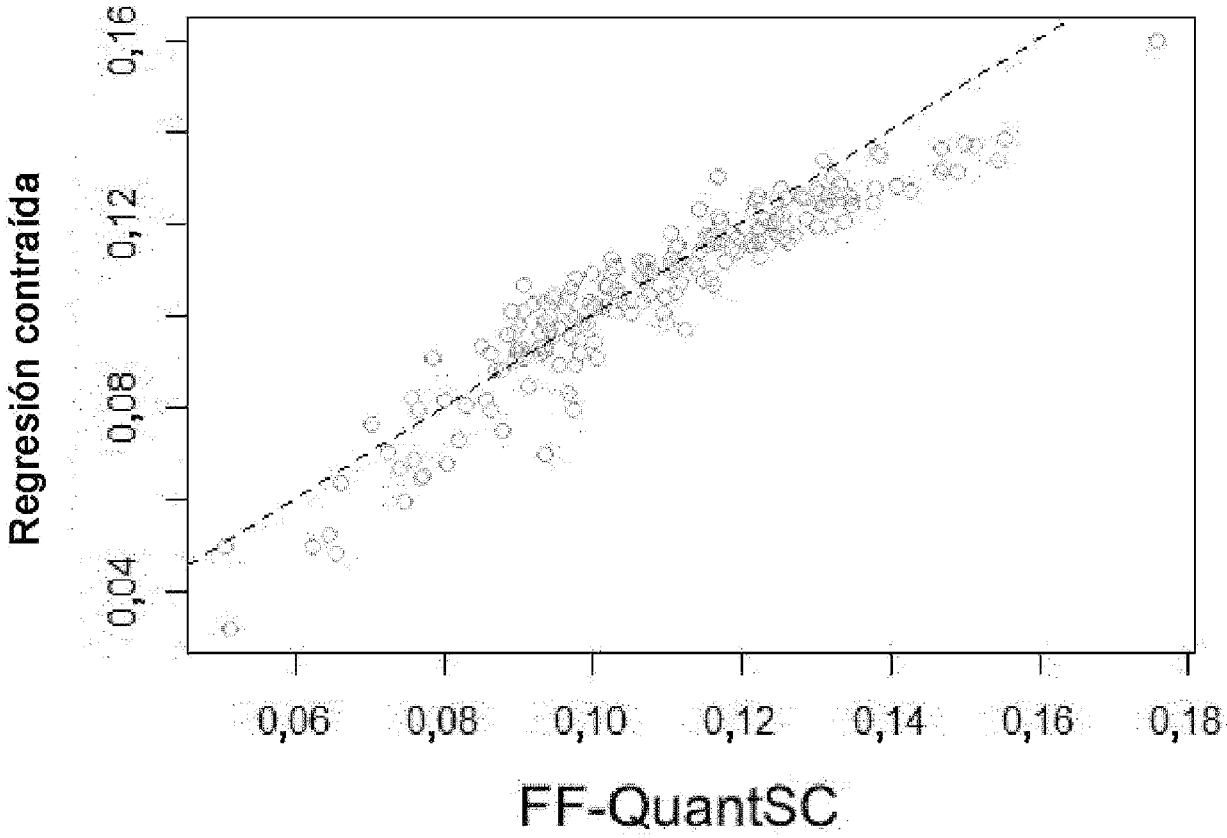
Regresión contraída
0,04
0,08
0,12
0,16
0,06
0,08
0,10
0,12
0,14
0,16
0,18
FF-QuantSC

Fig. 9